ISBN 978-7-5352-3854-2

9 787535 238542 >

PRICE RMB 560 YUAN
（Volume Ⅰ，Volume Ⅱ）

DCABES 2007 PROCEEDINGS

Editor in Chief Guo Qingping
Associate Editor in Chief Guo Yucheng

Volume Ⅱ

2007 International Symposium on
Distributed Computing and Applications to
Business, Engineering and Science

DCABES 2007

PROCEEDINGS

Volume Ⅱ

Editor in Chief Guo Qingping
Associate Editor in Chief Guo Yucheng

Hubei Science and Technology Press, Wuhan China

2007 International Symposium
On Distributed Computing and Applications
To Business, Engineering and Science

DCABES 2007 PROCEEDINGS

Volume II

Editor in Chief: Guo Qingping
Associate Editor in Chief: Guo Yucheng

Yichang, China

August 14-17, 2007

Hubei Science and Technology Press, Wuhan, China

图书在版编目(CIP)数据
2007 年电子商务、工程及科学领域的分布式计算和应用国际学术研讨会论文集 / 郭庆平、郭羽成 主编.
--武汉: 湖北科学技术出版社,2007.8
ISBN 978-7-5352-3854-2
Ⅰ.2… Ⅱ.郭… Ⅲ.分布式计算机-计算机应用-国际学术会议-文集 Ⅳ.TP338.8-53
中国版本图书馆 CIP 数据核字(2007)第 112318 号

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the Wuhan University of Technology, the Hubei Science and Technology Press, the Natural Science Foundation of China, the Ministry of Education, China, or other sponsors and organizers.

Organized by
WUT Wuhan University of Technology

Co-organized by
CAA Computer Academic Association of Hubei Province & Wuhan Metropolis

Sponsored by

WUT Wuhan University of Technology
MOE Ministry of Education, China
NSFC National Nature Science Foundation of China

DCABES 2007 PROCEEDINGS
Editor in Chief :**Guo Qingping,** Associate Editor in Chief: **Guo Yucheng**
Editorial Production by **Wu Ruilin, Tang Jie;** Cover Art Production by **Da Min, Wan Xiaofen**
Published by ***Hubei Science and Technology Press, Wuhan, China*** Tel: +86-(0)27-87679468
Address: The 13th Floor of Block B, 268 Xiongchu Avenue, Wuhan, Post Code: 430070 P. R. China
Printed in Wuhan, China by Youngster Union Printery, Post Code: 430063
880mm×1230mm sixteenmo 71.5 sheets 1 500 k Characters First Edition August 2007
First Impression September 2007

ISBN 978-7-5352-3854-2 Price RMB 280 Yuan

QoS of Book Printing and Bounding is guaranteed by the printery.

CONTENTS

Volume I

Distributed/Parallel Algorithms

Distributed/Parallel Applications

Network Techniques and Applications

Distributed System Architectures

Software Architecture/ Parallel Programming Language

Distributed Operating System Techniques

Network/Web Security

Neural Network Computing

Volume II

Cluster Computing, Parallel Processing and Grid Computing

Multi-agent Application

Distributed Database and Data Mining

Web Service Applications and Web-based Computing

E-Commence Techniques and Applications

E- Education

Graphics, Image, Vision and Voice Processing

Embedded System, Hardware Design and Diagnosis

PREFACE

The DCABES is a community working in the area of Distributed Computing and its Applications in Business, Engineering, and Sciences, and is responsible for organizing meetings and symposia related to the field. The DCABES 2007 is the Sixth International Conference on Distributed Computing and Applications for Business, Engineering and Sciences held on 14-17 August 2007 in the Three Gorges, Yichang, Hubei, China. It is the third time for the DCABES international conference to be organized by School of Computer Science and Technology, Wuhan University of Technology.

As in previous conferences, the DCABES intends to bring together researchers and developers in the academic field and industry from around the world to share their research experience and to explore research collaboration in the areas of distributed parallel processing and applications.

In recent years, more and more attentions have been put on to the distributed parallel computing. I am confident that the distributed parallel computing will play an even greater role in the near future, since distributed computing resources, once properly cooperated together, will achieve a great computing power and get a high ratio of performance/price in parallel computing. In fact the grid computing is a direct descendent of the distributed computing.

We are gratified that the DCABES 2007 has received more than 500 papers submission，which cover a wide range of topics, such as Grid Computing, Mobile Computing, Parallel/Distributed Algorithms, Image Processing and Multimedia Applications, Parallel/Distributed Computational Methods in Engineering, System Architectures, Networking and Protocols, Web-Based Computing & E-Business, E-Education, Network Security and various types of applications etc.

All papers contained in this Proceedings are peer-reviewed and carefully chosen by members of Scientific Committee and external reviewers. Papers accepted or rejected are based on majority opinions of the referees. All papers contained in this Proceedings give us a glimpse of what future technology and applications are being studied in the distributed parallel computing area in the world.

I would like to thank all members of the Scientific Committee, the local organizer committee, the external reviewers for selecting papers. Special thanks are due to Professor, Dr. Choi-Hong LAI, who co-chaired the Scientific Committee with me. It is indeed a pleasure to work with him and obtain his suggestions.

Also sincere thanks should be forward to Mr Tsui Y M Thomas, Chinese University of Hong Kong, Professor Xu W.B., Southern Yangtze University for their enthusiastically taking part in and supporting the DCABES conference.

I am also grateful to Prof. Souheil Khaddaj, Kingston University, London, UK.; Prof. V.P. Kutepov, Moscow Power Engineering Institute (Technical University), Russia; Prof. A J Davies, University of Hertfordshire, UK; Prof. Xiao-ChuanCai, University of Colorado at Boulder, USA; Prof. Choi-Hong Lai, University of Greenwich, London, UK for their contributions of keynote speeches in the conference.

Sincerely thanks should be forwarded to the Natural Science Foundation of China (NSFC), the China Ministry of Education (MOE), without their supports the DCABES 2007 could not be held in China successfully. We would also like to thank the WUT (Wuhan University of Technology, China), the National Parallel Computing Society of China (NPCS), the ISTCA (International Science and Technology Cooperation of Hubei Province, China), and the CAA (Computer Academic Association of Hubei Province & Wuhan Metropolis, China) for their supports as local organizers of the conference.

Finally I should also thank A/Professor Jian Guo for his efforts in conference organizing activities. The special thanks also should be given to my graduate students, Mr. Shadi Ibrahim for the conference website design, Mr. YeTian Li, Liang Huang and ZhiChao Yan for their efforts in organizing activities. It also should be mentioned that my graduate students, Mr YeTian Li, Liang Huang, HaiXiong An, Ms LiangLiang Wang, Yang Yang, Lin Chen, PengPeng Duan, Mr. YongQin Jia, Zhen Zhou, YuZhong Chao of the grade 2005; Mr. ZhiChao Yan, Peng Cui, Ms JuanJuan Zhao, Mr. Lin Hu, Wei Tang, GuangYou Zhou, YiFan Huang, Fan Yang of the grade 2006 spent a lot of time and efforts typesetting the proceedings. Without their help the proceedings could not looks so good.

Enjoy your stay in Three Gorges of the Yangtze River, China. Hope to meet you again at the DCABES 2008.

Guo, Professor Qingping
Chair of the DCABES2007
Dept. of Computer Science
Wuhan University of Technology
Wuhan, China

COMMITTEES

Honorary Chair
Zhou, Professor Zude, President of the WUT, China

Chair of Scientific Committee
Guo, Professor Q. P., Wuhan University of Technology

Co-Chair of Scientific Committee
Lai, Dr. Choi-Hong, University of Greenwich

Chair of Organizer Committee
Guo, Professor Q. P., Wuhan University of Technology

Steering Committee
Guo, Professor Q.P. (Co-Chair) Wuhan University of Technology, Wuhan, China
Lai, Professor C.-H. (Co-Chair) University of Greenwich, UK
Tsui, Thomas. Chinese University of Hong Kong, Hong Kong, China
Xu, Professor W.B. Southern Yangtze University, Wuxi, China

Scientific Committee (in alphabetical order)
Cai, Professor X.C. University of Colorado, Boulder, U.S.A.
Cao, Professor J.W. R&D Centre for Parallel Algorithms and Software,Beijing, China
Chi, Professor X.B. Academia Sinica, Beijing, China
Guo, Professor Q.P. Wuhan University of Technology, Wuhan, China
Ho, Dr. P. T. University of Hong Kong, Hong Kong, China
Jesshope, Professor C. University of Amsterdam, the Netherlands
Kang, Professor L.S. Wuhan University, China
Keyes, Professor D.E. Columbia University, USA
Lai, Professor C.-H. University of Greenwich, UK
Lee, Dr. John. Hong Kong Polytechnic, Hong Kong, China
Liddell, Professor H. M. Queen Mary, University of London, UK
Lin, Dr. H.X. Delft University of Technology, Delft, the Netherlands
Lin, Dr. P. National University of Singapore, Singapore
Loo, Dr. Alfred Hong Kong Lingnan University, Hong Kong, China
Ng, Professor Michael, Baptist University of Hong Kong, China
Pan, Professor Yi, Computer Science, Georgia State University, Atlanta, USA
Sloot, Professor P.M.A. University of Amsterdam, Amsterdam the Netherlands
Sun, Professor J. Academia Sinica, Beijing, China
Tsui, Mr.Thomas Chinese University of Hong Kong, Hong Kong, China
Xu, Professor W.B. Southern Yangtze University, Wuxi, China
Zhang, Professor Jun. University of Kentucky, USA
Zhou, Professor Jun, Computing Math, Chinese University of Hong Kong

Local Organizing Committee
Zhou, Professor Z.D. (Honorary Chair)
President of Wuhan University of Technology, Wuhan, China
Guo, Professor Q.P. (Chair) Wuhan University of Technology, Wuhan, China
Zhong, Professor L . (Co-Chair) Wuhan University of Technology, Wuhan, China
Liu, Professor Q. Wuhan University of Technology, Wuhan, China
Xu, Professor H.Z. Wuhan University of Technology, Wuhan, China
Chen, Professor H. Wuhan University of Technology, Wuhan, China
Xu, Professor N. Wuhan University of Technology, Wuhan, China
Zeng, Professor C.N. Wuhan University of Technology, Wuhan, China
Zhang, Professor H. M. Wuhan University of Technology, Wuhan, China
He, Professor Y. X. Wuhan University, Wuhan, China
King, Professor Hai, Hua Zhong University of Science and Technology, Wuhan, China
Tan, Professor L.S. Central China Normal University, Wuhan, China
Kang, Professor L.S. Wuhan University, Wuhan, China
Lu, Professor J.G. South Central China Nationality University

Cluster Computing, Parallel Processing and Grid Computing

A Framework for Data Management in the Grid

Thi-Mai-Huong Nguyen, Frédéric Magoulès
Applied Mathematics and Systems Laboratory, Ecole Centrale Paris
Grande Voie des Vignes, 92295 Châtenay-Malabry Cedex, France
Email: mai-huong.nguyen@ecp.fr, frederic.magoules@hotmail.com

ABSTRACT

In Grid Computing, a job could be executed on a node that is geographically far away from its data files. These files are stored in heterogeneous storage systems located at geographically distributed virtual organizations. The current approach includes explicit data file transfers to execution nodes, which forces users to deal with different administrative policies at each site and various data access mechanisms on each storage system. This implies a lot of human interventions in order to develop dedicated programs and scripts for data transfers for job execution. This paper presents GRAVY, a framework which enables the data management between distributed file systems irrespective of their heterogeneity. This feature enables high-level schedulers integrated with GRAVY to control data placements like computational jobs (i.e., they can be queued, scheduled and monitored). GRAVY supports multiple data transport protocols and can be extended easily.

Keywords: Distributed Computing, Data Management, Data Intensive Applications, Grid Computing, Virtual File System.

1. INTRODUCTION

In recent years, the data requirements for scientific applications have been growing dramatically in both volume and scale. Much scientific research is now data intensive. Today information technology must cope with an ever-increasing amount of data, which continues to increase rapidly each year, and they are expected to reach the exabyte (1 million terabytes) scale by around 2015 [1].

Grid technology [2] enables access and sharing of computing and data resources across distributed sites for execution of data intensive applications. However, the Grid is also a complex environment which is composed of various and heterogeneous machines distributed in many different administrative domains. The goal of Grid Computing is to provide transparent access to their resources in the way that the impact on applications is minimized from internal management mechanism of the Grid. This transparency feature must be applied to access and management of data for the execution of data-intensive applications in the Grid. The emphasis lies on providing common interfaces between existing data storage systems in order to make them work seamlessly. This will not only liberate novice Grid users (i.e., scientists) from data access related issues in order to concentrate to their problems in their fields but also limit the change of interfaces between existing applications. The contribution of this paper is a novel framework called *GRAVY (GRid-enAbled Virtual file sYstem)* that allows users to transparently access data in grid environment irrespective of transport protocols.

The next section of the paper describes data access problems in Grid environments, which lead to the motivation of our work. Then, we present in Section 3 the architecture of GRAVY framework. We have implemented a prototype of GRAVY framework that we are proposing in Java, which allows users to have the view of a unified location-transparent file system of the Grid and to access to this system without being familiar with protocols technical details. In Section 4, we present the experimental results performed to evaluate this prototype. Section 5 gives an overview of related work. Finally, the paper concludes with some final remarks in Section 6.

2. GRID DATA ACCESS CHALLENGES

2.1 Pre-Transfer Files for Computational Jobs

In Grid Computing, it is the role of grid schedulers to choose a computing node for job execution. Therefore, the job may execute on a chosen node that is geographically far away from its data files. In such situation, the user is expected to explicitly transfer the data files to the chosen node before the job is started (i.e., stage-in) and copy back the result files to a third-party storage (e.g., user local home storage) after the job is completed (i.e., stage-out).

Firstly, this file staging solution is somehow a burden for the user as it supposes the user has to know in advance all of the data files required by the computational job. Indeed, the user doesn't have the knowledge of the server that will be chosen for the computations. Generally, the choice of computational server is done by grid schedulers. Secondly, depending on the specific application, the user has to know how to map command-line arguments (i.e., job namespace) to physical location of data files (i.e., user namespace) for input files and for output files. In particular, the user has to know internal file system organization of the chosen node, which is error-prone. Thirdly, it is impossible to realize computational workflow in which the output of one job is the input of another job.

2.2 Heterogeneous Computing Environment

As indicated previously, a frequent obstacle to the creation of computational jobs that operate effectively in Grid environments is access to remote data. This problem is challenging because the Grid is a heterogeneous computing environment including multiple administrative domains. Data at each domain is accessed through different mechanism including how the data is organized, which transfer protocols are supported, and how the authentication is carried out. Users are forced to deal with such aspect whenever they want to access data at different administrative domains and it is difficult to efficiently share data between these domains.

Although GridFTP [3], which extends FTP has recently been promoted as the standard protocol for data movement in the Grid, heterogeneity is still to reign in this environment. For example, files can be accessible through other protocols, such as HTTP [4], FTP [5], SCP/SSH [6], etc. Each protocol has its own authentication mechanism and proposes its own data interaction style (e.g., GUI, command-line, APIs). Due to this diversity, users are obliged to manually transfer files by using different tools supported (e.g., SmartFTP [7], PuTTY [8]) or writing scripts and programs (e.g., RSL [9] for Globus [10], JDL [11] for LCG2 [12]) to perform file staging. The data management task appears too complicated for non-technical users.

3. GRAVY FRAMEWORK

The architecture of GRAVY framework is presented in the Fig. 1. GRAVY framework is composed of five major components namely *Data Access Interface (DAI)*, *Grid File System (GFS)*, *Global Naming Manager (GNM)*, *Data Request Broker (DRB)*, and *Wrapper File System Interface (WFSI).*

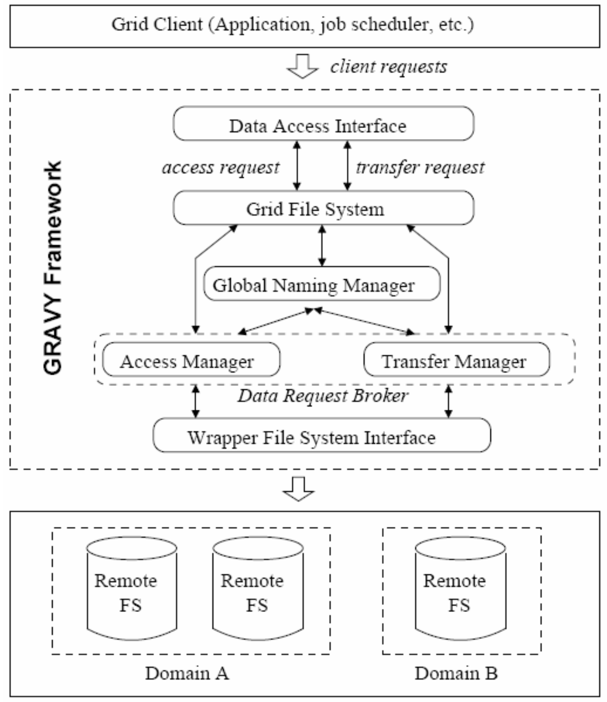


Fig.1. Architecture of GRAVY framework. The dashed rectangle is the core components of GRAVY. The rectangle below is underlying file systems belonging to multiple administrative domains, or virtual organizations.

3.1 Data Access Interface (DAI)

DAI is responsible for transforming client requests in a specific protocol used by the client to and from common requests understood by GFS. The fact that DAI supports multiple protocols allows GRAVY to be easily and flexibly deployed according to client needs. For example, HTTP support allows GRAVY to integrate easily into Grid web portals, FTP support allows clients to use popular FTP client tools to communicate with GRAVY, local access via APIs and Web-Services support allow GRAVY to integrate into applications and job scheduler for data movement control.

Besides local access, DAI currently supports three protocols: FTP [5], HTTP [4], and Web-Services. The implementation of FTP access is based on [13]. The protocol Web-Services is deployed using WSRF framework implemented in GT4 [14].

3.2 Grid File System (GFS)

GFS organizes a unified logical view of data above distributed heterogeneous file systems. Logical view of data consists of virtual directories corresponding to physical directories on remote file systems. It should be noted that virtual directories may not necessarily correspond to any physical data locations on remote file systems. This decoupling of data from its physical locations gives users the ability to create their own view of data grid.

GFS functionality is based on a grid file object *GridFile*. The fundamental purpose of *GridFile* object is to abstract file operations so that users can interact to data in different file systems without being concerned with the underlying protocol. *GridFile* provides standard interfaces for invoking data access and data transfer requests irrespective of protocols that underlying file systems support.

3.3 Global Naming Manager (GNM)

Storage administrators are able to configure the logical view of GFS to adapt to changes in policy and available data. GNM provides mechanism to specify the mapping from a virtual file reference to its location on remote file system in a XML file. In this way, heterogeneous file systems distributed across different sites are unified through a global namespace, which provides users with a single, logical view of distributed files, enabling intuitive access to data. As the physical location of the data is irrelevant to clients, they can access files without knowing their location.

In fact, GNM adapts the first level of indirection of the Resource Namespace Service (RNS), which is a specification of the Grid File System working group of the Global Grid Forum [15] to implement the global namespace for GFS. RNS is proposed to provide a naming mechanism to link existing data sources. RNS describes a three-tier naming architecture that consists of human interface names, logical reference names, and end-point references. Mapping from a human readable name to an actual data location can be realized in two levels of indirection. The first level is mapping human interface names directly to end-point references. The second level is realized by mapping human interface names to logical reference names (that may not be very readable by humans), which in turn map to end-point references.

3.4 Data Request Broker (DRB)

The primary goal of DRB is to handle requests received from GFS. DRB interprets user requests in generic file operation format provided by *GridFile* into the correct invocation using the protocols supported by the remote file system. DRB consults GNM to translate logical file name to physical file locations. After the actual file address is obtained, DRB invokes file operations in a specific protocol by interacting through WFSI. DRB is composed of *Access Manager* and *Transfer Manager*, which are responsible for executing access requests and transfers requests respectively.

Access Manager: generally, the execution time of access operations (e.g., directory creation, file rename) is very short (in the order of milliseconds), so the *Access Manager* is designed to execute access operations synchronously. The *Access Manager* converts generic access requests of GFS into specific protocol supported by the remote file systems and accomplishes it. Finally, it returns the result of execution to GFS.

Transfer Manager: transfer requests need to be treated differently from access requests, since transfer requests generally have long execution time and they can fail for a variety of reasons at anytime during the execution. They need to be monitored and rescheduled for restart in case of failure. Hence, the *Transfer Manager* is designed to execute transfer requests asynchronously. The *Transfer Manager* performs the actual movement of files from one remote file system to the other. In case of transfer failure due to dropped connections, machine reboots or temporary network outages, the *Transfer Manager* will restart the transfers at another time in order to ensure the successful completion of transfers.

The *Transfer Manager* uses the simple "first-come, first-served" principle to schedule file transfers. It performs inter-protocol transfers using a memory buffer or third party transfers whenever available. Inter-protocol transfers are performed when the transfers are done between file systems which don't support the same protocol. In this case, two consecutive data connections are used. The first connection performs transfer from the source file system to the memory buffer of GRAVY, and the second connection performs the transfer from the memory buffer of GRAVY to the destination file system.

3.5 Wrapper File System Interface (WFSI)

WFSI is a set of well-defined interfaces designed to make GRAVY framework completely modular. In grid environment, it is not expected that all file systems support the same transport protocol to communicate to each other. Thanks to WFSI, it is straightforward to add support to GRAVY for a particular transport protocol.

In the current version of GRAVY, we have already added support for different transport protocols, which are local access, FTP, GridFTP and SSH/SCP protocol. GRAVY can be used for data management with these protocols without any extra work. We have used client-side libraries provided in GT4 [14] to implement WFSI for FTP and GridFTP protocol, and JSch [16] for SSH protocol.

3.6 Security

Each transport protocol may use different security mechanisms and policies. Establishing the confidence across different protocols is challenging since each protocol has its own authentication mechanism and it enforces its own access control policy. We develop our own security packages based on the Grid Security Infrastructure (GSI) provided by Globus [17], which is utilized for authentication in GridFTP protocol. GSI avoids a centrally-managed security system and supports *single sign-on* for users of the Grid. In the whole system, there is a Certificate Authority (CA) signing certificates for users. With their own certificates, users can generate proxy certificates for authenticating to different administrative domain. For FTP protocol, authentication is performed through anonymous access. For SSH/SCP, we utilize password-less RSA/DSA public key authentication so it does not have the password typing overhead.

4. EVALUATION

We have implemented a prototype of GRAVY framework that we are proposing on Java 5.0. So, our prototype runs on any platform that supports the JVM 5.0. This section evaluates the performance of this prototype. Real experiments have been carried out on three Pentium 4 3.2 GHz machines with 512 MB of RAM, each running Linux with kernel 2.4.x. They are directly connected to 100 Mbps network adapter.

4.1 Inter-Operability Test

In order to test the GRAVY's feature of supporting multiple protocols, we perform transfers for a range of file sizes from 256KB to 1024MB between two machines. Each machine hosts a *ProFTP server*, a *GT4 GridFTP server* and a *sshd server*. GRAVY, which is deployed on the third machine, launches transfers between these two servers in GridFTP, FTP and SSH/SCP protocol respectively. We evaluate the bandwidth of these transfers and the results are shown in Fig. 2, each value is averaged from ten experiments.

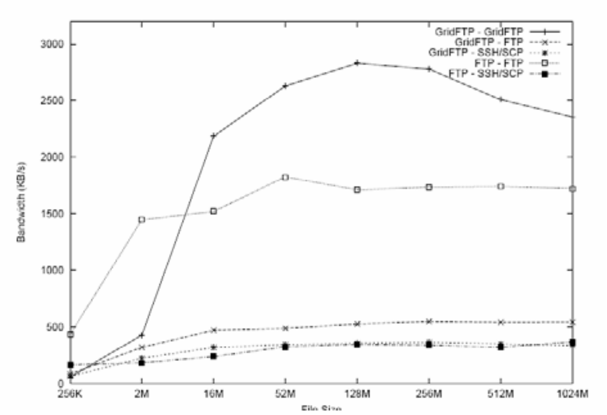


Fig 2. Transfer results in multiple protocols

We observe that the bandwidth varies a lot across each of the protocols. In general, third-party transfers show superior performance over other transfers. This is predictable as GRAVY has to pay the overhead for connecting two data streams for transfers in two different protocols (e.g., GridFTP and FTP). For small files (less than 2M in these experiments), the GridFTP and FTP third-party transfers don't have good performance due to the overhead of security processing (note that GridFTP uses X509 certificates and FTP has to exchange some messages to do anonymous authentication/authorization). As the file size increases, the performance of third-party transfers in GridFTP and FTP is improved clearly. The best performance of GridFTP is obtained around file size of 128MB - 256MB. The poor performance of SSH/SCP can be explained by the fact that SCP uses the symmetric key encryption/decryption and encrypts all the traffic on the channel.

4.2 Performance

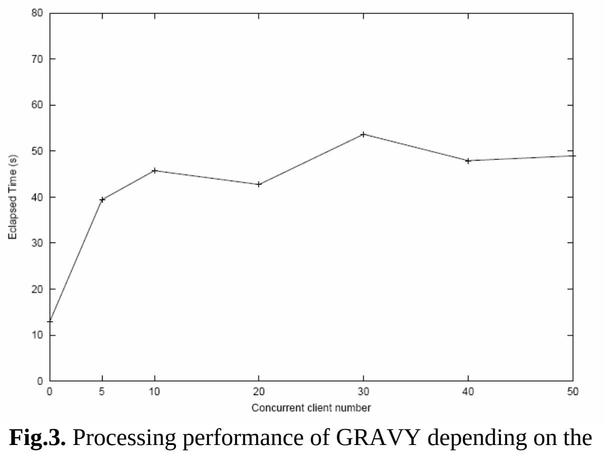


Fig.3. Processing performance of GRAVY depending on the number of clients concurrently transferring files

This experiment evaluates the impact of the number of concurrent transfers on the expected performance. We use GRAVY to launch several concurrent transfer processes. Our measurements were aimed at testing the stability and processing efficiency of GRAVY. The transfers were performed with files of 10MB. Fig. 3 shows the transfer's elapsed time depending on the number of concurrently connecting clients. Each value is an average of 5 tests. As the number of concurrent client increases, obtained elapsed time appears to increase linearly but its performance remains relatively stable.

5. RELATED WORK

The continual increasing requirement for sharing data in wide-scale scientific applications has lead to the development of multiple file systems, data access facilities and middlewares in grid environments in recent years. We present below some of these solutions.

Based on the basic Globus services [10], the DataGrid [18] is a large and complex project that defines a layered architecture of service components for transferring large datasets in heterogeneous environment. This architecture is similar to ours in the sense that both try to separate the physical location of data from its logical view, which is called metadata.

GT4 [14] provides a number of components for data management. These components fall into two basic categories: data movement, which is composed of GridFTP tools and Reliable File Transfer (RFT) service, and data replication, which consists of Replica Location Service (RLS). An important related component, OGSA-DAI [19], provides data access and integration capabilities to data resources, such as databases, within a Web-Service based framework.

Within the EGEE project [20], the data management system (DMS) [21] is composed of several components. The first is storage elements (SEs) which are the real element doing the storage of files. In the framework of the DMS, files are available through two namespaces: logical (Logical File Name - LFN) and physical (Storage File Name - SFN). The DMS is responsible for mapping an LFN to one or more SFNs. Other components of DMS are data catalogs that offer access to file replicas using LFN and data scheduler, which assures the availability of data at the chosen site for computation.

The Punch Virtual File System (PVFS) [22] allows standard NFS clients to connect to standard NFS servers by using NFS-forwarding proxies. PVFS enables a client executing on a computer server to access files stored within another security domain. However, it only supports NFS systems and is difficult to extend.

In [23], Garca-Carballeira et al. describe a global and parallel system for grids. This file system, which used the RNS specification [15] for offering a global namespace in the Grid, provides a generic distributed partition on various remote GridFTP servers.

6. CONCLUSIONS

In this paper, we have introduced GRAVY, a framework for data management in the Grid, which enables the inter-operability between heterogeneous file systems in the Grid. GRAVY framework integrates underlying heterogeneous file systems into a unified location-transparent file system of the Grid. This virtual file system provides to applications and users, a uniform global view and a uniform access through standards APIs and interfaces.

Since GRAVY is at the user-level and is overlayed on top of existing systems, it can be easily deployed on Grid nodes and integrated with high-level scheduler for handling data transfer between heterogeneous file systems for grid jobs. GRAVY supports multi-protocol and can be extended easily.

REFERENCES

[1] PPDG Deliverables to CMS. Available online at: http://www.ppdg.net/archives/ppdg/2001/doc00017.doc.

[2] Ian Foster, Carl Kesselman, and Steven Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International *Journal of High Performance Computing Applications,* 15(3):200–222, 2001.

[3] Bill Allcock, Joe Bester, John Bresnahan, Ann L. Chervenak, Ian Foster, Carl Kesselman, Sam Meder, Veronika Nefedova, Darcy Quesnel, and Steven Tuecke. Data Management and Transfer in High Performance Computational Grid Environments. *Parallel Computing Journal*, 28(5):749–771, May 2002.

[4] R. Fielding, U. Irvine, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. RFC-2068: Hypertext Transfer Protocol - HTTP/1.1, 1997. Available online at: http://www.w3.org/Protocols/rfc2068/rfc2068.

[5] Jon Postel and Joyce Reynolds. RFC-959: File Transfer Protocol. Available online at: http://www.w3.org/Protocols/rfc959/.

[6] T. Ylonen and C. Lonvick. RFC-4251: The Secure Shell (SSH) Protocol. Available online at: http://www.ietf.org/rfc/rfc4251.txt.

[7] SmartFTP. Available online at: http://www.smartftp.com.

[8] PuTTY. Available online at: http://www.putty.nl.

[9] The Globus Resource Specification Language (RSL), specification 1.0, 2000. Available online at: http://www.globus.org/toolkit/docs/2.4/gram/rslspec1.html.

[10] Globus Project. Available online at: http://www.globus.org.

[11] The EDG Job Description Language (JDL). Available online at: http://server11.infn.it/workload-grid/docs/DataGrid-01-TEN-0142-0%5F2.pdf.

[12] Large Hadron Collider Computing Grid Project (LCG), 2006. Available online at: http://lcg.web.cern.ch/LCG.

[13] Rana Bhattacharyya. Java FTP server. Available online at: http://www.myjavaserver.com/~ranab/ftp.

[14] Ian Foster. Globus Toolkit Version 4: Software for Service-Oriented Systems. In IFIP International Conference on Network and Parallel Computing, volume 3779 of *Lecture Notes in Computer Science, Springer-Verlag,* pages 2–13, 2005.

[15] Manuel Pereira, Osamu Tatebe, Leo Luan, and Ted Anderson. Resource Namespace Service specification, May 2006. Available online at: http://www.ggf.org/GGF17/materials/272/Resource_Namespace_Service_Refactored.pdf.

[16] JSCH - Java Secure Channel. Available online at: http://www.jcraft.com/jsch.

[17] Ian Foster, Carl Kesselman, Gene Tsudik, and Steven Tuecke. A Security Architecture for Computational Grids. In Proceedings of the *5th ACM Conference on Computer and Communications Security*, pages 83–92, San Francisco, California, USA, November 2-5 1998.

ACM Press.

[18] Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury, and Steven Tuecke. The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets. *Journal of Network and Computer Applications*, 23:187–200, 1999.

[19] Mario Antonioletti, Malcolm Atkinson, Rob Baxter, Andrew Borley, Neil P Chue Hong, Brians Collins, Neil Hardman, Ally Hume, Alan Knox, Mike Jackson, Amrey Krause, Simon Laws, James Magowan, Norman W. Paton, Dave Pearson, Tom Sugden, Paul Watson, and Martin Westhead. The Design and Implementation of Grid Database Services in OGSA-DAI. *Concurrency and Computation: Practice and Experience*, 17(2-4):357–376, February 2005.

[20] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. Enabling Grids for E-sciencE (EGEE), 2006. Available online at: http://www.eu-egee.org.

[21] Peter Kunszt and Paolo Badino. EGEE gLite User's Guide - Overview of gLite Data Management. *Technical Report* egee-tech-570643-v1.0, CERN, Geneva, Switzerland, 2005.

[22] Renato J. O. Figueiredo, Nirav H. Kapadia, and Jose A. B. Fortes. The PUNCH Virtual File System: Seamless Access to Decentralized Storage Services in a Computational Grid. In *Proceedings of the 10th IEEE International Symposium on High Performance Distributed Computing (HPDC'01)*, pages 334–344, San Francisco, CA, USA, August 2001. IEEE, IEEE Press.

[23] Flix Garca-Carballeira, Jess Carretero, Alejandro Caldern, J. Daniel Garca, and Luis. M. Sanchez. A Global and Parallel File System for Grids. *Future Generation Computer Systems 23*, 23(1):116–122, January 2007.

Grid Services for Research Computing at CUHK

Frank Ng[1], Sammy Tang[2]
Information Technology Services Center
The Chinese University of Hong Kong
[1]frank-ng@cuhkedu.hk, [2]sammy-tang@cuhk.edu.hk

ABSTRACT

Grid computing [1] is a tool which contributes to the generation of valuable results that cannot be produced in individual laboratories and to the success of advanced research studies. Thus, it significantly reduces the total cost and time needed for the research works. At the Chinese University of Hong Kong (CUHK), researchers in popular computational fields (e.g., computational physics and chemistry) normally demand huge computing power to conduct heavy computational tasks such as the first-principle calculation within deep earth[2]. To meet the demand for huge computing power, we built a computational grid[3].environment at CUHK to pool and integrate various centralized computing clusters and interdepartmental computing resources into a massive computing system with efficient and secure resources sharing in an on-demand fashion. Apart from meeting user demands, the next prime objective of the CU grid computing services is to facilitate the creation of knowledge researchers in grid technology and user community, which in turn, would be able to accelerate research and technology development in the University. Grid services with a robust, reliable, and efficient infrastructure can also be used to motivate an adoption of grid technology to increase the international competitiveness of the University. This would also help to simulate more research opportunities by using the grid as a platform for nationwide collaborations. We have successfully built the CU Grid environment, supported some major findings in a few research areas, and connected to major grid sites such as the Pacific Rim Applications and Grid Middleware Assembly (PRAGMA)[4].In this paper, we intend to share our experiences in building the CU Grid, our users' research areas which utilize the grid, their research findings, and the future development of the CU Grid.

Keywords: CU Grid, OGSA, OpenCA

1. INTRODUCTION

The Chinese University of Hong Kong (CUHK) supports major information technology (IT) services such as academic computing, administrative computing needs, and application development through the Information Technology Services Center (ITSC).

In the early 1990s, ITSC entered into a new era when it started providing Higher Performance Computing (HPC) services after decommissioning a large mainframe system.

That period also signalled a major ITSC shift from centralized mainframe systems to distributed computing services. The same period also saw ITSC's introduction of the first SGI Origin 2000 supercomputer system, the first of its kind in South-East Asia. The SGI Origin 2000 helped launch the pioneering Cache-Coherent Non-Uniform Memory Access (CC-NUMA) system for CUHK researchers.

For the next few years after the introduction of the SGI Origin 2000 system, ITSC also installed some SMP servers and an IBM RS/6000 SP system. However, coping with the HPC demands was still a difficult task.

With the success of the introduction of commodity processors and clustering technology, a Linux-based cluster was later built which helped to confirm the feasibility of deploying the distributed cluster systems and to save investment costs in providing an easy-to-use HPC environment with such processor and technology.

Recently, with the continuous users' demands and needs, the recent advancement on the grid technology, and the reduction on hardware costs, ITSC adapted the Grid computing technology as one more tool to keep up with the demand. The grid technologies were utilized to build an easy-to-scale and a high Return on investment (ROI) grid supercomputer system, namely, the ITSC Core Computational Grid[3]. This was done by integrating high-end dual-core Intel PCs, resources from interdepartmental computer laboratories, and ITSC central grids/clusters over the campus network.

2. DEMAND ON COMPUTING POWER

In the past decades, the computing power needs that were put forward by our researchers have jumped tenfold every three to four years. Moreover, it was always higher than those that could be provided by the previous system. As an example, in the first quarter of 2002, we launched a 100 GFLOPS computing cluster, but since that time, it was only in 2006 that we acquired the ability to facilitate the additional needs with a 1,100 GLOPS computational grid[3].

Among the different types of HPC systems available to the researchers, a computation-intensive system is the most-widely accepted by our researchers in the University. It is followed by a medium-size data storage system for data-intensive computing tasks.

As a comprehensive-based research university, we understand that computing power is a basic, fundamental, and important requirement for our researchers in carrying out high-quality research studies. Last year, the University announced a 10-year strategic plan[8]. that focused its research on five major areas: Bio-medical Sciences, Chinese Studies, Information Sciences, Earth Sciences, and Financial Sciences.

With the University's strategic plan in mind, ITSC is also preparing to scale-up the existing grid power by 5,000-10,000 GFLOPS (CU Grid) to 2.5 to 4TFLOPS in the coming two years. The CU Grid environment will serve as a major computational services environment for our researchers.

One example on the application of our CU Grid services and HPC would be the project entitled "Probing the Early Universe with Cosmic Microwave Background Anisotropies"[9], led by Prof. Chu Ming-Chung[c] of the Physics Department, CUHK.

The project involves building codes for carrying out multi-dimensional minimization and multi-parameter fittings on Cosmic Microwave Background Anisotropies (CMBA) data. The estimated computing resources needed for the projects are about 15,360 processor days on a CMBA analysis varying on fundamental 'constraints' and f® gravity runs. The project was supported by an RGC Earmarked Grant and the initial study result was passed through a scientific review.

3. NEW TECHNOLOGY ADOPTION

Our vision of grid services is similar to that of a power grid. That is, the grid services should be highly stable, reliable, and able to adapt new demands with regard to users needs, changes in requirements, and research interests or deployment of new technologies and applications.

The basic benefits of our grid services architecture is that once we have the designed grid infrastructure properly in place, our user will have access to a virtual computing environment that consists of diverse computing resources.

To reach this vision, we built our grid based on the common standards that would ensure a resulting infrastructure that is secure and robust. We therefore adapted the standards from the Open Grid Services Architecture[10] (OGSA), related OGSA Grid Security Infrastructure (GSI)[11] components, and a Globus ToolKit[12] (GTK Version 3.2.1 initially and 4 later on). The standards help us to build, verify, and provide the necessary framework and tools for connecting, deploying, monitoring, and managing the computing services.

4. IMPLEMENTATION OF CU GRID

Initially, ITSC built the grid services infrastructure in an intra-grid fashion and focused mainly on ITSC's own computing resources. That quickly formed the base of the initial ITSC Core Computational Grid[3] (e.g., specific purpose machines in central machines room) with limited targeted applications available (e.g., VASP, Amber, and NAMD).

Over time, with the initial groundwork, the foundation later built on the standards and the experiences gained previously. These initial ITSC Core Gird resources were integrated and interconnected with resources from other departments and a number of remote ITSC computing resources (e.g., IT Examination Center and remote user areas). The current CU Grid is currently converged on a 240 nodes grid services environment based on the OGSA[10] model.

The applications coverage was also expanded. Currently, the services environment contains a larger variety including the IMSL, MPIBlast, GotoBLAS, Cactus, Cosmomc, and Geant. The computing power also grew from the initial 0.5TFLOPS to the current 1.1TFLOPS. The processing power is expected to grow to 2.5TFLOPS in the next year.

Our grid topology is scalable. It is based on the gigabit network technology that uses mostly commonly available Cisco switches. The network can easily be expanded using inexpensive, low-port-count switches to accommodate additional servers for increased horizontal and overall performance on bandwidth and capacity. Our grid network is resilient to failure because the failure of nodes would be isolated easily by simply not delivering jobs into the specific nodes. This resilience means maintenance services and replacement of nodes or components has minimal impact on the availability of the grid services.

5. COLLABORATION WITH OTHER GRID SITES

With the existing grid services, we believe that our services could be further expanded to cover larger computing needs if we could be able to join forces with the world's leading research and grid efforts.

We are in the process of connecting to various testbeds of the Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) [4]. For example, the testing on a G/Farm with different types of applications has already begun and would be further explored. We are also studying the possibility of joining ChinaGrid for more opportunities to connect and work with other institutions.

Through the participation in PRAGMA and the collaboration with other workgroups, we would be able to expand our grid activities, including the integration of new applications, grid middleware development, cross-grid integration, and related services validation. Moreover, all these would finally help us to fine-tune our CU Grid to a more robust, easy, and simple to use grid service for the researcher and scientist in CUHK.

6. FUTURE WORKS

At the system level, there are a number of improvements that could be done in order to maintain the overall goals.

For example, the integration of OpenCA as a replacement of the SimpleCA (embedded in GT4) would be vital and would further be expanded to allow and support user delegation through the user's own certificate. The validation of user credential in this case would be done via the CU Grid (OpenCA) CA or CUHK existing CU CA.

We are also targeting the integration of a Web Services Architecture on application level according to the WS Resources Framework (WSRF) specification[14]. This would enrich the CU Grid services following the integration of the GT4's Web services. A UDDI server is on the way to support and house the related grid services on a design root on Service-Oriented Architecture (SOA)[15].

Furthermore, to test the WS services, we plan to develop a Multi-Stream JPEG Plotting Scheme on Large-Scale Data that would be based on our CU Grid facilities to generate accurate JPEG images for various related projects. The project would initially help to plot images from satellite data with a GS database and would be deployed into our WS server as a WS service.

7. RELATED RESEARCH RESULTS

For one example, the consensus sequences of human chromosomes are compared with the result of those found out by a common program (e.g., PILER, Repeatscout, ReAS). The 24M and 64M human genome sequences with an overlapping length of 50, 75, and 100 and a similarity of 70%, 80%, and 95% were used to generate the results. It has been tested to obtain a result with time reduction to 1-1.5 days for the 24M file and slightly less than 4 days for the 64M file.

8. CONCLUSIONS

Grid computing proves to be a robust and versatile tool. In addition, it quickly responds to the demand for large computing power in a research-based university. With a well-structured design in systems, it has proven that the grid middleware and its related tools are mature and scalable.

We have illustrated that it can be further enhanced to support 1,000 Intel-based (or AMD) systems[3] by integrating central grids/clusters and interdepartmental computer laboratories. Consequently, the resulting system would have a computing power between 6 to 20 TFLOPS[5]. Such a configuration would still be a low-cost but effective approach and would not cause any financial pressure to the Grid services provider and/or users.

As the development of Grid Toolkits (GT) are still in progress, the development of grid middleware components such as WS services, Grid RPC, Globus XIO, GridFTP[6], etc. can be put to support the massive data movement over the next-generation campus network[7].

We anticipate that a data grid would be required to deal with projects that are focused on the areas of earth and bio-medical sciences as projects are underway and initial testing is being carried out in our grid environment.

ACKNOWLEDGEMENT

The authors would like to express their thanks to the Steering Committee of DCABES 2007 Conference for the acceptance of their paper. Furthermore, the CUHK ITSC HPC support team would also like to express their appreciation to the researchers and postgraduates who continuously worked with the team and encouraged the team to produce high-quality HPC services under a limited manpower structure and tight budget scheme.

REFERENCES

[1] Ian Foster, “The Anatomy of the Grid: Enabling Scalable Virtual Organizations,” *IJSA*, 2001

[2] Parallel Computing and Environment in the Chinese University of Hong Kong, *International Algorithms and Computing Environment, ICPACE* 2003.

[3] Internal Document: “An Implementation Plan of Computational Grid,” 2003.

[4] “Pacific Rim Applications and Grid Middleware Assembly,” *PRAGMA* 2003 http://www.pragma-grid.net/Brochure

[5] Intel Woodcrest: “An Evaluation for Scientific Computing,” Philip C. Roth and Jeffrey S.Vetter, C*omputational Sciences and Mathematics Division*, Oak Ridge National Laboratory, One Bethel Valley Road, Oak Ridge, TN 37831.

[6] “Effective Use of Multi-Core Commodity Systems in HPCKent Milfield,” Kazushige Goto, Avi Purkayastha, Chona Guiang and Karl Schulz, Texas Advanced Computing Center, The University of Texas at Austin.

[7] Globus Toolkit Version 4: *Software for Service-Oriented Systems*, Ian Foster, Math & Computer Science Division, Argonne, National Lab, Argonne, IL 60439, U.S.A. Department of Computer Science, University of Chicago, Chicago, IL 60637, U.S.A.

[8] “Experiences Deploying a 10 Gigabit Ethernet Computing Environment to Support Regional Computational Science,” Jason Cope, Theron Voran, Matthew, Woitaszek, Adam Boggs, Sean McCreary, Michael Oberg, and Henry M. Tufo, University of Colorado, Boulder, CO, National Center for Atomspheric Research, Boulder, CO.

[9] “Probing the Early Universe with Cosmic Microwave Background Anisotropies,” Prof. Chu Ming-Chung[c], Physics Department, Departmental Project, CUHK.

[10] “The Physiology of the Grid, An Open Grid Services Architecture for Distributed Systems Integration,” *OGSA* Ian Foster, Carl Kesselman, Jeffrey M. Nick and Steven Tuecke

[11] *Open Grid Services Infrastructure*, version 1.0, GGF 2003

[12] *GLOBUS*, Toolkit version 3.2.1 and version 4, Globus Alliance.

[13] WSRF, *Web Services Resource Framework*, V1.2 Specification, *OASIS* 2006.

[14] SOA, OASIS (Organization for the Advancement of Structured Information Standards), SOA reference model, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=soa-rm

A Study of COTS Middleware Reuse and Integration in C2 Architectural Style

Jubo Luo[1,2], Wei Liu[3],Junfeng Yao[1] ,Xiaojian Li[1],Dan Xie[1]
[1](State Key Laboratory of Software Engineering (Wuhan University), Wuhan 430072, China)
[2](School of Management, Wuhan University of Science and Technology, Wuhan, 430081, China)
[3](School of Computer, Wuhan University of Technology, Wuhan, 430063, China)
Email: whuluocheng@126.com

ABSTRACT

With the rapid development of software development technologies using model and reuse,dependence of COTS(Commercial-off-the-shelf) in software development has increased considerably. To manage some risks in using COTS components, it is necessary to increase the reusability of the code so that the problematic or failural COTS components can easily be replaced by other components. Successful component reuse and integration depends on qualities of the components reused and the software context. Architectural styles have the potential for providing structure for offthe-shelf (OTS) component reuse. We propose an approach to reuse and Integration of DCOM components based on the C2 style. An On-Line Book Sales System Case Study was performed to confirm that the C2 style offers significant reuse and integration potential to application developers.

Keywords: COTS Components, Reuse, Integration, C2 Architectural Style, DCOM Component

1. INTRODUCTION

Commercial-off-the-shelf (COTS) software is developed by a third party. Usually a commercial vendor, and intended to be put of a new softwm system. Developed by professionals in the sea, COTS software can possess high quality and provide very sophisticated packaged functionality. COTS products reuse can help software developers to reduce the development effort and increase the product's quality [1],[3],[4],[5]. Other benefits are quick feasibility of demonstrations and support of COTS products by their vendors [2]. These benefits of COTS reuse make it an important issue in software engineering.

Apart from the generd reuse and integration problems such as selection, integration,maintenance. COTS products are plagued by their own specific problems: A COTS product may not be compatible with in-house software or other COTS products.Usually the source code of COTS software is not provided. So it cannot be modified.Different versions of the same COTS product may not be compatible,causing more problems tbr developers[13].

In this paper we are mainly concerned with COTS components reuse and Integration in C2 architectural style.In section 3, we present the results of how well “off-the-shelf” components could be reused in applications designed in accordance with the C2 software architectural style. In section 4,We propose an approach to reuse and Integration of DCOM components based on the C2 style. An On-Line Book Sales System Case Study was performed to confirm that the C2 style offers significant reuse and integration potential to application developers.

2. BACKGROUND

We can now consider specific COTS integration methods that have been developed and applied. The first two integration methods use specific software architectures that are more suitable for COTS integration than the conventional ones,while the third integration method is based on a special COTS-oriented life-cycle.In the first integration approach, it is suggested [6] that all components must be wrapped so that all interactions rule performed only through the wrappers .

Another type of architecture suitable for COTS integration is C2 [7], which is a component- and message-based architectural style. C2 allows the use of heterogeneous components with their internal architecture. It has asynchronous message passing and makes no assumption on shared address space, or a single thread of control.These features allow reusing COTS products with different characteristics. Research has been conducted on using different off-the-shelf middleware in C2 architecture [11] . However, C2 uses a layered style, i.e., components of upper layers can send messages only to components of lower layers, thus limiting the number of COTS that can be used within this architectural style.

3. OTS COMPONENT REUSE

In this section we will discusses the reuse of OTS components in C2-style architectures.

3.1 C2’s Suitability for OTS Component Reuse

C2 is architecture of highly-distribute evolvable and dynamic system.

A C2 architecture is a hierarchical network of concurrent components linked together by connectors in accordance with a set of style rules [7]. C2 can allow a variable number of components to be attached to a connector. The rule of C2 is relatively simple, because every component just responds to the component above themselves and each component have at most two communication links.

The C2 architectural style is a component-based style that supports largegrain reuse and flexible system composition, emphasizing weak bindings between components. Components communicate only through asynchronous messages mediated by connectors. Both components and connectors have a top interface and a bottom interface [8].

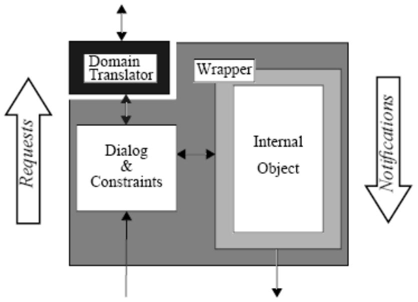


Fig.1. The Internal Architecture of a C2 Component.

Systems are composed in a layered style, where the top interface of a component may be connected to the bottom interface of a connector and its bottom interface may be connected to the top interface of another connector.Each side of a connector may be connected to any number of components or connectors. Requests from client components flow up through the system layers and responses from server components, called notifications, flow down (Fig.1)[8].

4. AN ON-LINE BOOK SALES SYSTEM CASE STUDY

Currently available component infrastructure frameworks such as CORBA，DCOM and Enterprise JavaBeans，support distributed applications by providing a transparent communication mechanism between components across hosts.

Dynamically configured distributed applications rely on the underlying framework to determine each component’s location. In DCOM and CORBA, this type of distribution is achieved through the use of trader services. A trader accepts requests for component instances, servicing the requests with references toexisting components, or by creating new components in the system [16].

Compose various DCOM components provided by third-parties and reconfigure flexibly an architecture through the plug-and-play technique,and compose DCOM components without modification through the plug-and play technique based on the C2 style[15].

4.1 DCOM wrapper for DCOM Composition and Reuse

Components usually consist of multiple objects and thus have a larger granularity. This characteristic enables them to combine functionalities of the objects as a single service.The role of DCOM wrappers are as follows:To communicate with components By the logic of message passing instead of the logic of method invocation.The logic of message passing enables to send a message to the target component which is not explicitly specified.We give the representative form of the message handling logic as following code:

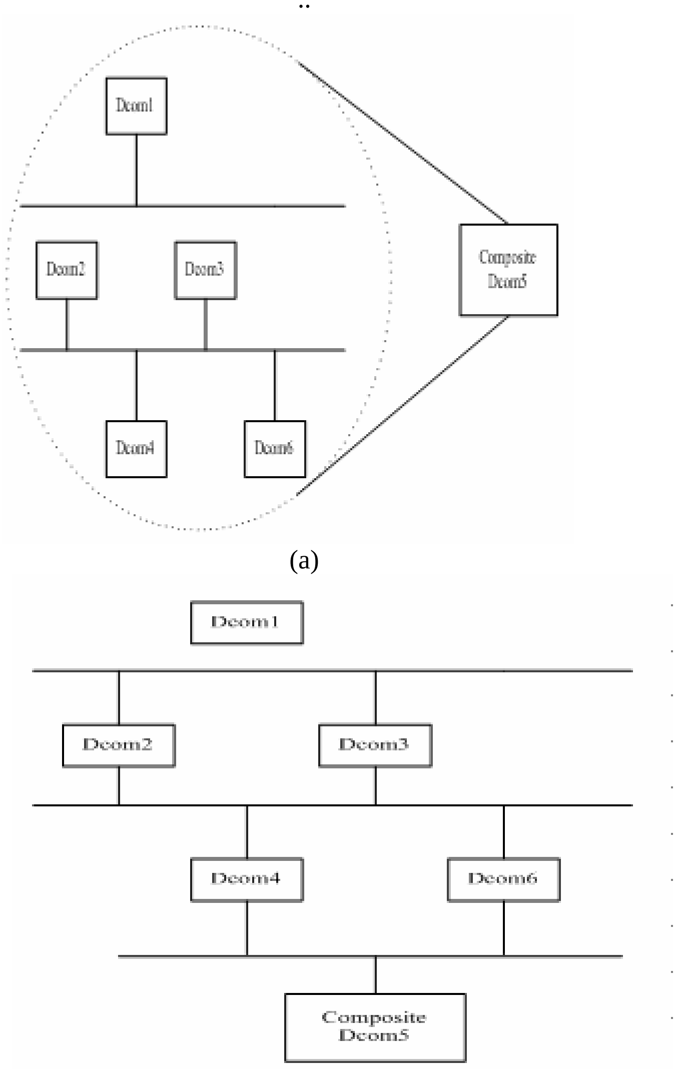
```
While(received_message is true)
   {//upload parameters
..........
//invoke functions of DCOM
..........
//create notification/request
..........
//send notification/request
..........
   }
```

4.2 Generating a Composite DCOM Component

Distributed component technologies combine the characteristics of components with the functionality of middleware systems to provide inter-process communication between components. That is to say components that can communicate across machine boundaries.

In Fig.2 (a)(b),we propose a conceptual framework of generating a composite DCOM component. The relationship of the C2 architecture which is composed of various DCOM components and a composite DCOM. A new C2 component which links the C2 architecture with the composite DCOM is needed. [12].

..



(a)

(b)

Fig.2.(a)(b) generating a composite DCOM component framework

The representative form of the message handling logic of the new C2 component is illustrated as the following code:

```
Method_remote(){..........
                    R new_request=new R(“request_message”);
                    ..........
                    Send (new_request);
                    Return message_values;
                    }

Void processing (Message m) {//processing message
                                ..........
                               }
```

4.3 DCOM integration and Reuse in C2 style Architecture of an On-Line Book Sales System

Middleware is responsible for providing transparency layers that deal with distributed systems complexities such as location of objects, heterogeneous hardware/software platforms and numerous object implementation programming languages.

A simple example using the component-assembly tool to assembly DCOM components through the plug-and-play technique.The example system is an On-Line Book Sales system. In Fig.3, we give the C2 style architecture of an On-Line Book Sales that is composed of 7 DCOMcomponents:postoffice,enterprise,member,payment,books,houset,salesorder[12].

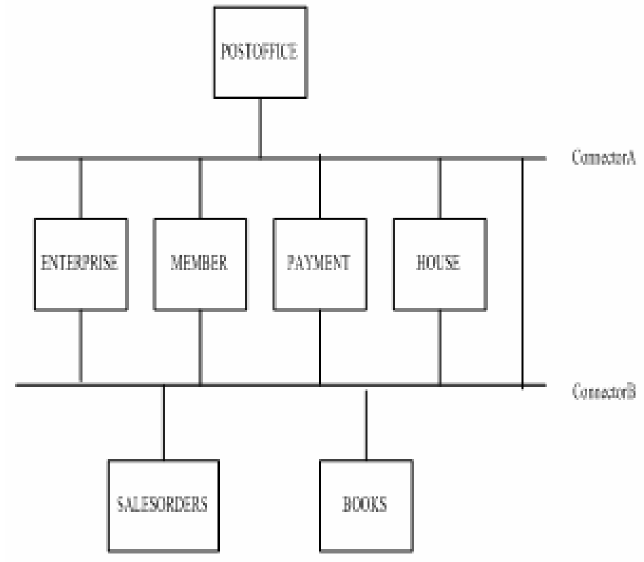


Fig.3. C2 style architecture of an On-Line Book Sales

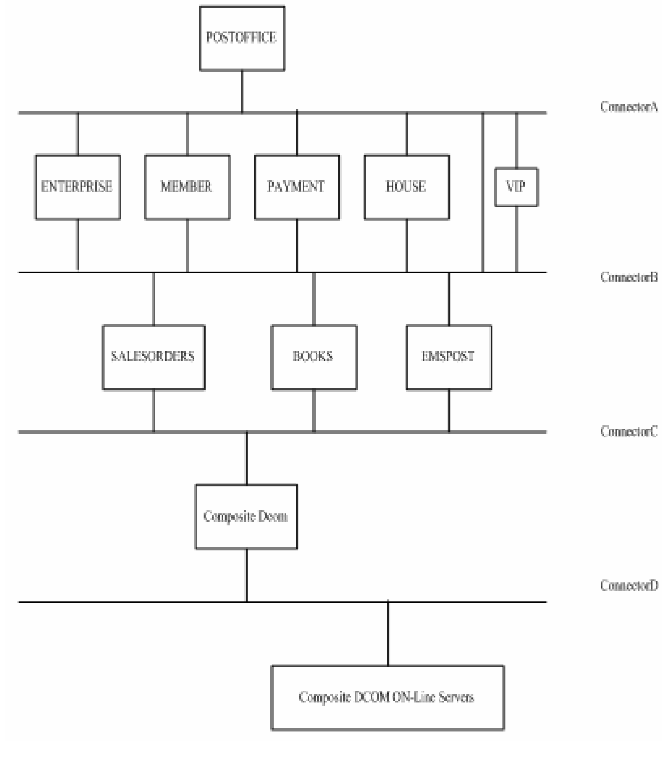


Fig.4. Integrating a DCOM component Post and a VIP component into the On-Line Book Sales system

In fig.4,we integrating a DCOM component EMSPOST and a VIP component into the On-Line Book Sales system without modification through the plug-and-play technique based on the C2 style. And create a composite DCOM component and a composite DCOM servers component for a C2 architecture which is provided by third parties through the plug-and-play technique.

5. CONCLUSIONS AND FUTURE WORK

In this paper we are mainly concerned with COTS components reuse and Integration in C2 architectural style. We present the results of how well “off-the-shelf” components could be reused in applications designed in accordance with the C2 software architectural style. We propose an approach to reuse and integration of DCOM components based on the C2 style. An On-Line Book Sales System Case Study was performed to confirm that the C2 style offers significant reuse and integration potential to application developers.

The explicit and flexible nature of C2 connectors directly contributes to our ability to implement distribution, mobility, and runtime structural change. the C2-style does not, by itself, guarantee that a change will leave the application in a consistent state.A C2 architecture is a hierarchical network of concurrent components linked together by connectors in accordance with a set of style rules.The rules are strong enough to make reuse tractable but broad enough to enable integration of OTS components, is a key issue in formulating and adopting architectural styles.

Architectures may be required to describe software systems at different levels of detail, where complex behaviors are either explicitly represented or abstracted away into single components and connectors.
The integration problems classification is should be further developed..Our future research can target at devetopment of the COTS integration process embracing integration，selection, maintenance,security.The proposed COTS integration approach should be applicable in a real development environment.

REFERENCES

[1] Gentleman, W.M., "Effective Use of COTS (Commercid-Off-the-Shelf)Software Components in Long Lived Systems", *The Proceedings of ICSE 97*, Boston, USA, 1997, pp. 635-636.

[2] Sreeram Bayana, “Learning to deal with COTS(Commercial off the Shelf) College of Engineering and Mineral Resources at West Virginia University,” Master’s thesis,2005,pp40-46.

[3] Fox. G., Lmtner. K., Marcom. S., "A Software Development Process for COTS-based information System Infrastructure". The Twenty-second Software Engineering Workshop. NXSNGoddard Space Flight Center Software EngineeringLaboratory (SEL). Greenbelt, MD, December 1997, pp. 133-153.

[4] Voas, J.M. “The Challenges of Using COTS Software in Component-Based Development,” *Computer*. June 1998, pp. 44-45.

[5] Voas, J.M. "Certifying Off-the-shelf Software Components", Computer, June 1998. pp. 53-59.

[6] Vidger. M.R. Dean, J., "An Architectural Approach to Building Systems from COTS Software Components", The Twenty-second Software Engineering Workshop. NASAIGoddard Space Flight Center Software Engineering Laboratory (SEL), Greenbelt, MD, December 1997, pp. 99- 131.

[7] Medvidovic. N..Oreizy, P., Taylor, R.N., “Reuse of

Off-the-shelf Components in C2-Style Architectures," *Proceedings of the 1997 Symposium on Software Reusability (SSR'97)*. Boston, USA. May, 1997. pp. 190-198.

[8] Nenad Medvidovic, "Architecture-Based Specification-Time Software Evolution," Dissertation of PH.D.1999, University of California, pp.69-75.

[9] K. Brockschmidt. *Inside OLE 2*. Microsoft Press, 1994.

[10] Davis. M.J. Williams. R.B. "Software Architecture Characterization," *Proceedings of the 1997 Symposium on Software Reusability (SSR'97)*. Boston, USA, May, 1997. pp. 30-38.

[11] Dashofy. E. Medvidovic, N. Taylor, R.N., "Using Off-The-Shelf Middleware to Implement Connectors in Distributed Software Architectures". *Proceedings of the 21' International Conference on Software Engineering*, Los Angeles.

[12] http://www.etri.re.kr, "Electronics and Telecommunications Research Institute.An Approach to Composition of EJB Components Using C2 style," 2005.

[13] Daniil Yakimovich, "A Comprehensive reuse model for COTS software products," Dissertation of PH.D, University of Maryland, 2001.

[14] W. Lam and V. Shankararaman, "An Enterprise Integration Methodology", *IT Professional*, Volume: 6, Issue: 2, IEEE Computer Society Press, 2004, pp. 44-48.

[15] Woods, E.: "Experiences Using Viewpoints for Information Systems Architecture: An Industrial Experience Report," F. Oquendo (Ed) *Proceedings of the First European Workshop on Software Architecture*, LNCS 3047, Springer Verlag, 2004.

[16] Michael Richmond, "Component Migration with Enterprise JavaBeans," *ACM*, 2000, pp.3.

Jubo Luo is presently a PhD student in the State-Key Lab.of Software Engineering, WuHan University, Wuhan, China. His research interests in software engineering include component based software development, software component software architecture, software reuse and web services. Mailing address: State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China.

A Novel Approach to Remote File Management in Grid Environments*

Haili Xiao, Hong Wu, Xuebin Chi
Supercomputing Center, Computer Network Information Center
Chinese Academy of Sciences, Beijing 100080, China
Email: haili@sccas.cn , wh@sccas.cn, chi@sccas.cn

ABSTRACT

File management is usually one of the most important services in Grid. In this paper, a novel and convenient approach to managing remote files and directories in a Grid environment will be introduced. Adding a new driver to the client machine, ordinary file operations, such as creating new folders, copying and removing, as well as drag'n'dropping files, are done locally inside the file manager which is bundled with the operating systems. In contrast with file services provided by Grid portals, it is more convenient and efficient for Grid users and more features are supported.

Keywords: Grid,File Management,Remote File

1. INTRODUCTION

Web-based Grid portals are widely used in Grid environments nowadays. A Grid portal provides a single web interface to all the computing, storage and other resources distributed in the entire Grid [7, 8, 32]. Projects like the GridPort Toolkit [25, 30], the GridSphere portal framework [18, 24, 31] and the GENIUS Grid portal [2,29] supported by EGEE [26] enable developers to quickly develop and package web applications.

Almost all these Grid portals, especially computing Grid portals, provide a file service component or alike, which is also one of the most important services[9–12, 16]. The file management functionality enables execution of file-related tasks on the underlying Grid systems. With GridFTP being operational on the remote server, file management operations allow users to upload and download files to/from remote systems. [20, 22, 25] As part of a portal, operations are all limited within the web browser.

In the following paragraphs, a novel approach to managing remote files and directories in a Grid environment will be introduced. Actually, it can be considered as adding a new driver to the client machine. And all the data of the new driver are retrieved from remote servers via GSI-enabled SSH connections. File operations are not done within a web browser, but locally within the file manager which is bundled with the operating system. Consequently, it is more convenient and more features are supported.

First, an overview of adding a new SSH remote filesystem is introduced. Then, details of design and implementation of file management in a Grid environment are given. Finally, conclusions of this approach and its features are made, and future work is also included.

2. OVERVIEW

To add a new SSH remote filesystem, two parts should be included: retrieving remote file information and integrating them into the operating system as part of filesystem. As for retrieving, file information is fetched from remote servers by SSH client tools or other wrapped applications. In Grid environments, GSI-enabled OpenSSH [33] adds GSI (Grid Security Infrastructure) to the list of available authentication mechanisms for SSH protocol 2. Since GSI-enabled SSH is deployed in almost all Grids, service of file management can be build upon it. It is out of question. On the other hand, however, for integrating there are different ways for different systems.

2.1 Filesystem As A Kernel Module

For GNU/Linux systems, many kernel modules are developed to add new features. Currently SSH remote filesystem is implemented as a kernel module, and file information is retrieved from remote servers by SSH client commands or other wrapped applications.

The FUSE [27] kernel module is developed to support new filesystem. With USE it is possible to implement a fully functional filesystem in a userspace program. The FUSE kernel module and the FUSE library communicate via a special file descriptor which is obtained by opening /dev/fuse. This file can be opened multiple times, and the obtained file descriptor is passed to the mount syscall, to match up the descriptor with the mounted filesystem. It is the best userspace filesystem framework for linux.

SSH Filesystem [28] is a filesystem client based on the SSH File Transfer Protocol. On the server side there's nothing to do. On the client side mounting the filesystem is as easy as logging into the server with SSH. It is based on FUSE above mentioned and supports multithread plus directory contents cache.

Gmail Filesystem [36] is a Python application which provides a mountable Linux filesystem using Gmail account as its storage medium. It also uses the FUSE infrastructure to help provide the filesystem, and libgmail [34] to communicate with Gmail. Most common unix commands are supported to operate on files stored on Gmail (e.g. cp, ls, mv, rm, etc.).

Shfs (secure SHell FileSystem) [37] is a simple and easy to use Linux kernel module which mounts remote filesystems using a plain ssh connection. Shfs sources include a GNU/Linux kernel module and several userspace utilities (to mount or unmount remote filesystems).

All these filesystems are based on a Linux kernel module (FUSE or SHFS).Fig.1 illustrates the structure of SSH remote filesystem under Linux.

*This work is supported by the 863 Program "Research of Management Technology in China National Grid" (2006AA01A117), the 863 Program "Environment of supercomputing service face to scientific research" (2006AA01A116), the Ministry of Science and Technology of China "Construction of the network computing application system of astronomy bioinformatics and computational chemistry" (2004DKA640002), the 973 program "Research on high performance scientific computation" (2005CB321702).

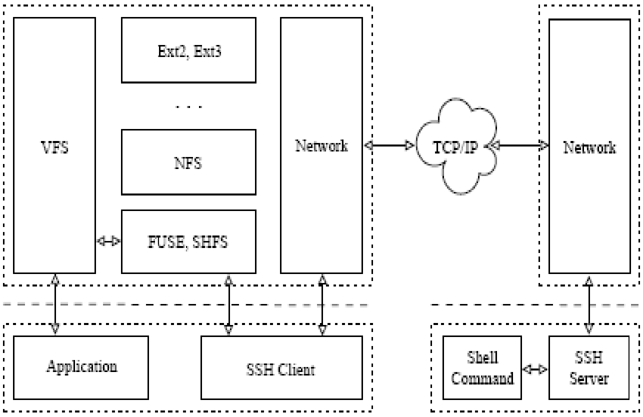


Fig.1. SSH remote file system under Linux

From the structure (Fig.2) of Windows filesystem, kernel module can be inserted at VFAT/NTFS layer, which is one level above the I/O subsystem layer. [15] However, New kernel module for the Windows system maybe not a good choice when taking into account stability.

2.2 Filesystem As A Shell Extension

In stead of kernel module, for Microsoft Windows system, Shell namespace extension is an alternative to adding a remote filesystem to the system. For example, GMail Drive [40] is a virtual filesystem to use Gmail as a storage medium. It literally adds a new drive under the My Computer folder, where folders are created, copied or drag'n'droped to. Saving and retrieving files stored on Gmail account are directly from inside Windows Explorer. GMail Drive is a Shell namespace extension.

2.3 Windows Shell And Shell Namespace

The Windows Shell manages and provides access to the wide variety of objects, ie. folders and files that reside on computer disk drives. the Shell also manages a number of non-filesystem, or virtual objects, which include Network printers, Other networked computers, Control Panel applications, The Recycle Bin. These virtual objects may even be located on remote computers. [5, 6]

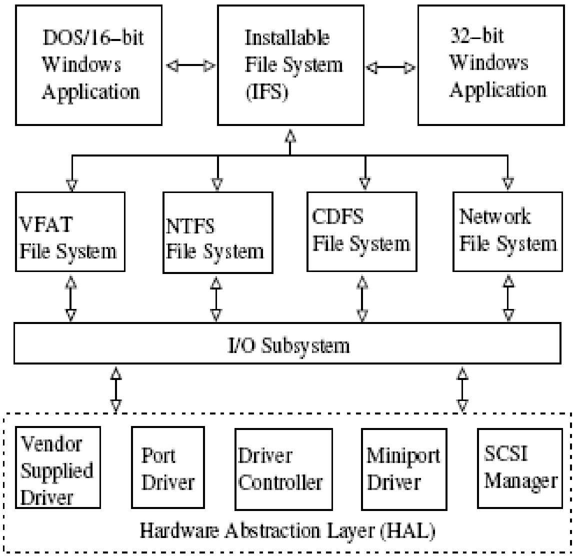


Fig.2. Windows file system structure.

The Windows Shell namespace is a larger and more inclusive version of the file system. It organizes the file system and other objects into a single tree-structured hierarchy. The ultimate root of the namespace hierarchy is the desktop. Immediately below the root are several virtual folders such as My Computer and the Recycle Bin. My Computer is the root of various disk drives. We can add our virtual folders below the desktop, My Computer or elsewhere. [1, 23]

3. DESIGN AND IMPLEMENTATION

3.1 Design Diagram

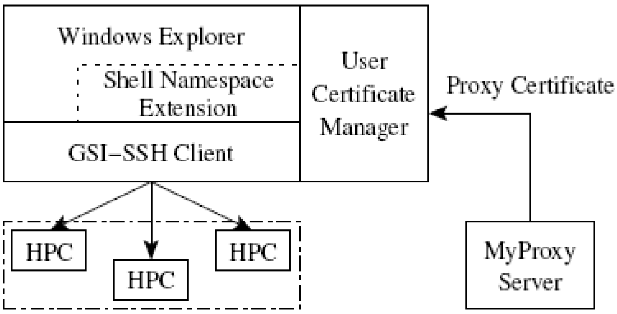
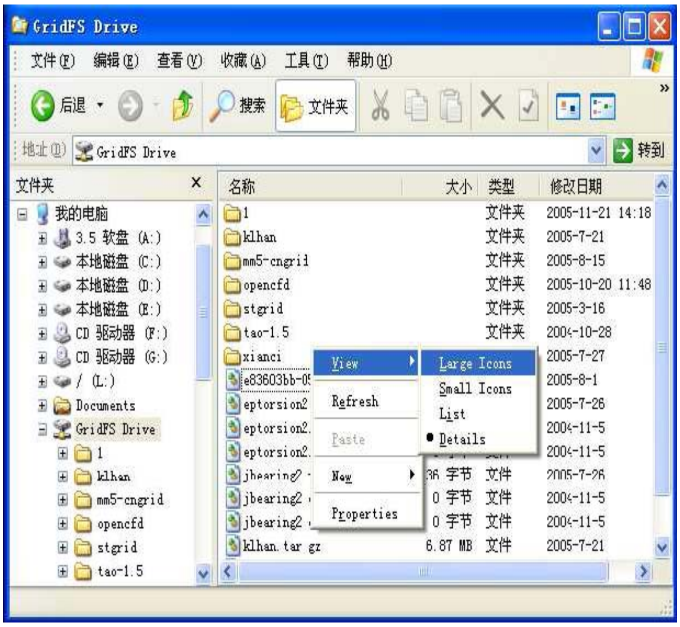


Fig.3. Remote file management in Grid.

From the above analysis, Fig.3 is chosen as the design diagram. First, the user needs to acquire a delegation or proxy certificate from MyProxy server before any further operations. When the user wants to manage remote files, he is authenticated by remote server with his certificate. If successful, data of file information are fetched from server by GSI-enabled SSH client, which is then integrated into the operating system as an extension to the Shell namespace. From the user's point of view, there is no difference between managing local files and remote one. There is a snapshot in Fig.4.

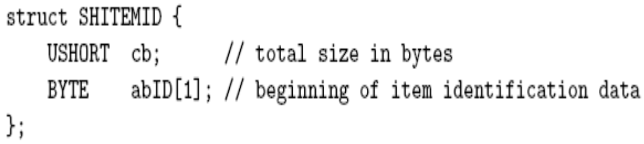


名称	大小	类型	修改日期
1		文件夹	2005-11-21 14:18
klhan		文件夹	2005-7-21
mm5-cngrid		文件夹	2005-8-15
opencfd		文件夹	2005-10-20 11:48
stgrid		文件夹	2005-3-16
tao-1.5		文件夹	2004-10-28
xianci			2005-7-27
e83603bb-05			2005-8-1
eptorsion2			2005-7-26
eptorsion2.			2004-11-5
eptorsion2.			2004-11-5
jhearing2	36 字节	文件	2005-7-26
jbearing2	0 字节	文件	2004-11-5
jbearing2	0 字节	文件	2004-11-5
klhan.tar gz	6.87 MB	文件	2005-7-21

Fig.4. Snapshot with a context menu.

3.2 Important Data Structures

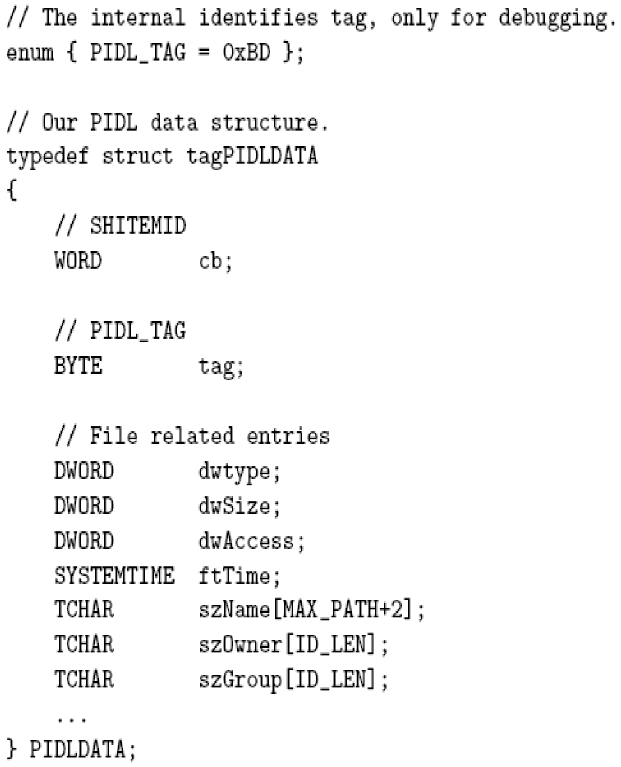
The Windows Shell tracks every object in its namespace by associating an item identifier with it. These are binary identifiers that consist of one or more SHITEMID structures (Fig.5):



```
struct SHITEMID {
    USHORT  cb;       // total size in bytes
    BYTE    abID[1]; // beginning of item identification data
};
```

Fig.5. SHITEMID structure.

Since the parent of an item has its own item identifiers, any item can be uniquely identified by a list of item IDs, called an ID list (ITEMIDLIST). A pointer to an ITEMIDLIST is called a PIDL, and is used extensively in namespace extensions [3, 14]. ITEMIDLIST is a variable-length structure. Only the first two bytes are defined, which contain the size of the identifier list. The rest of the identifier list contains data that is implementation-specific. Fig.6 lists the PIDL data in our example.



```
// The internal identifies tag, only for debugging.
enum { PIDL_TAG = OxBD };

// Our PIDL data structure.
typedef struct tagPIDLDATA
{
    // SHITEMID
    WORD        cb;

    // PIDL_TAG
    BYTE        tag;

    // File related entries
    DWORD       dwtype;
    DWORD       dwSize;
    DWORD       dwAccess;
    SYSTEMTIME  ftTime;
    TCHAR       szName[MAX_PATH+2];
    TCHAR       szOwner[ID_LEN];
    TCHAR       szGroup[ID_LEN];
    ...
} PIDLDATA;
```

Fig.6. Remote file information in the PIDLDATA structure.

This structure begins with the cb member which holds the total length in bytes (cb == sizeof(PIDLDATA)), followed by the real file information. These are data about file type, size, name and modification date, etc.

4. CONCLUSIONS

File management is one of the most important services in a Grid, especially a computing Grid. Traditionally, Grid portals enable execution of file-related tasks on the underlying Grid systems.

In this paper, a novel and convenient approach to remote file management in a Grid environment is introduced. Adding a new driver to the client machine, ordinary file operations are done locally inside the file manager which is bundled with the operating systems. In contrast with file services provided by Grid portals, it is more convenient and efficient for Grid users. It also has been tested to be quite stable when used in the CNGrid (China National Grid) [38] project and the ScGrid (Scientific Computing Grid) [39] project.

The current version is a single thread implementation, so it is not suitable for transferring large files. For the time being, remote to remote is not supported neither. And performance can still be improved by supporting directory contents cache. Besides, more features can be included in the future, the most exciting part of which would be submitting jobs and checking jobs, not only managing files, inside the file manager locally. That could be a completely new and wonderful experience for all Grid users.

5. ACKNOWLEDGEMENTS

We'd like to thank the National Development and Reform Commission, the Ministry of Science and Technology of China, the 973 program, and the National Natural Science Foundation of China for their financial support. And we also thank the collaborators and users from the CNGrid project and the Sc-Grid project. Help from colleagues of the Supercomputing Center is very much appreciated.

REFERENCES

[1] *David Campbell Extending the Windows Explorer with Name Space Extensions Microsoft Systems Journal*,Jul, 1996.

[2] Roberto Barbera, GENIUS and GILDA, *GENIUS/GILDA Tutorial*, May,2004.

[3] Nikos Bozinis, *Shell Explorer's Cookbook*,2004.

[4] Markus Friedl, Niels Provos, William A. Simpson, “Diffie-Hellman Group Exchangefor the SSH Transport Layer Protocol,” *Internet-Draft, IETF Network Working Group*,Jul 2003.

[5] Michael Dunn, “The Complete Idiot's Guide to Writing Namespace Shell Extensions,” *The Code Project,* Dec 2001.

[6] Michael Dunn, “The Complete Idiot's Guide to Writing Shell Extensions,” *The Code Project,* Jun 2002.

[7] I. Foster, C. Kesselman, “The Grid: Blueprint for a New Computing Infrastructure,” Morgan Kaufmann, 1998.

[8] I. Foster, C. Kesselman and J. Nick et al., “The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration,” January, 2002.

[9] D. Gannon, G. Fox, and M. Pierce, et al.,” Grid Portals: A Scientist's Access Point for Grid Services,” *Global Grid Forum,* 2003.

[10] Dennis Gannon, Oliver Wehrens, “Grid Portals C A Gateway to Grid Applications,” *Global Grid Forum* 13, 2005.

[11] gLite Installation and Configuration Guide, v. 1.5 (rev.4), Members of the EGEE Collaboration, February 07, 2006.

[12] Gregor von Laszewski, Beulah Alunkal, Kaizar Amin, et al, *The Java CoG Kit User Manual Draft Version* 1.1a Mar 14, 2003.

[13] S. Meder, V. Welch, S. Tuecke, D. Engert, GSS-API Extensions, *Grid Security Infrastructure (GSI) Working Group of the Global Grid Forum,* Feb 2001, Revised May 2003.

[14] Krishna Kotipalli, Shell Namespace Extensions, *MSDN Magazine*, Mar 1999.

[15] Rajeev Nagar, *Windows NT File System Internals - A Developer's Guide*, O'Reilly, September, 1997.

[16] J. Novotny, The Grid Portal Development Kit, Concurrency: Pract. Exper. 2000; 00:1-7, 2000.

[17] Novotny J., Tuecke S., Welch V. “An Online Credential Repository for the Grid:MyProxy,” *Proc. 10th IEEE Symp. On High Performance Distributed Computing*, 2001.

[18] Jason Novotny, Michael Russell, Oliver Wehrens, GridSphere: A Portal Framework For Building Collaborations, February 13, 2003.

[19] Abraham Silberschatz, Peter Baer Galvin and Greg Gagne, *Operating System Concepts*, Sixth Edition, John Wiley & Sons, Inc., 2002.

[20] Antonio Peris, Patricia Lorenzo and Flavia Donno, et al., *LHC Computing Grid:LCG-2 User's Guide*, LCG Experiment Integration and Support, August 4, 2005.

[21] J. Linn, Generic *Security Service Application Program Interface* Version 2, Update1, RFC 2743, IETF Network Working Group,Jan 2000.

[22] Giuseppe Rocca, Job Services with Genius Portal, *GENIUS/GILDA Tutorial*,May 2004.

[23] Eamon Tuathail,*Windows Shell Namespace Extensions Developer's Guide*, Clip-code Shell SDK,2000.

[24] Kurt Mueller,Steve Mock,*GridSphere Tutorial*,Aug 2004.

[25] Maytal Dahan, John Boisseau and Eric Roberts et al., *Build grid portals with Grid Portal Toolkit 3*,Oct 19, 2004.

[26] http://www.eu-egee.org

[27] http://fuse.sourceforge.net

[28] http://fuse.sourceforge.net/sshfs/

[29] https://genius.ct.infn.it

[30] http://gridport.net

[31] http://www.gridsphere.org

[32] http://grid.ucla.edu

[33] http://grid.ncsa.uiuc.edu/ssh/

[34] http://libgmail.sourceforge.net

[35] http://msdn.microsoft.com

[36] http://richard.jones.name/google-hacks/gmail-filesystem/

[37] http://shfs.sourceforge.net

[38] http://www.cngrid.org

[39] http://www.scgrid.cn

[40] http://www.viksoe.dk/code/gmail.htm

Fcrsf: An Application-Oriented Framework for Grid Resource Selection*

Liang Hu, Dong Guo, Bingxin Guo and Shilan Jin
College of Computer Science and Technology, Jilin University
Changchun, Jilin Province 130012, China
Email: guodong@jlu.edu.cn

ABSTRACT

Different applications have different characteristics and their requirements on resources may differ significantly. How to select suitable resources for applications is a critical and complex issue. This paper presents an application-oriented resource selection framework that provides a general-purpose resource selection service for different kinds of applications. This service adopts the approach of classifying the available resources into many clusters with different performance according to application preference and then selecting suitable resources to support the application execution. A preference-based fuzzy clustering technique is adopted to reach this goal. Besides meeting application requirements, this framework can select resources efficiently by reducing the search space of available resources. Furthermore, using this framework to select resources can avoid heavy loads' centralizing on only a few resources so as to improve load balance in grid environments.

Keywords: Grid, Fuzzy Clustering, Application Preference, Resource Selection

1. INTRODUCTION

In grid environments, the discovery and configuration of suitable resources for applications remain challenging problems [1] due to applications' multiple requirements and grid resources' features of being distributed, heterogeneous, autonomous, and dynamically available.

To address this issue, Fcrsf (the Fuzzy Clustering Resource Selection Framework) has been designed to locate grid resources that match application requirements. This framework is application-oriented which combines application preference and real-time resource information to identify a suitable set of resources. Its resource selection approach is to classify the available resources into many clusters with different performance according to application preference, and then to select suitable resources to support the application execution. A preference-based fuzzy clustering technique, which is proposed and proved in our previous work [2], is adopted to realize resources' reasonable clustering of resources. An algorithm has been designed for this framework to evaluate and select resources. Grid information service is used to collect resource information which is continually updated so that resources' selection decisions may be made at run time based on resources' deliverable performance. Meanwhile, this framework is also a general-purpose framework that can be used by different kinds of applications and grid environments.

The rest of this paper is organized as follows. Section 2 describes related works. In Section 3, the architecture of Fcrsf is described and Section 4 introduces grid resources' fuzzy clustering. In section 5, the resource selection approach is described. Section 6 describes experiments and performance evaluation and Section 7 concludes.

2. RELATED WORKS

Condor [3] provides a general resource selection mechanism based on the ClassAd language, which allows users to describe arbitrary resource requests and resource owners to describe their resources. A matchmaker is used to match user requests with appropriate resources. When multiple resources satisfy a request, a ranking mechanism sorts available resources based on user-supplied criteria and selects the best match. Because the ClassAd language and the matchmaker were designed for selecting a single machine on which to run a job, however, they can not easily be applied when a job requires multiple resources. To overcome ClassAd language's limitation, in [1], a general-purpose resource selection framework is presented to select multiple resources which meet an application's requirement. It requires users to provide application-specific mapping modules and ranking mechanism to personalize the resource selector, which is beyond the reach of common users. Our framework can select multiple suitable resources for applications, only requiring users to provide an application preference vector and the required resources' amount.

Plenty of resource brokers are available with their respective grid environments/scheduling systems such as Legion [4], European DataGrid [5], UNICORE [6], Nimrod/G [7], and GridLab [8]. These brokers were designed for their specific grid environment, thus are not easily extensible, nor are easily incorporated into other environments. In some cases, they are dependent on a specific job model or job submission functionality. In other cases, they are dependent on specific information sources that may not be available elsewhere. Finally, many of them have built-in selection policies such as a specific load-balancing scheme that may be unsuitable and/or undesirable in other environments [9].

Some researchers have applied fuzzy clustering method to solve resource selection problems. In [10], a machine selection algorithm classifies all available nodes into several logic clusters. According to the ratio of application's computing quantity to communication quantity, one or more suitable logic clusters, are selected separately by using the λ matrix method. This method considers no other application requirements but computing performance and communication performance, while our method can describe any application requirements by only defining application preference vector when constructing resource clusters. In [11], the target system, also called processing cell network, is pre-treated by fuzzy clustering method in order to realize the reasonable clustering of processor network. In the scheduling stage, the cluster with better synthetic performance will be chosen first. In this method, the resource clusters are not constructed for any specific kind of applications, so they may not be the most suitable resource sets for applications. However, our method

* Supported by Program for New Century Excellent Talents in University of China under Grant No. NCET-06-0300 and by National Natural Science Foundation of China under Grant No. 60473099. Corresponding author: Dong Guo, Phn: +86-431-85159404.

makes use of application preference to construct clusters, therefore, the clusters would be more suitable for applications.

3. ARCHITECTURE

Different applications have different characteristics, and their demands on the resources may differ significantly. The goal of our resource selection framework is to provide an application-oriented, general-purpose and easy-to-use selection approach.

Its resource selection approach includes two steps: firstly, resources' features which meet different kinds of applications are calculated and then they are classified into many clusters according to application preference so that suitable resource cluster(s) can be identified easily. Therefore, it is an application-oriented approach. Secondly, resource clusters are evaluated and then suitable resources are selected. Details will be described in Section 4 and Section 5.

The architecture of Fcrsf is shown in Fig. 1. Grid information service is provided by the Monitoring and Discovery Service (MDS4)[12], a component of the Globus Toolkit4^{TM}[13]. Except the grid information service, it comprises three modules: Feature Computation Center，Fuzzy Clustering Center and Resource Selector.

(1). Feature Computation Center receives resource information provided by MDS4, and uses them to compute the resource features. The results are sent to the Resource Feature Center and updated periodically.

(2). Fuzzy Clustering Center acquires information from Resource Feature Center and creates resource clusters by using grid resources' preference-based fuzzy clustering method according to application requirements.

(3). Resource Selector uses an algorithm, which is designed for this framework, to select resources and returns results to application.

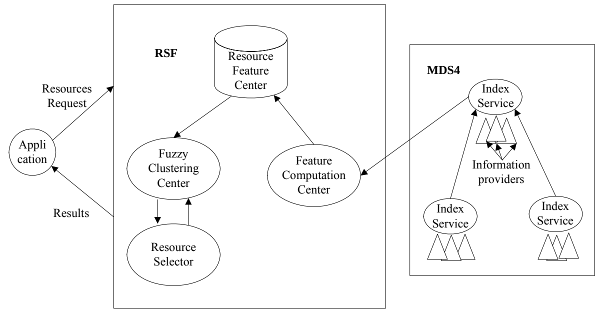


Fig.1. The Architecture of Fuzzy Clustering Resource Selection Framework

In this framework, the input are the required resources' amount and application's preference which should be defined as a vector by users. The output are the resources which meet application requirements.

4. RESOURCE FUZZY CLUSTERING

Resource performances' identification is the basis of resources' selection. Fcrsf adopts fuzzy clustering technique to classify resources into many clusters so that suitable resources can be identified easily.

4.1 Feature Computation

In grid environments, resources are heterogeneous and suit to different applications. For example, resources with highly powerful floating-point and fixed-point performance may suit to computing-intensive applications; resources with highly powerful I/O performance may suit to data-intensive applications; resources with highly powerful communication performance may suit to communication-intensive applications.

Therefore, it is necessary to compute resource features such as “computing performance”, “communication performance” and “trust” so that they can be allocated accurately. Feature Computation Center of this framework uses resource monitoring information which are provided by MDS4 to calculate resource features. Lots of research works have defined many functions that illustrate resource features. For example, in [10] the computing performance of grid resources (C_k) is defined as the function of three parameters: L_k, D_k, W_k. C_k = f (L_k, D_k, W_k),where L_k is system load, D_k is the result obtained from Dhrystone benchmark and W_k is the result obtained from Whestone benchmark. In our previous work, CPU performance tool has been designed to measure grid resources' computing performance [14]. By using formulas introduced in previous literature, some resource features can be defined or redefined to illustrate resources' performances such as “computing performance” and “communication performance”.

The results are sent to the Resource Feature Center and updated when new resource features are calculated out.

4.2 Resource Clustering Method

In [2], we propose and then prove a preference-based fuzzy clustering method. This method is adopted by Fcrsf to create resource clusters. It is described as follows.

Let $X=\{X_1,\ldots,X_n\}$ be a set of n resources and $X_i=\{X_{i1},\ldots,X_{im}\}$,where X_{ik} ($1 \le k \le m$) is the k^th^ feature of the i^th^ resource. The features of a resource set can then be represented as a $n \times m$ matrix, where m is the number of features which are used to describe resources.

Construct m similar matrixes $R^1 \ldots R^k \ldots R^m$ by applying Eq. (1) for every resource feature where $R^k = (r_{ij})^k \;\; n \times n$ ($1 \le k \le m$):

$$r_{ij}^{\ k} = \frac{min(x_{ik}, x_{jk})}{max(x_{ik}, x_{jk})} \qquad (1)$$

Let application preference (w) be a vector with m elements, one element for one feature:

$$w = \{w_1, w_2, \ldots, w_m\}$$

where w_k($1 \le k \le m$) is the application preference for the k^th^ feature and:

$$\sum_{k=1}^{m} w_k = 1 \qquad (2)$$

Construct matrix R which combines all features and preferences (w) by applying Eq. (3):

$$R = \sum_{k=1}^{m} w_k \times R^k \qquad (3)$$

that is,

$$r_{ij} = \sum_{k=1}^{m} w_k \times r_{ij}^{\ k} \qquad (4)$$

Then R satisfies reflexivity and symmetry and can be used as similar matrix to construct fuzzy clusters. The proof is in [2].

By using vector *w* as application preference, our method can not only take into consideration all features but also emphasize special features. When $w_k = 1$ and $w_i = 0$ ($i \neq k$), it clusters grid resources according to the kth feature while ignores all the other features which application does not interest; when $w_1 = w_2 = \ldots = w_m = 1/m$, it clusters grid resources with all the features considered.

4.3 Application Preference

Application A_i's preference w_i can be obtained by code analysis, previous execution experience and the tracing of application process. Applications, which include plenty of floating-point computation, fixed-point computation etc., may probably have the preference for computing performance, such as linpack [15] and IS of NPB [16]. Many tools, such as PAPI [17], can count the operations in application process and the obtained results can be used to analyze application preference. For example, suppose applications' preferences include resources' “computing performance”, “communication performance” and “trust”, according to the requirement for “trust” and different ratio of application's computing quantity to communication quantity, application preference can be defined. if application A_i ignores the “trust” and the ratio of computing operations to communication operations is 1:3, then its preference can be defined as $w_i = \{0.25, 0.75, 0\}$.

4.4 Clustering Steps

Fuzzy Clustering Center uses the λ matrix method to create the resource clusters. The clustering steps are as follows:

1. Standardize the initial resources' features so that they are in the interval [0,1] by applying Eq. (5)
$$r'_{ij} = \frac{r_{ij} - A_j}{B_j - A_j} \qquad (5)$$
where $A_j = \min(x_{ij})$, $B_j = \max(x_{ij})$.
2. Create m similar matrixes $R^1 \ldots R^k \ldots R^m$ by applying Eq. (1) for every resource feature, and then create the fuzzy similar matrix R_i under preference w_i by applying Eq. (3).
3. Calculate the transitive closure matrix R_i^* by using R_i. Let λ vary from 1 to 0, different results can be obtained. The nearer λ towards 1, the more similar are the resources in a cluster; the nearer λ towards 0, the more different are the resources in a cluster. At last a proper clustering result can be adopted according to application requirements.

As soon as resource clusters generate, they are sent to resource selector to identify suitable resources.

5. RESOURCE SELECTION

5.1 Evaluation of Clusters' performance

Clusters' performance evaluation is the basis of resource selection. Resource selector applies Eq. (6) to calculate the average performance of every cluster which is provided by Fuzzy Clustering Center, and the results can be used to select resources. The bigger the $\overline{P(Ci)}$ is, the more suitable the resources in the cluster are for application.

$$\overline{P(Ci)} = \frac{1}{n} \sum_{p_k \in Ci} \sum_{j=0}^{m} w_j * p_{[k][j]} \qquad (6)$$

Where $\overline{P(Ci)}$ is the average performance of the ith cluster; $p_{[k][j]}$ is the jth performance of the kth resource in the ith cluster; w_j is the application preference for the jth performance.

5.2 Resource Selection Algorithm

To select suitable resources from clusters, a resource selection algorithm is designed. It evaluates clusters' performance and seeks to identify resources in one or more clusters to support the application execution. This algorithm's input are resources' amount (m) which application requires, application preference (w) and the clusters ($C_1,\ldots,C_n$) which are provided by Fuzzy Clustering center. Let the available resource set be ASourceSet, thus

$$\text{ASourceSet} = C_1 \cup C_2 \cup \ldots \cup C_k \ldots \cup C_n$$

Below is the algorithm.

```
BetterSet=NULL
For each $C_i \in \{C_1,\ldots C_n\}$
   $P(C_i)$=CalAverPerf($C_i$, $w_i$)
   //calculate the average performance of $C_i$ by using (6)
Sort $C_1,\ldots,C_n$ by descending $P(C_i)$,and the result is
   $<C_1',\ldots, C_n'>$
i=1;
flag=false;
Do
{
      If count($C_i'$)≥m
         //count resources' amount of $C_i'$
         {
            BetterSet=BetterSet ∪ RandomSelect($C_i'$, m)
            //select m resources from $C_i'$ randomly and
               add them to BetterSet
            flag=true
         }
      else
         {
            BetterSet = BetterSet ∪ SelectedAll($C_i'$)
      //select all resources in $C_i'$ and add them to BetterSet
            m=m-count($C_i'$);
            i=i+1;
         }
}
while (flag or i>n)
ASourceSet= ASourceSet - BetterSet
If BetterSet =NULL
      return failure
else
      return BetterSet
```

This algorithm firstly calculates the average performance of every cluster, and then sorts the clusters by descending the average performance. Followingly, it repeatedly removes the resources in the best cluster into the “BetterSet” until the amount of resources reaches the requirements or available resources are all selected, or fail if no resources are available. When removing resources into the “BetterSet”, resources are selected randomly from one cluster, since resources in the same cluster have similar performance. This can enhance the load balance, too.

This algorithm is not guaranteed to find a best solution if it does exist. The set-matching problem can be modeled as an optimization problem under some constraints. Since this problem is NP-complete in some situations, it is difficult to find a general algorithm to solve the problem efficiently when the number of resources is large [1]. Our work provides an efficient algorithm with whole complexity O(n × logn) (including computing average performance of every cluster, sorting the clusters and resources matching), where n is the number of resource clusters.

6. EXPERIMENTS AND PERFORMANCE EVALUATION

6.1 Experiments

Ten resources are used to create clusters in our experiment. Suppose $A = \{A_1, A_2, \ldots, A_n\}$ be a set of grid applications and their preferences include resources' "computing performance", "communication performance" and "trust". By using formulas introduced in previous literature, ten resources' performance are calculated and illustrated in Table 1. where n=10,m=3.

6.1.1 Clustering Results Under Different Preference： According to different applications' preferences, different clusters can be obtained to meet applications' requirements. Suppose A_i is a computing-intensive application and its preference vector is w_i={1,0,0};A_j is another application that requires resources more trustworthy and its preference vector is w_j={0,0,1}, the fuzzy clustering results are illustrated in Table 2.,when λ=0.89,0.81 and 0.75.

As shown in Table 2., the fuzzy clustering results are different under different preferences, even if the λ values are the same. Therefore, it is necessary to carry out fuzzy clustering under application preference to select more suitable resources for application.

Table 1. Ten Grid Resources' Initial Performance

resources	computing performance	communication performance	trust
x_1	95	120	87
x_2	72	30	72
x_3	88	98	82
x_4	56	69	69
x_5	76	93	92
x_6	62	115	88
x_7	89	81	94
x_8	39	36	78
x_9	77	49	81
x_{10}	59	159	75

Table 2. The Results of Different Preference-Based Fuzzy Clustering

	λ = 0.89	λ = 0.81	λ = 0.75
w_i ={ 1, 0, 0}	$\{x_1, x_3, x_7\}$, $\{x_2, x_5, x_9\}$, $\{x_4\}$,$\{x_6\}$,$\{x_8\}$, $\{x_{10}\}$	$\{x_1, x_3, x_7\}$, $\{x_2, x_5, x_9\}$, $\{x_4, x_6, x_{10}\}$, $\{x_8\}$	$\{x_1, x_2, x_3, x_5, x_7, x_9\}$ $\{x_4, x_6, x_{10}\}$, $\{x_8\}$
w_j ={ 0, 0, 1}	$\{x_1, x_6\}$, $\{x_3, x_9\}$, $\{x_5, x_7\}$, $\{x_2\}$,$\{x_4\}$,$\{x_8\}$, $\{x_{10}\}$	$\{x_1, x_5, x_6, x_7\}$, $\{x_3, x_9\}$, $\{x_2\}$,$\{x_4\}$,$\{x_8\}$,$\{x_{10}\}$	$\{x_1, x_5, x_6, x_7\}$, $\{x_3, x_8, x_9\}$ $\{x_2\}$, $\{x_4\}$, $\{x_{10}\}$

By changing the value of λ, different scale clusters can be obtained so as to meet the requirements of different scale applications. As can be seen from Table 2., when λ decreases, the amount of clusters decreases, while the resource amount in some clusters increases. If the application's scale is small, the resource cluster which conforms to application preference best can be allocated to the application; while the application's scale is large, the value of λ can be decreased to acquire a proper cluster with a large amount of resources, or several clusters with higher specific performance to be allocated so as to increase the application performance.

6.1.2 Clusters' Performance： Table 3. illustrates the average performance of every cluster and its resources' related performance which contributes to its average performance under different preferences when λ = 0.81.

As can be seen from Table 3., under a certain preference, resources in the same cluster have similar specific performance; resources of different clusters have different specific performance.

Table 3. The Average Performance of Different Clusters When λ=0.81

w_i={1,0,0}		w_j={0,0,1}	
$C_1=\{x_1=95, x_3=88, x_7=89\}$	$\overline{P(C_1)}$=90.67	$C_1=\{x_1=87, x_5=92, x_6=88,x_7=94\}$	$\overline{P(C_1)}$=90.25
$C_2=\{x_2=72, x_5=76, x_9=77\}$	$\overline{P(C_2)}$=75	$C_2=\{x_3=82, x_9=81\}$	$\overline{P(C_2)}$=81.5
$C_3=\{x_4 =56, x_6=62, x_{10}=59\}$	$\overline{P(C_3)}$=59	$C_3=\{x_2=72\}$	$\overline{P(C_3)}$=72
$C_4=\{x_8=39\}$	$\overline{P(C_4)}$=39	$C_4=\{x_4=69\}$	$\overline{P(C_4)}$=69
		$C_5= \{x_8=78\}$	$\overline{P(C_5)}$ = 78
		$C_6= \{x_{10}=75\}$	$\overline{P(C_6)}$ = 75

6.2 Performance Evaluation

As illustrated in above sections, this framework is designed to be application-oriented which selects resources according to application requirements. That is, computing-intensive applications can obtain resource clusters with higher computing performance, while communication-intensive applications can obtain clusters with higher communication performance. By using preference-based fuzzy clustering method and resource selection algorithm, it reaches this goal. Meanwhile, according to different applications, different resource features can be defined or redefined to select suitable resources for them, so it has good flexibility.

The time complexity and space complexity of resources selection can be improved. Once resource features' similar matrices (the number is m) have been created, all applications can use them to create their own similar matrix (R) under their preferences until the resource features are updated. It is not necessary to search resources, which are enormous in amount, for every application, hence the time complexity and space complexity for resource selection can be improved.

Scalability can be enhanced by changing the value of λ so as to obtain proper amount of resource cluster(s), as is discussed in above subsection.

Furthermore, by defining application preference vector, grid resources can be divided into clusters according to different criteria. Although clusters created by different application preferences may have some intersection of resources, the load may not centralize on only a few best resources, thus load balance can be improved.

7. CONCLUSIONS

Grid applications usually have special requirements for grid resources which are geographically distributed, heterogeneous in nature, owned by different individuals or organizations with their own policies, different access, and dynamically varying loads and availability. How to select suitable resources for applications is a critical and complex undertaking.

This paper presents an application-oriented resource selection framework that provides a general-purpose resource selection service for different kinds of applications. This framework combines application preference and real-time resource information to identify a suitable set of resources. A new technique called preference-based fuzzy clustering is used to realize reasonable clustering of resources so that suitable resources can be identified easily. Besides meeting application requirements, this framework can select resources efficiently by reducing the search space of available resources. Furthermore, using this framework to select resources can avoid heavy loads' centralizing on only a few resources so as to improve load balance. Our future work will focus on designing a scheduling system based on this framework.

REFERENCES

[1] CH. Liu, LY. Yang, I. Foster and D. Angulo, “Design and Evaluation of a Resource Selection Framework for Grid Applications,” *Proceedings of IEEE International Symposium on High Performance Distributed Computing (HPDC-11)*, Edinburgh, Scotland, July 2002.

[2] D. Guo, L. Hu, SL. Jin, BX. Guo, “Applying Preference-based Fuzzy Clustering Technology to Improve Grid Resources Selection Performance,” to appear in the the 4th *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'07)*, IEEE CS Press. Haikou, China.(accepted)

[3] R. Raman, M. Livny, M. Solomon, “Matchmaking: Distributed Resource Management for High Throughput Computing,” *Proceedings of 7th IEEE International Symposium on High Performance Distributed Computing,* Jul. 1998.

[4] S.J. Chapin, D. Katramatos, J. Karpovich, A.S. Grimshaw, “The Legion Resource Management System,” *Proceedings of the 5th Workshop on Job Scheduling Strategies for Parallel Processing*, Apr. 1999.

[5] W. Lee, S. McGough, S. Newhouse, J. Darlington, “Load-balancing EU-DataGrid Resource Brokers,” *UK e-Science All Hands Meeting*, Sep. 2003.

[6] J. MacLaren, “Resource Management and Resource Brokering Using UNICORE,” Global Grid Forum 7th Workshop on Grid Scheduling Architecture, Mar.2003. Available at
http://www.gridsched.org/ggf7/GGF7_Talk3.ppt.

[7] R. Buyya, D. Abramson, J. Giddy, “Nimrod/G:An Architecture for a Resource Management and Scheduling System in a Global Computational Grid”. *Proceedings of 4th International Conference on High Performance Computing in Asia-Pacific Region (HPC Asia 2000)*, Beijing, China, May 14-17, 2000.

[8] J. Nabrzyski, “GridLab Resource Management System,” Global Grid Forum 7th Workshop on Grid Scheduling Architecture, Mar. 2003. Available at
http://www.gridsched.org/ggf7/GGF7_Talk2.ppt.

[9] P.Z. Kolano, Proceedings of 2004 IEEE International Symposium on Cluster Computing and the Grid (CCGrid'04), IEEE Computer Society, April 19-22, 2004, Chicago, USA, pp. 563- 571.

[10] XL. Gui, QJ. Wang, WQ. Gong, DP Qiang, “Study of a Machine Selection Algorithm for Grid Computing,” *Journal of Computer Research and Development*, Peking, vol. 41, no.12, pp.2189-2194, 2004.

[11] XL. Du, CJ Jiang, GR Xu, ZJ Ding, “A Grid DAG Scheduling Algorithm Based on Fuzzy Clustering,” *Journal of Software,* vol.17, no.11, pp.2277-2288, 2006.

[12] J. M. Schopf, L. Pearlman, N. Miller, C. Kesselman, I. Foster, M. D'Arcy, and A. Chervenak, “Monitoring the Grid with the Globus Toolkit MDS4,”*Proceedings of SciDAC 2006*, June 2006.

[13] Globus Toolkit Version 4: Software for Service-Oriented Systems. I. Foster. IFIP International Conference on Network and Parallel Computing(NPC 2006), Springer-Verlag LNCS 3779, pp 2-13, 2006.

[14] D. Guo, L. Hu, M. Zhang, ZP. Zhang, “GcpSensor: a CPU Performance Tool for Grid Environments,” *Proceedings of the 5th International Conference on Quality Software (QSIC 2005)*, Melbourne, Australia, September 19-20, 2005, IEEE Computer Society Press, pp. 273-278.

[15] J. J. Dongarra, P. Luszczek and A. Petitet, “The LINPACK Benchmark: Past, Present, and Future,” *Concurrency and Computation: Practice and Experience*, vol. 15, no. 9, 2003, pp. 803-820.

[16] NAS. NAS PARALLEL BENCHMARKS. http://www.nas.nasa.gov/Resources/Software/npb.html. January 31, 2007.

[17] S. Browne et al, “A Portable Programming Interface for Performance Evaluation on Modern Processors,” *The International Journal of High Performance Computing Applications*, vol. 14, no. 3, 2000, pp. 189-204.

Extending Role-based Access Control Model with Context for Grid Applications

Yanfen Cheng, Hanbing Yao
School of Computer Science and Technology,WHUT, Wuhan 430063, China
Email: chengyanfen@whut.edu.cn

ABSTRACT

The Grid infrastructure presents many challenges due to its inherent heterogeneity, multi-domain characteristic and highly dynamic nature. One critical challenge is providing authentication, authorization and access control guarantees. Despite the recent advances in access control approaches applicable to Grid computing, there remain issues that impede the development of effective access control models for Grid applications. Amongst them are the lack of context-aware models for access control, and reliance on identity or capability-based access control schemes. In this paper, we propose RCBAC model that extends the RBAC with context constraints. The RCBAC mechanisms dynamically grant and adapt permissions to users based on a set of contextual information collected from the system and user's environments, while retaining the advantages of RBAC model. We also describe the implementation architecture of RCBAC for the grid application.

Keywords: RCBAC, RBAC, Context, Access control, Grid Security

1. INTRODUCTION

The Grid Security Infrastructure (GSI) has been accepted as the primary authentication mechanism for the Grid computing. Developed as part of the Globus project, GSI defines single sign-on algorithms and protocols, cross-domain authentication protocols, and temporary credentials called proxy credentials. GSI is widely used and has been integrated into a number of Grid environments and applications[1]. However, the authorization and access control challenges are not fully addressed by existing approaches.

While many research efforts address important aspects of the overall authorization and access control problem in a Grid environment, these efforts focus on relatively static scenarios where access depends on the user's identity (or role)[2-4]. They do not address access control issues for Grid applications where the access capabilities and privileges of a subject not only depend on its identity but also on its security-relevant contextual information, such as time, location, or environmental state available at the time the access requests are made, and incorporate it in its access control decisions. These context parameters capture the dynamically changing access requirements in Grid application, and hence are critical to the effectiveness of the resulting access control scheme. In order for the access control to be effectively exercised in such scenarios with context-aware access requirements, the traditional access control models must be extended to make them context-aware. To this end, we propose a D-RBAC model for Grid applications.

The remainder of the paper is organized as follow: Section 2 presents RBAC model. Section 3 describes our approach, including a brief presentation of security context, and presents a formal define for D-RABC. Section 4 describes D-RABC framework for grid application. Section 5 concludes this paper.

2. Role Based Access Control

RBAC model was first presented by Sandhu and has recently received increasing attention in the security community[5]. As opposed to DAC and MAC model based on a simple subject-object relation, RBAC model is based on three sets of entities called Users (U), Roles (R), and Permissions (P). A user (U) is a human being or an autonomous agent. A role (R) is a job title or a job function in the organization with, associated semantics concerning responsibility and authority. The permission (P) is a description of the type of authorized interactions a subject can have with one or ore objects.

Access control policy is embodied in RABC components such as user-role, role-permission, and role-role relationships. These RBAC components determine whether a particular user is allowed access to a specific piece of system data. A user can be assigned many roles, and a role can be assigned to many users. The many-to-many assignment relation User-Assignment (UA) captures this property. A role can be assigned much permission, and permission can be assigned to many roles. The many-to-many assignment relation Permission-Assignment (PA) captures this property. The formal definition for RBAC is as follows:

1. U, R, P, S which are, respectively, the sets of users, roles, permissions, sessions.
2. UA$\subseteq U \times R$, which is a many-to-many User-Assignment relation assigning a user to roles.
3. PA $\subseteq P \times R$, which is a many-to-many, Permission-Assignment relation assigning permissions to roles.
4. RH$\subseteq P \times R$ is a partial order on R called role hierarchy.
5. user: S→U, is a function mapping each session s_i to the single user(s_i) and is constant for the session's lifetime.
6. roles: S→2R is a function mapping each session s_i to a set of roles $roles(s_i) \subseteq \{r | (\exists r' \geq r)[(user(s_i), r') \in UA]\}$ so that session s_i has the permissions $U_r \in roles(s_i)\{p | (\exists r'' \leq r)[(p, r'') \in PA]\}$

Sandhu defines a comprehensive framework for RBAC models that are characterized as follows:

•RBAC0: the basic model where users are associated with roles and roles are associated with permissions.

• RBAC1: RBAC0 with role hierarchies.

•RBAC2: RBAC1 with constraints on user/role, role/role, and/or role/permission associations.

RBAC allows to express and to enforce enterprise-specific security policies and which simplifies the administration of access rights. Users can be made members of roles as determined by their responsibility and qualification and can be easily reassigned from one role to another without modifying the underlying access structure. Roles can be granted new permissions, or permissions can be revoked from roles as needed. RBAC can be used by the security administrator to enforce the principle of least privilege as well as static, dynamic, and operational policies of separation of duties.

Recently RBAC has been found to be the most attractive solution for providing security in a distributed computing infrastructure [6]. Although the RBAC models vary from very simple to pretty complex, they all share the same basic structure of subject, role and privilege. Other factors, such as relationship, time and location, which may be part of an access decision, are not considered in these models. The D-RBAC model presented in this paper extends RBAC to provide context-aware access control mechanisms for Grid applications.

3. RCBAC MODEL

In this paper, we present a Dynamically Authorized Role-Based Access Control (D-RBAC) mechanism. The D-RBAC mechanisms dynamically grant and adapt permissions to users based on a set of contextual information collected from the system and user's environments. The D-RBAC model extends the RBAC with context and content-based constraints, while retaining its advantages (i.e. ability to define and manage complex security policies). RBAC addresses many other issues such as role activation, revocation, role hierarchies and separation of duty constraints. These issues apply to D-RBAC as well.

3.1 Context-Aware Security

As its name suggests, context-based security is all about considering "context" explicitly in the specification of access control models[7-9]. Fig.1 illustrates the idea behind context-based security in the grid application. The grid environment is initially controlled with a specific configuration of the security policy in an initial context. This context is continually changing in request to triggers (dynamic changes in the environment). The security policy must then adapt itself to the new context.

By a security policy, we mean a specification that expresses clearly and concisely what access request are authorized and what are those that are denied for each type of user in each situation. Formally, a situation is what we call a security context.

Context-aware security adapts to cope with the new types of security problems introduced by the heterogeneous, dynamic and multi-domain nature in grid environments.

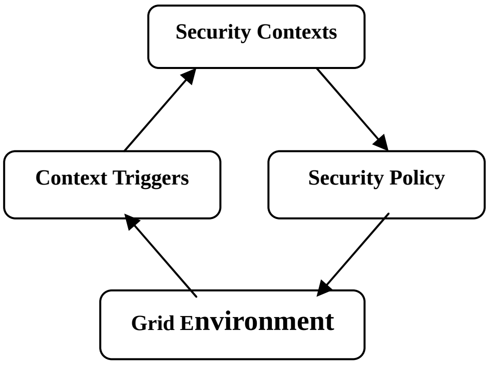


Fig.1. Context-aware security in grid application

3.2 A Formal Define For Security Context

This section defines the set of specifications needed to define D-RBAC for context-aware access control in Grid applications. In the following, we provide the formal definition of security context. In order to formalize the security context, we introduce a type system to allow specifying domains of legal values for various context parameters. The D-RBAC model relies on the components we define below:

Definition 1. Context Parameter (CP): A context parameter is represented by a data structure p, having the following fields: name ∈CN, type ∈CT, and a context function getValue(). The CN is a set of the possible names of context parameters, and the CT is a set of types of context parameters, and the context function of getValue()is a mechanism to obtain runtime values for specific context parameter. CP represents a certain property of the environment whose actual value might change dynamically (like time, date, or session-data for example). For example, the set CN may be defined as: CN = {time, location, duration, system_load}, with the corresponding set CT defined as: CT = {Time, String, Long, Integer}.

Context Parameter is separated from the main business logic of target applications. Since every context type definition is independent of the specification of the access rules, any change to them has no effect on other parts of the system.

Definition 2. Context Set (CS): A context set CS consists of n context parameters {CT1, CT2 ···, CTn}, n≥0, for any CTi, CTj, with i≠j and 1≤i, j≤n, we have that CTi.name≠ CTj.name (i.e. the parameter names must be distinct). By analyzing the grid application security requirements, application designers determine which context types will be used to specify access policy. Although the context set is determined before the application implementation, system administrators can dynamically add new ones when needed.

3.3 A Formal Define For RCBAC

Based on the formalization of the RBAC model, we present a precise description of D-RBAC model that includes security-relevant contextual information. Both role hierarchies and separation of duty in RBAC are meaningful in the D-RBAC, though they are omitted here in our description. We only consider flat user and security-relevant contextual information. This formalization can be extended to hierarchies and constraints similar to the RBAC1 and RBAC2 models. An overview of the D-RBAC is presented in Fig.2. We keep USERS, ROLES, OBS, OPS, PRMS and SESSIONS in the RBAC.

Definition 3. Context Condition (CN): CN = <CT > <OP> <Value>, CT∈CS, OP is a standard comparison and logical operator, VALUE is a specific value, and the type of VALUE is CP.type.

A Context Condition is a predicate (a Boolean function) that compares the current value of a context parameter either with a predefined constant. The corresponding comparison operator must be an operator that is defined for the respective domain. All variables must be ground before evaluation. Therefore each context parameter is replaced with a constant value by using the according context function prior to the evaluation of the respective condition. Examples for context conditions can be CN1: Date()<"2006-01-01", CN2: age(subject)>18.

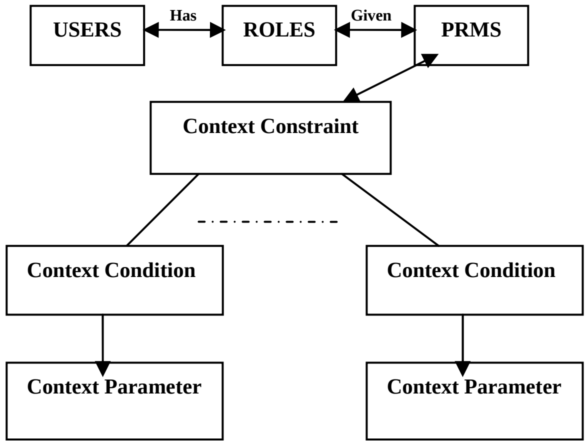


Fig.2. RCBAC Model

Definition 4. Context Constraints (CC): CC = CL1 $\cup$ CL2 $\cdots$ $\cup$ CLn, CL = CN1 $\cap$ CN2 $\cdots$ $\cap$ CNn. CN1,CN2 $\cdots$,CNn. CC are context condition. Based on this format, our access control schema is capable of specifying any complex context related constraint to describe all kinds of security requirements. System administrators can dynamically adapt context constraint.

Context Constraints are conditions that an object must satisfy in order that the user's attempt to perform an operation succeeds. These conditions involve security-relevant parameters of the attempted operation. This may include information gleaned from environment (such as the time of day, or whether it's a holiday), or state contained in the target object itself. These constraints are distinct from those defined in the base RBAC model, which constrain role definitions in order to avoid conflicting roles, promote separation of duties, etc. Systems such as[7] allow constraints, in the form of environment roles that are purely dependent on external properties rather than the properties of the objects or subjects involved in the operation. The Role Object Model defines a role as a set of policies. Constraints involving properties of the objects are used to limit the applicability of those policies over object instances[10].

Definition 5. D-RBAC: D-RBAC = {USERS, ROLES, OBS, OPS, PRMS, SESSIONS, CC}. The USERS, ROLES, OBS, OPS, PRMS and SESSIONS are defined in RBAC, the CC is context constraint.

Definition 6. Access Policy (AP): We define an access policy as a triple, AP = (R, P, C), R$\in$ROLES, P$\in$PRMS, C$\in$CC. If C is empty then this policy reverts to simple RBAC.
Definition 7. Access Request (AR): We define access request as a triple, AR = (R′ , P′ , RC), R′ $\in$ROLES, P′ $\in$PRMS, RC(runtime context) is a set of values for every context type in the Context Set. That is, RC = {CT1.getvalue(), CT2.getvalue(), $\cdots$, CTn.getvalue()}, {CT1, CT2 …, CTn} is the context set (CS) of the grid application.

An access request is granted only if there exists an access policy AP (R, P, C), such that R′= R, P′ = P, and C evaluates to true under RC (that is, when all CPi in context constraint C are replaced with their values in RC, then the resulted Boolean expression is true).

3.4 Dynamic Context Evaluation Algorithms

We can design the basic algorithm to determine whether an access request is authorized or not based upon the context parameter in our model.

The application passes an Access Request (ar) to the algorithm Request Permission, and receives a Boolean value in return - indicating whether the attempted operation should be allowed, or not. The ar contains the caller's roles and permissions and context constraints. The access control system first checks whether the application's Access Policy contain the user's Access Request. Then the context constraints are populated through plugging in values from the application's runtime environment. For each context condition, it is examined if the corresponding runtime value can be captured by an actual context function of the context parameter. If, however, no appropriate context function is available, one can implement a new context function in order to enforce the corresponding context condition without any effect on other parts of the system.

```
Algorithm 1: RequestPermission (AccessRequest ar)
CPS = {} //initialize candidate access policy set
for each AP in PS//PS are policy set
    if (ar.R' ∈ AP.R) and (ar.P' = AP.P)
      put AP into CPS
    end if
end for
result = false
for each AP in CPS
      if (EvaluateContexts(AP.C) is true)
        result = true
        break
      else
        result = false
      end if
end for
return result
```

```
Algorithm 2: EvaluateContexts(Constraint rc)
for each CL in rc
   for each CN in CL
         if (<CP.getvalue()><OP><VALUE> = false)
              //CP.getvalue() get CP'S runtime value
              //OP is specific operator of CN
          CL = false
          break
         end if
   end for
   if (CL = true)
         return true
   else
         continue
   end if
end for
return false
```

4. RCBAC FRAMEWORK FOR GRID APPLICATIONS

A prototype of the D-RBAC model has been implemented as

part of our lab's Grid system on the top of OGSI. It is a Grid-based computational collaboration that enables geographically distributed scientists and engineers to collaboratively access, monitor, and control distributed applications, services, resources and data on the Grid using grid portal. Key components of D-RBAC framework include:
Grid Portals: Providing users with pervasive and collaborative access to Grid applications, services and resources. Using these portals, users can discover and allocate resources, configure and launch applications and services, and monitor, interact with, and steer their execution. The grid portals include authentication module and global authority service module [11].

User Context Agent: Capturing all security-relevant information about a particular user.

Object Context Agent: Capturing all security-relevant information about the target object.

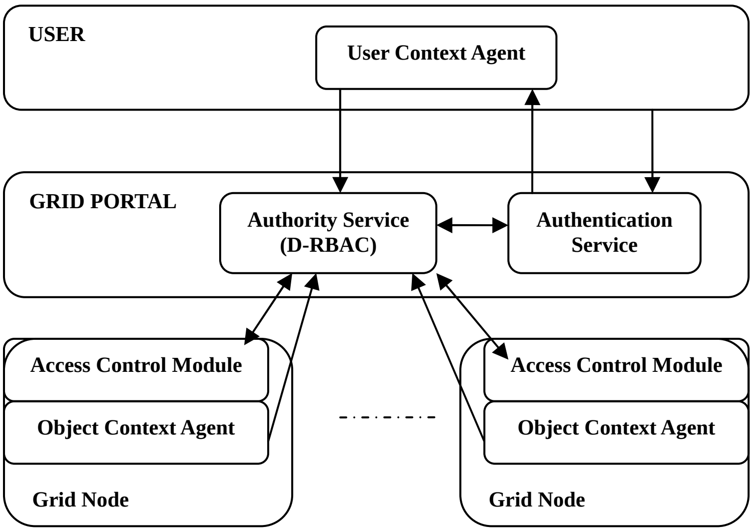


Fig.3. RCBAC framework Grid application

An overview of the D-RBAC for Grid applications is presented in Fig.3. The D-RBAC model ensures the users can access, monitor and steer Grid resources/applications/services only if they have appropriate privileges and capabilities. As the Grid environment is dynamic, this requires dynamic context aware access management. Note that authentication services are provided by GSI.

In our implementation, users entering the Grid application using the portal are assigned a set of roles when they log in. A user context agent is then locally set up for each user, which dynamically adjusts the user context. Similarly, the object context agents are set up at the application (or service/resource) for each role that will access it. The object context agents similarly adjust the object context.

As an illustration, assume that the following access request is submitted for evaluation to the grid application:
<R="guest", P="view", C={p1{time, Time}, p2{location, String}, p3{duration, Long}, p4{system_load, Integer}>.

The context recorded at the time of access request is captured by context agent, and provided to the system as part of the request. Now, assume that the following AP is applicable to the permission P:

<R="guest", P="view", C=CC>
CC = CL1∩CL2∩CL3∩CL4
CL1: {time > 8:00} AND {time < 18:00)
CL2: {location = "admin1"} OR location = "admin2"}
CL3: {duration≤600s}
CL4: {system_load != "high">

Based on this information, the system would return authorization decision for this access request. The available contextual information indicates that the access conditions are satisfied.

5. CONCLUSIONS AND FUTURE RESEARCH

In this paper we described a Dynamically Authorized Role-Based Access Control infrastructure that extends the traditional RBAC model to gain many advantages from its context-aware capability. Our research motivation comes from the complicated access control requirements in Grid application. Traditional RBAC is not able to specify a sufficiently fine-grained authorization policy or specify constraints that should be applied to an access policy. Our new access control infrastructure is dynamic and distributed with these advantages:

1 The D-RBAC model extends traditional RBAC by associating access permissions with context-related constraints. Every constraint is evaluated dynamically against the current context of the access request. Therefore, the model is capable of making authorization decisions based upon context information in addition to roles.
2 Our context-aware access control is applied dynamically. At design time, administrators have great flexibility to specify complex context-aware authorization policies. At run-time, our authority service can enforce any context-aware policy automatically because it is not statically bound to any application.
3 Context information is separated from the main business logic of target applications. Since every context type definition is independent of the specification of the access rules, any change to them has no effect on other parts of the system. Thus our security infrastructure is flexible and permits easy extensibility.

Although context constraints can be modeled and used in a straightforward manner, they can potentially add a great deal of complexity to access control policies. On the other hand, they add much flexibility and expressiveness, and allow for the definition of fine-grained access control policies as they are often needed in real-world applications. We intend to report the detailed results of our on-going implementation efforts in some future work. We also plan to explore the interplay of contextual conditions in the presence of separation of duty constraints. Separation of Duty principles is a type of access control policy that require that two or more users be responsible for the completion of a business process. By distributing the responsibility of a business process between numerous users, there are fewer opportunities for one user to commit a fraudulent act without being discovered. It is critical to ensure that the access to grid resources based on context constraints do not violate any separation of duty constraints.

REFERENCES

[1] Foster I., Kesselman C., Tsudik G.et al, “A security architecture for computational grids,” in *Proceedings of the 5th ACM Conference on Computer and Communications Security*, San Francisco, CA, USA, 1998, 83-92.

[2] Foster I., Kesselman C., Tsudik G.et al, “A security architecture for computational grids,” in *Proceedins of the 5th ACM Conference on Computer and Communications Security*, San Francisco, CA, USA, 1998, 83-92.

[3] Johnston W., Mudumbai S, Thompson M, “Authorization and attribute certificates for widely distributed access control,” in *Proceedings of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*, WET ICE.Stanford, 1998, 340-345.

[4] Pearlman L., Welch V., Foster I. et al, “A community authorization service for group collaboration,” in *Proceedings Third International Workshop on Policies for Distributed Systems and Networks*, Monterey, 2002, 50-59.

[5] Alfieri R., Cecchini R., Ciaschini V. et al, “VOMS, an authorization system for virtual organizations,” in *Proceedings of the First European Across Grids Conference*, Santiago de Compostela, 2004, 33-40.

[6] Sandhu Ravi S.,Coyne Edward J.,Feinstein Hal L.et al, “Role-based access control models,” *IEEE Computer*, 1996,29(2):38-47.

[7] Moyer M. J., Ahamad M, “Generalized role-based access control,” in *Proceedings of International Conference on Distributed Computing Systems*, Mesa, AZ:Institute of Electrical and Electronics Engineers Inc.,2001.391-398.

[8] Lupu Emil, Sloman Morris, “Policy based role object model,” in *Proceedings of the International Enterprise Distributed Object Computing Workshop*, EDOC.Gold Coast, Aust: IEEE, Piscataway, NJ, USA,1997.36-47.

[9] Mcdaniel Patrick, “On context in authorization policy,” in *Proceedings of ACM Symposium on Access Control Models and Technologies*, Villa Gallia, Como, Italy,2003.80-89.

[10] Kumar A., Karnik N. , Chafle G, “Context sensitivity in role-based access control,” *Operating Systems Review*, 2002, 36(3):53-66.

[11] HePing Hu and HanBing Yao, “A Scheme for Authentication and Authorization in a Grid Application,” in *Proceedings of the 19th International Conference on Advanced Information Networking and Applications*, Taipei, Taiwan, 2005,.383-387.

A Job Assignment Scheme Based on Auction Model and Particle Swarm Optimization Algorithm for Grid Computing

Xingwei Wang, Lin Han, Min Huang
College of Information Science and Engineering, Northeastern University
Shenyang, 110004, P.R.China
Email: wangxw@mail.neu.edu.cn

ABSTRACT

In order to use the distributed, autonomous, heterogeneous and dynamic resource in grid efficiently and co-ordinately, the effective job assignment scheme for grid computing is necessary. Thus, in this paper, by introducing the knowledge of microeconomics and swarm intelligence, a job assignment method for grid computing is proposed, considering both deadline and cost simultaneously. It determines the resource trading price between the resource buyer and the resource provider based on the auction model, and then finds the optimal job assignment solution based on ICPSO (improved chaos search hybrid particle swarm optimization) algorithm. Simulation results have shown that the proposed method is both feasible and effective with higher utility and much balanced job assignment to resource compared with some other existing schemes.

Keywords: Grid, Job Assignment, Resources Pricing, Auction, Particle Swarm Optimization

1. GRID MODEL

The grid resources are aggregated by various resource domains interconnected by network. In each domain, there is one broker responsible for its resource management and providing resources to its users. When its resources are deficient, the broker could buy resources from other domains, doing inter-domain resource transactions.

The domain brokers of buying or providing resources are denoted as RB and RP respectively. When RB wants to buy some resources, it firstly sends a request to GIS (Grid Information Server). The request is described by a triplet $< t_s, t_d, C_b >$, t_s is job starting time, t_d is job completion deadline, and C_b is the required lower bound for computing capability with MIPS (Million Instructions Per Second) as its unit.

GIS sends these requests to each RP. If a RP would like to participate in resource trading, it will send a reply to GIS. The reply is described by a triplet $< p_r, t_e, C_p >$, p_r is CPU price with G$/MIPS as its unit, G$ is grid dollar [4], t_e is the earliest job completion time, C_p is computing capability RP could provide. GIS then sends these replies to each RB.

2. RESOURCE PRICING BASED ON ACUTION MODEL

Auctions can speedup trading [5]. Usually, RB knows the resource value to itself more clearly than RP, thus RP could get more information through auction. In this paper, the necessary improvement has been made to the classical auction model [5] in order to enable RB to make an offer more rationally, especially when resources are sufficient.

2.1 Symbol Definition

Suppose that there are m RPs and n RBs in grid. Symbols used in this paper are defined as follows:

$R_{p\,j}$: $j=1,2,\ldots,m$, the jth RP.

$R_{b\,i}$: $i=1,2,\ldots,n$, the ith RB.

P_j: $j=1,2,\ldots,m$, the current highest offer to $R_{p\,j}$ from all RBs.

O_j : $j=1,2,\ldots,m$, the base price of $R_{p\,j}$ offered at the beginning of the auction.

P_{ij}: $j=1,2,\ldots,m$, $i=1,2,\ldots,n$. At the end of each auction round, an offer sequence $(P_{1j}, P_{2j},\ldots, P_{ij},\ldots, P_{nj})$ to $R_{p\,j}$ is formed, where P_{ij} is the offer of $R_{b\,i}$ to $R_{p\,j}$.

$C_{p\,j}$: $j=1,2,\ldots,m$, the current computing capability of $R_{p\,j}$.

l_i: In order to prevent fraud, $R_{b\,i}$ gives its offer to both GIS and $R_{p\,j}$ simultaneously. GIS serves as an auctioneer. It usually sets a minimum cost l_i for $R_{b\,i}$ and gets the highest offer in each auction round.

C_{wij}: the value of $R_{p\,j}$ to $R_{b\,i}$ and defined as follows:

$$C_{wij} = (H_{ij} + \rho \cdot C_{p\,j}) \times e^{-\delta t} \tag{1}$$

Here, H_{ij} is the history value of $R_{p\,j}$ to $R_{b\,i}$; ρ and δ are regulation parameters, $\delta > 0$, ρ is a real number; t is auction duration, and the value of C_{wij} declines continuously as time passing.

2.2 Utility

The utility U_{Bij} of $R_{p\,j}$ to $R_{b\,i}$ is calculated as follows:

$$U_{Bij} = C_{wij} - P_j - l_i \tag{2}$$

When $R_{b\,i}$ gives its first time offer, $P_j = O_j$. If $U_{Bij} \le 0$, $R_{b\,i}$ quits from the auction; otherwise, the offer P_{ij} of $R_{b\,i}$ to $R_{p\,j}$ is calculated as follows:

$$P_{ij} = P_j + th(C_{wij}) \tag{3}$$

*This work is supported by the National High-Tech Research and Development Plan of China under Grant No. 2006AA01Z214; the National Natural Science Foundation of China under Grant No. 60673159; Program for New Century Excellent Talents in University; Specialized Research Fund for the Doctoral Program of Higher Education; the Natural Science Foundation of Liaoning Province under Grant No. 20062022.

$$th(C_{wij}) = \frac{e^{C_{wij}} - e^{-C_{wij}}}{e^{C_{wij}} + e^{-C_{wij}}} \quad (4)$$

Here, $th(C_{wij})$ is a hyperbolic tangent function of which variable field is $[0,+\infty)$ and function field is $[0,1)$. Obviously, the larger the value of C_{wij} is, the larger the value of P_{ij} will be.

Suppose that $C_{w1j} = \max_{1\le i\le n}\{C_{wij}\}$ and the value C_{w2j} of R_{pj} to R_{b2} is uniformly distributed on $[0, C_{w1j}]$, then

$$P(C_{w2j} = C_w) = 1/C_{w1j} \quad (5)$$

$$P(C_{w2j} \le C_w) = C_w / C_{w1j} \quad (6)$$

Thus, the probability that C_{w2j} equals to C_w and C_w is the second highest value of R_{pj} to all R_{bi} is calculated as follows:

$$P(C_{w2j} = C_w)\prod_{i=3}^{n} P(C_{wij} \le C_w) = \left(\frac{1}{C_{w1j}}\right)\left(\frac{C_w}{C_{w1j}}\right)^{n-2} \quad (7)$$

Based on Eq. (7), the expectancy value of C_w is calculated as follows:

$$P(C_{w2j} = C_w)\prod_{i=3}^{n} P(C_{wij} \le C_w) = \left(\frac{1}{C_{w1j}}\right)\left(\frac{C_w}{C_{w1j}}\right)^{n-2} \quad (8)$$

In this paper, suppose that each R_{bi} does not know the values of each R_{pj} to other R_{bi} s and does not know the order of C_{wij} in sequence $(C_{w1j}, C_{w2j}, \cdots, C_{wnj})$. Thus, assume that the highest-value is C_{w1j}, the offer P'_{1j} of R_{b1} is calculated as follows:

$$P'_{1j} = \frac{(n-1)}{n} C_{w1j} + \Delta \quad (9)$$

Here, Δ is the offer-adjusting step and $\Delta > 0$. Therefore, assume that C_{wij} is the highest value in sequence $(C_{w1j}, C_{w2j}, \cdots, C_{wnj})$, the rational offer P'_{ij} of R_{bi} to R_{pj} is calculated as follows:

$$P'_{ij} = \frac{(n-1)}{n} C_{wij} + \Delta \quad (10)$$

If $P_{ij} - P'_{ij} > \theta$ (θ is a preset threshold), R_{bi} should quit from the auction rationally. Thus, the following improvement is made to the classical auction procedure: only when $U_{Bij} > 0$ and $\left|P_{ij} - P'_{ij}\right| < \theta$, does R_{bi} give its offer at next round.

When each auction round ends, GIS gets the current highest offer and assigns it to P_j, and according to Eq. (2) and Eq. (3) recalculating U_{Bij} and P_{ij}.

The utility U_{Pij} of R_{bi} to R_{pj} is calculated as follows:

$$U_{Pij} = P_{ij} - O_j \quad (11)$$

If $U_{Pij} > 0$, R_{pj} trades with R_{bi}, otherwise R_{pj} aborts it.

2.3 Auction Procedure

For any R_{pj}, $1 \le j \le m$, the auction procedure is described as follows:

Step 1: R_{pj} declares its C_{pj} and O_j; set the times of auction rounds to be D; set counter d to be 0.

Step 2: According to Eq. (1), calculating C_{wij} for each R_{bi} ($1 \le i \le n$) and according to Eq. (2), calculating U_{Bij}: if $U_{Bij} \le 0$, R_{bi} aborts trading with R_{pj}, otherwise according to Eq. (3), calculating P_{ij}.

Step 3: GIS counts the number w of RBs who have done trading with R_{pj} this round. R_{bi} gets the value of w from GIS and according to Eq. (10) calculating P'_{ij}. If $\left|P_{ij} - P'_{ij}\right| < \theta$, R_{bi} gives its offer P_{ij}, otherwise R_{bi} aborts trading with R_{pj}.

Step 4: According to Eq. (11), R_{pj} calculating $U_{P_{ij}}$, $1 \le i \le n$. If $U_{Pij} > 0$, R_{pj} continues trading with R_{bi}, otherwise aborts it.

Step 5: GIS gets the highest offer P_j this around, $d = d + 1$.

Step 6: If $d > D$, R_{bi} with the offer P_j wins the resource R_{pj} and P_j is the final trading price, the auction ends; otherwise, go to Step 2.

Time complexity of this algorithms is $O(mDn)$.

3. JOB ASSIGNMENT METHOD

In [6], PRIMAL (Price based Optimal workload allocation scheme) is proposed. In this paper, a job assignment method is proposed based on PRIMAL, considering both time limit and cost. It uses ICPSO and the trading price determined by the auction procedure in section 2.3 to find a job assignment solution with the shortest completion time and the smallest cost under the specific time limit.

3.1 Particle Position and Rate Coding

$x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ represents the position of particle i in each dimension of D-dimension space. This integral queue with length D corresponds to a job assignment solution. Here, D is the number of RPs from which RB buys resources; $x_{id} \in [0, x_{\max}]$ is the amount of jobs assigned to R_{pd} with MI (Million Instruction) as unit and $x_{\max}$ is the most amount of jobs assigned to R_{pd}; $\sum_{d=1}^{D} x_{id} = \Omega$, $\Omega > 0$ is the total amount of jobs.

$v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$ represents the speed of the particle i in each dimension of D-dimension space, $v_{id} \in [-v_{\max}, v_{\max}]$, $v_{\max}$ is the adjusting step of the most amount of jobs assigned to R_{pd}. This integral queue with length D is used to adjust the amount of jobs assigned to R_{pd} .

$p_i = (p_{i1}, p_{i2}, \dots, p_{iD})$ represents the by far optimal position of the particle i, denoted by *pBest*. This integral queue with length D is used to record the by far optimal job assignment solution corresponding to the particle i and its fitness is denoted as *pBest*.

$p_g = (p_{g1}, p_{g2}, \dots, p_{gD})$ represents the best among all p_i . This integral queue with length D is used to record the global optimal job assignment solution corresponding to the whole particle swarm, denoted as *gBest* .

3.2 Particle Position and Speed Update

Updating process is described as follows:

Step 1: According to Eq. (12), updating inertia weight value $\omega(t)$, which changes with the iteration times.

$$\omega(t) = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{iter_{\max}} \cdot t \tag{12}$$

Here, $\omega_{\max} = 0.9$, $\omega_{\min} = 0.4$, $iter_{\max}$ is the maximum iteration times.

Step 2: According to Eq. (13), updating the position and speed of the particle i.

$$v_{id} = \omega v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \tag{13}$$

$$x_{id} = x_{id} + v_{id} \tag{14}$$

Here, c_1 and c_2 are two learning factors, usually their values are 2; r_1 and r_2 are two random numbers between 0 and 1.

3.3 Boundary Mutation Strategy

When the position of the particle i has been updated, for any $x_{id} \notin [0, x_{\max}]$: if $x_{id} > x_{\max}$, $x_{id} = x_{\max} - c \times r \times x_{\max}$; if $x_{id} < 0$, $x_{id} = 0 + c \times r \times x_{\max}$.

When the speed of the particle i has been updated, for any $v_{id} \notin [-v_{\max}, v_{\max}]$: if $v_{id} > v_{\max}$, $v_{id} = v_{\max} - 2 \times r \times v_{\max}$; if $v_{id} < -v_{\max}$, $v_{id} = -v_{\max} + 2 \times r \times v_{\max}$.

Here, r is a random number between 0 and 1. c is a number between 0.01 and 0.5 and its value has significant influence on the result, thus it should be determined with regard to the objective function of the algorithm. Thus, the better value of c is determined by comparing those different values of the algorithm objective function under different values of c .

3.4 Particle Position Adjustment

It is possible that the particle position does not satisfy $\sum_{d=1}^{D} x_{id} = \Omega$ after update or being carried out the boundary mutation strategy. Suppose $x'_i = (x'_{i1}, x'_{i2}, \dots, x'_{iD})$ is the new position of the particle i and $\sum_{d=1}^{D} x'_{id} = \Omega' \neq \Omega$, do the following transformation to x'_i :

Step1: For $x'_i = (x'_{i1}, x'_{i2}, \dots, x'_{iD})$, do $x''_{id} = \left\lceil \frac{x'_{id}}{\Omega'} \cdot \Omega \right\rceil$, $1 \le d \le D$, then get $x''_i = (x''_{i1}, x''_{i2}, \dots, x''_{iD})$;

Step2: For $x'_i = (x'_{i1}, x'_{i2}, \dots, x'_{iD})$, choose any $x''_{it} (1 \le t \le D)$ and do $x''_{it} = x''_{it} - (\sum_{d=1}^{D} x''_{id} - \Omega)$, then use $x''_i = (x''_{i1}, x''_{i2}, \dots, x''_{it}, \dots x''_{iD})$ as the adjusted position of the particle i . Obviously, $\sum_{d=1}^{D} x''_{id} = \Omega$.

3.5 Initial Particle Swarm Generation

Step1：Set the particle swarm size to be m and set counter i to be 1;

Step2：Generate a integral queue $x_{i1}, x_{i2}, \dots, x_{iD}$ on the range $[0, x_{\max}]$ randomly with $\sum_{d=1}^{D} x_{id} = \Omega$ satisfied, and use it as the initial position of the particle i ;

Step3: Generate a integral queue $v_{i1}, v_{i2}, \dots, v_{iD}$ on the range $[-v_{\max}, v_{\max}]$ randomly as the initial speed v_i of the particle i .

Step4: $i = i + 1$; if $i > m$, the initial particle swarm $\{x_1, x_2, \dots, x_m\}$ has been generated, otherwise go to Step2.

3.6 Fitness Function

The fitness function of the particle i is defined as follows:

$$f_i = w_1 \cdot \frac{1}{\sum_{d=1}^{D} (x_{id} \cdot p_d)} + w_2 \cdot Z \cdot \lambda \tag{15}$$

$$Z = \left| \sum_{d=1}^{D} t_d - T \right| \tag{16}$$

$$t_d = \frac{x_{id}}{C_{pi}}, \quad d = 1, 2, \dots, D \tag{17}$$

Here, w_1 and w_2 denote cost and deadline preference weight respectively, reflecting their relative significance to the job. $x_{id} \cdot p_d$ is the cost paid under the specific job assignment solution corresponding to the particle i , p_d is the trading price with R_{pd} . T is the deadline for job completion and t_d is the time consumed by R_{pd} to complete its allocated jobs. $\sum_{d=1}^{D} t_d$ is the sum of the time consumed by each R_{pd} to complete its allocated jobs, $\sum_{d=1}^{D} t_d \le T$. λ is the adjustment coefficient, $\lambda > 0$.

Obviously, the smaller the cost, the shorter the job completion time, then the smaller $\sum_{d=1}^{D} t_d$, the larger the value of Z , the larger the fitness value, thus the better the particle.

3.7 Procedure of the Proposed Job Assignment Scheme

The procedure of the proposed job assignment scheme is described as follows:

Step0：Set the particle swarm flying space to be D -dimension and its size to be m . According to section 3.1 and section 3.5, generating the initial particle swarm. Set the maximum iteration times to be Q

and the counter q to be 1. Set parameter $\omega_{\max}$, $\omega_{\max}$, $iter_{\max}$, $x_{\max}$, $v_{\max}$, C, c_1, c_2 and c. Input the total amount of jobs to be assigned Ω.

Step1: For each particle, set its current position to be its *pBest* and compute its fitness value according to f Eq. (15), then select that particle position corresponding to the largest fitness value as *gBest*.

Step2: If $q > Q$, go to Step6, otherwise go to Step3.

Step3: For each particle i in the particle swarm, $i = 1,2,\ldots,m$, do as follows:

Step3.1: According to Eq. (13) and Eq. (14), update its speed and position respectively.

Step3.2: If its speed and position exceed the corresponding boundary, applying the boundary mutation policy described in section3.3 to them.

Step3.3: If $\sum_{d=1}^{D} x_{id} = \Omega$, applying the position adjustment policy described in section 3.4.

Step3.4: According to Eq. (15), computing its fitness value, if better than the fitness value of its corresponding *pBest*, update *pBest* with $x_i(q) = (x_{i1}(q), x_{i2}(q), \ldots, x_{iD}(q))$.

Step3.5: If the fitness value of *pBest* better than *gBest* 's, update *gBest* with *pBest*.

Step4: According to Eq. (18) and Eq. (19), computing the fitness value variance σ^2 of the particle swarm: if $\sigma^2 < C$, go to Step5; otherwise, $q = q + 1$, go to Step2.

$$\sigma^2 = \sum_{i=1}^{m} \frac{f_i - f_{avg}}{f} \tag{18}$$

Here, f is a normalization scaling factor to confine σ^2.

$$f = \begin{cases} \max\limits_{1\le i\le m} \left| f_i - f_{avg} \right| & \max\limits_{1\le i\le m} \left| f_i - f_{avg} \right| > 1 \\ 1 & others \end{cases} \tag{19}$$

Step5 : Set the optimal solution vector $z = (z_1, z_2, \ldots, z_d, \ldots z_D)$ for chaos search, let $z = p_g$ and $zBest = gBest$, do the following chaos search:

Step5.1: Set counter d to be 1;

Step5.2: Set the iteration times L and set counter l to be 1;

Step5.3: According to Eq. (20) and Eq. (21), get $z_d^{'}$: if $f(z_d^{'}) > zBest$, $zBest = f(z_d^{'})$ and $z_d = z_d^{'}$.

$$y_d^{'} = \mu \cdot y_d \cdot (1 - y_d) \tag{20}$$

$$z_d = y_d^{'} \cdot p_{gd} \tag{21}$$

Step5.4: $l = l + 1$;

Step5.5: If $l > L$, go to Step5.6, otherwise go to Step5.3;

Step5.6: $d = d + 1$; if $d > D$, the chaos search ended, get $z_d^{''}$ by adjusting $z_d^{'}$ as described in section 3.4 if necessary, $p_g = z^{''}$, $gBest = zBest^{''}$, $q = q + 1$, go to Step2; otherwise go to Step5.2.

Step6: Output $p_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$ and *gBest*, the algorithm ends.

4. SIMULATION RESEARCH

The proposed job assignment method in this paper has been implemented by simulation on Gridsim [7] and performance evaluation has been done. In the simulation, the number of *RB* is 3, the numbers of *RP* are 5、10、15、20 and 25, the relative parameters of every resource region range come from WWG [7]. In Fig.1, *RB* utility is a relative value with its minimum under the classical auction model as 1. Obviously, the proposed auction scheme is more effective.

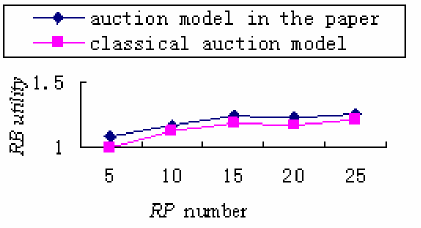


Fig.1. Comparison of *RB* Utility

Fig.2 shows the relation between the total amount of jobs (Ω) and their completion time ($\sum_{i=1}^{10} t_i$)under the proposed job assignment method and the PRIMAL. Table 1 show the amount of jobs($\beta_i, i = 1,\ldots 10$) assigned to resources with different price($R_{p_j}, j = 1,\ldots 10$). Obviously, the former is advantageous over the latter, the former assigns jobs to resources more even.

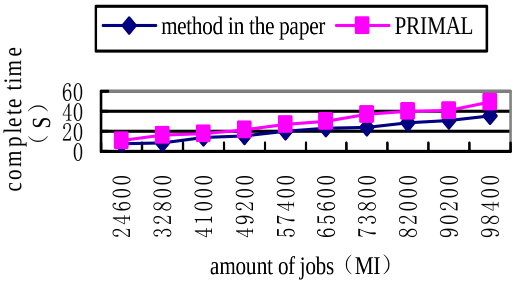


Fig.2. Comparison of completion Time

Table 1 amount of jobs comparison

RP	Price （G$/MI）	Amount of jobs β_i （MI）	
		Method in this paper	PRIMAL
R_{p_1}	0.2198	3079	3772
R_{p_2}	0.5014	2015	395
R_{p_3}	0.1845	3451	4301
R_{p_4}	0.3256	2928	3004
R_{p_5}	0.1395	3375	4578
R_{p_6}	0.5719	1672	256
R_{p_7}	0.2045	2959	3974
R_{p_8}	0.3784	2326	2001
R_{p_9}	0.4801	2314	985
$R_{p_{10}}$	0.2700	2824	3635

5. CONCLUSIONS

In this paper, a job assignment method in grid is proposed. It introduces microeconomics and swarm intelligence, considering both time limit and cost simultaneously. It determines the trading price between resource buyer and resource provider based on the auction model and then find the optimal job assignment solution using ICPSO. Simulation results have shown that the proposed method is both feasible and effective. It is advantageous over the PRIMAL. Its prototype system is being developed now and will be tested in ChinaGrid [1].

REFERENCES

[1] Duo Y., Liu P, *Grid Computing [M]*, Beijing: TsingHua University Press, 2002.

[2] Penmatsa S, Chronopoulos AT, “Job Allocation Schemes in Computational Grids Based on Cost Optimization[C]. IPDPS,” in *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS’05)* - Workshop 4 - Volume 05. Washington DC, USA: IEEE Computer Society, 2005: 180a-180a.

[3] Yang J., Zhou J., Yu J., et al, “Hybrid Particle Swarm Optimization Algorithm base on Chaos Search[J],” *Computer Engineering and Application*, Vol. 41. No.7. (2005):69-71.

[4] Buyya R, “Economic-based Distributed Resource Management and Scheduling for Grid Computing[D],” Melbourne, Australia: School of Computer Science and Software Engineering Monash Universityg, 2002.

[5] Buyya R, Abramson D, Venugopal S, “The Grid Economy[J],” in *Proceedings of the IEEE*, 2005, 93(3): 698-715.

[6] Ghosh P, Roy N, Das SK, et al, “A Game Theory based Pricing strategy for Job Allocation in Mobile Grids[C]. IPDPS,” in *Proceedings of the 18th IEEE International Parallel and Distributed Processing Symposium (IPDPS’04)*, Santa Fe, New Mexico, USA: IEEE Computer Society, 2004: 82-87.

[7] Buyya R, Murshed M. GridSim, “a toolkit for the modeling and simulation of distributed resource management and scheduling for Grid Computing [J],” *Concurrency and Computation: Practice and Experience*, 2002, 14(3): 1175-1220.

A Study on Modeling Supply Chain Management Based on Knowledge Management in Grid Computing Environment*

Tian Lan, Runtong Zhang, Shengbo Shi
School of Economics and Management, Beijing Jiaotong University
Beijing, 100044, China
Email:LT500800@163.COM

ABSTRACT

Although Supply Chain Management (SCM) is developed in the existing Internet environment, there are still some problems that need to be solved by a new technology. For example, SCM in the existing Internet environment can not meet the requirements of dealing with large amount of complex information, data statistics, analytical model analyses and management approaches. Now, the emergence of Knowledge Management (KM) and grid computing brings in the hope to solve such problems. This paper studies the SCM based on KM in grid computing environment from the viewpoints of both theory and model. Specifically, an optimized model of information transmission for SCM based on KM in grid computing environment is obtained and examined.

Keywords: Supply Chain Management, Knowledge Management, Grid Computing, Information Transmission, Optimization Model

1. INTRODUCTION

During the 1990s, the economic globalization, intense competition, consumer demand diversification, product life cycle shortening, rapid development and widely application of information technology make enterprises faced with a new environment, under which, enterprise management practice has been developed rapidly, as Supply Chain Management (SCM) became one of the important means to improve the competitiveness of enterprises. Particularly, with the development of Internet and E-Commerce, Supply Chain Management also has been developed with new forms and content.

At present, the emergence of Grid technology and Knowledge Management brings in a new opportunity for the development of Supply Chain Management. Grid's characteristics such as effective data-processing capability and fully use of idle resources can meet the demand of Supply Chain Management in the future. The nature relationship of “Supply and Demand” makes Supply Chain Management based on Knowledge Management in Grid computing environment become a new focus on the SCM research.

This paper is organized as follows: Section2 introduces some related theories such as Supply Chain Management and Grid computing. Sections3 proposes a new viewpoint of SCM based on KM in Grid computing environment. Section4 shows a model framework of the overall study. Section5 designs and examines an optimized model of information transmission for SCM based on KM in Grid computing environment. Last section concludes this work.

2. RELATED THEORIES

2.1 Supply Chain (SC), Supply Chain Management (SCM) and Supply Chain Management based on Knowledge management (SCM based on KM)

The viewpoint of Supply Chain (SC) was proposed in the late 1980s at first. With the emergence of global manufacturing, Supply Chain Management (SCM) has been universally applied, and finally becomes a new management model in manufacture.

Generally, the most comprehensive and widely accepted definition of SC is that, SC is a whole functional network chain, around its core business, structured from procurement of raw materials to manufacture of intermediate products and final products, until final products are consumed by the end-users. The function of SC is to link the suppliers, manufacturers, distributors, retailers and end-users through controlling the information flow, logistics and capital flow The Fig.1 below shows a typical model of Supply Chain.

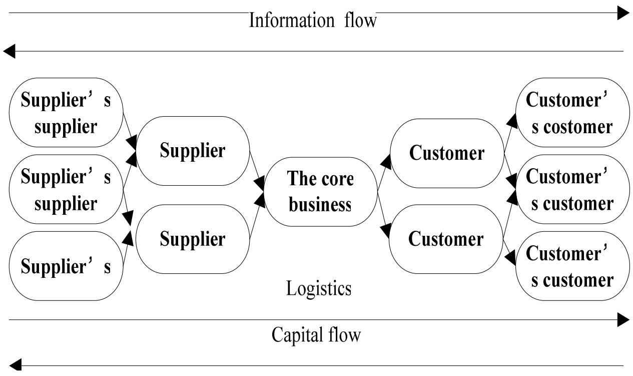


Fig.1. A typical model of Supply Chain structure

Therefore, Supply Chain Management is a process to design, plan and control the logistics, information flow, capital flow, value flow and traffic flow in the production and circulation with the management functions such as planning, organizing, coordination, controlling and incentives. The aim of SCM is to achieve the best combination to strengthen the core enterprise's competitiveness, improve the efficiency and effectiveness of SC and provide the greatest value to end-users at a minimum cost.

Supply Chain Management shows a series of factors which related to the production and circulation process and have different kinds of enterprises jointed to make them become a tight organism to defeat rivals. [2-6]

* Supported by Beijing Jiaotong University's Technology Funds (2005SM018) and Research Center for Beijing Transportation Development.

In current, the integration of Supply Chain Management and Knowledge Management has become a trend in SCM and the purpose of Supply Chain Management based on Knowledge Management is to improve the knowledge sharing and the efficiency cooperation between enterprises in Supply Chain.

2.2 Grid and Grid Computing

Grid technology can manage the Internet very well as it can together all the geographically dispersed resources such as computing resources, storage resources, data resources, information resources, software resources, knowledge and expert resources and so forth to make Grid environment become a fully shared information system. Grid Computing is a new distributed computing model aimed at complex science computing, which organizes the scattered computers in different geographic locations to form a "virtual super computer" by the Internet. [7-8]

The obvious advantage of the Grid computing Internet is cheap as Grid technology can maximize the value of resources through sharing, collect the idle and wasted resources for Grid users and avoid extra expenses arising from the geographical restrictions, which obviously have cost-reduced potential to users.

3. VIEWPOINT ON SCM BASED ON KM IN GRID COMPUTING ENVIRONMENT

Currently, there are still a lot of deficiencies in the existing Internet-based Supply Chain Management as it can't meet the requirements of dealing with large amount of complex information, data statistics, analytical model analyses and management approaches. Enterprises have to allocate the most amount of system resource to meet the peak computing application or outsource to the specialized enterprises in the traditional distributed computing environment, which results in a waste of resources in the non-peak data processing. Meanwhile, in the traditional distributed computing environment, the distributed resources can only be statically linked while the different types of systems can not be dynamic synergies connected. [9]

SCM based on KM in grid computing environment can solve the above problems so as to meet the demand of logistics management for the enterprise group and real-time query for the remote enterprise managers.

Besides, SCM based on KM in grid computing environment can provide a consistent response to customers, real-time access to information and make good use of the enterprises' resources beyond the internal management (such as ERP: Enterprise Resource Planning) and external management.

Consequently, the enterprises can benefit from a global SCM to connect the manufactures and customers around the core business.

Generally, SCM requires real-time information sharing and effective resources distribution to maximize the profits of the Supply Chain partners while Grid computing can meet the demand precisely, which reflects the properties such as superior processing capability, resources integration, low-cost, virtual union and so forth. So the nature "Supply and Demand" relationship make SCM based on KM in grid computing environment become a hotspot in future.

4. MODEL FRAMEWORK OF THE OVERALL STUDY

4.1 SCOR Model

Supply Chain Operation Reference model (SCOR) is the most influential and widely applied Supply Chain Management model, proposed by Supply Chain Council (SCC). [10]
SCOR model contains some well–known concepts such as Business Process Reengineering, Benchmarking and Process Evaluation which are integrated into a multi-functional framework in SCOR. The framework of SCOR consists of five basic management processes, namely, plan, procurement, production, distribution and return. The Fig.2 below shows the framework of SCOR model.

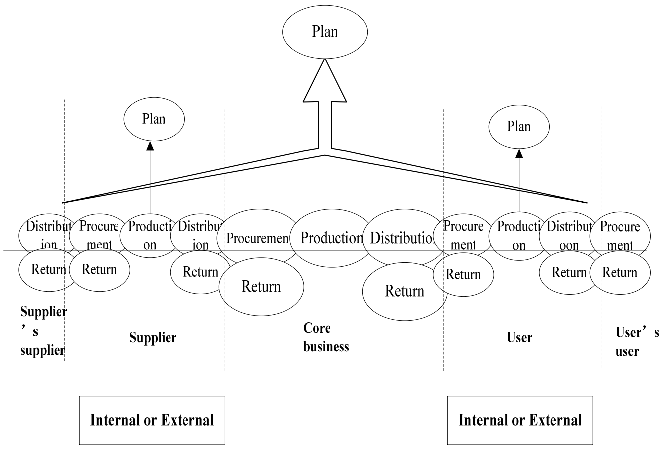


Fig.2. Five basic management processes in SCOR

4.2 Model Framework for SCM Based on KM in Grid Computing Environment

On one hand, SCM arises from integration management. Firstly, it is a functional integration related to the procurement, production, distribution, storage, and other activities in business. Secondly, it is the space integration related to the dispersed suppliers, facilities and markets. Finally, it is inter-phase integration related to the strategic level, the tactical level, the operational level management. On the other hand, Grid's features such as high capacity and optimal allocation of resources can promote the integration of Supply Chain Management effectively.

This paper presents a model framework for SCM based on KM in Grid computing environment due to the advantages of Grid computing as the Fig.3.

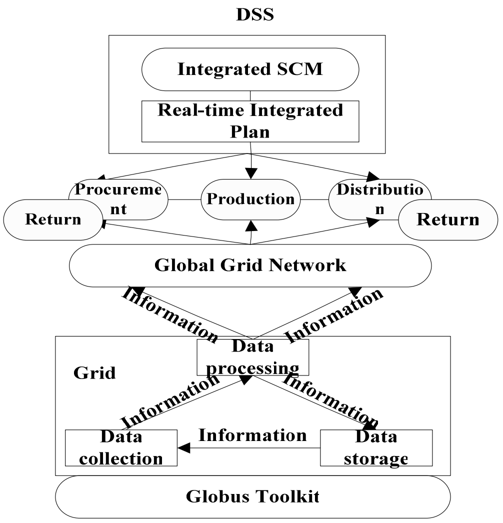


Fig.3. SCOR model framework for SCM based on KM in Grid computing environment

5. OPTIMIZED MODEL OF INFORMATION TRANSMISSION FOR SCM BASED ON KM IN GRID COMPUTING ENVIRONMENT

According to the SCOR model framework for SCM based on KM in Grid computing environment (Fig.3), it is clear that the data transmission is an important point to achieve the effective the integrated Supply Chain Management. This section focuses on the information transmission modeling for SCM based on KM in Grid computing environment.

5.1 Modeling Background

Generally, Bullwhip effect can't be well-solved in the existing Internet environment as the restrictions of the Internet infrastructure platform. But when comes to grid computing environment, these difficulties can be solved more perfectly due to the two main advantages of grid computing, the super data processing capabilities and the ability to make full use of the Internet idle resources. In Fig.3, it is clear that the functions such as planning, procurement, production, distribution and return are not only in the internal enterprise, but also among enterprises, which makes SCM more complex. The Fig.4 shows a typical SCM system. [7] In order to establish an optimized model of information transmission for SCM based on KM in Grid computing environment, this paper will omit the unimportant tips on SCM.

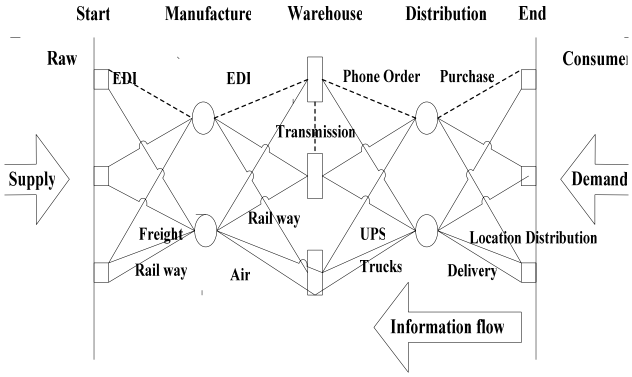


Fig.4. Typical SCM system

In Supply Chain, every chain enterprise needs to share information, including the product information, the inventory information and the sales information etc. Therefore, it is very important to publish and transmit the information as soon as possible.

In this model, it is supposed that there is a communication network in the Supply Chain, transmitting the information among computers in different enterprises. And according to the simple Supply Chain structure model (Fig.1), there are 13 companies, including a core enterprise, two suppliers, two users, four supplier's suppliers and four user's users.

Assuming every enterprise (including user) has only one computer, which is denoted as a vertex in Fig.5. and there are 12 files transmitted among enterprises, which are regards as sides in Fig.5. And for all x and y in Fig.5, T(ex)=1, C(Vy)=1. So the question is how to calculate the corresponding competition time to show the efficiency of information transmission in SCM.

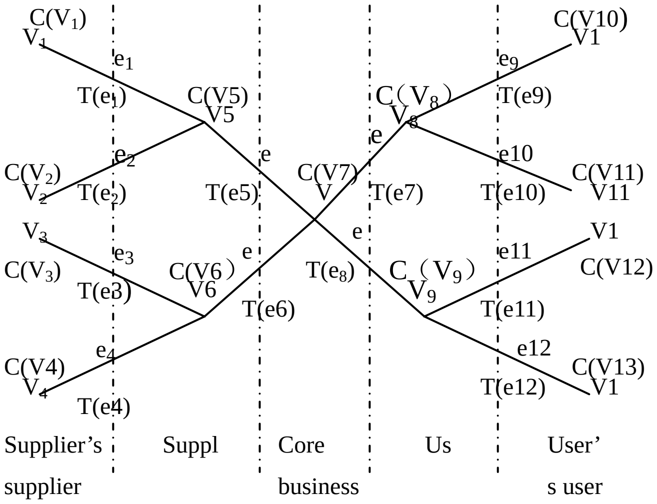


Fig.5. Model in non-grid computing

5.2 Modeling in Non-grid Computing Environment

Assumed the transmission time of each document, T (ex) = 1, and the processing capacity of every computer, C (Vy) = 1, the problem becomes relatively simple to solve.

In Fig.5, the documents (information) are processed batch by batch, so the minimum number of batches shows the competition time.

In the model, there are always some documents been transmitted at per unit-time corresponded to the sides without the same vertexes in Fig.5, called "Matching" in Graph Theory. Different "Matching" contains different sides and all of the "Matching" forms the whole graph (Fig.5). Therefore, the number of "Matching" can delegate the optimal competition time.

Through the above model analysis, it is clear that 4 documents are transmitted in the first unit time, 4 in the second unit time, 3 in the third unit time and 1 in the forth unit time. So the optimal competition time is 4 units.

5.3 Modeling in Grid Computing Environment

In grid computing environment, all Supply Chain enterprises can share the date processing capability and all the information are integrated as a Global Repository in Fig.6, regardless of the information allocation and data types.

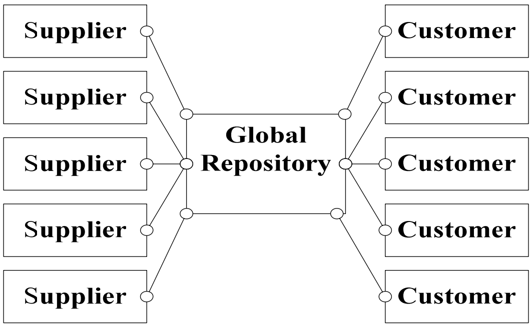


Fig.6. Global Repository

To simplify the model, this paper assumes that the Global Repository is in the core business. So Fig.5 can be changed into Fig.7.

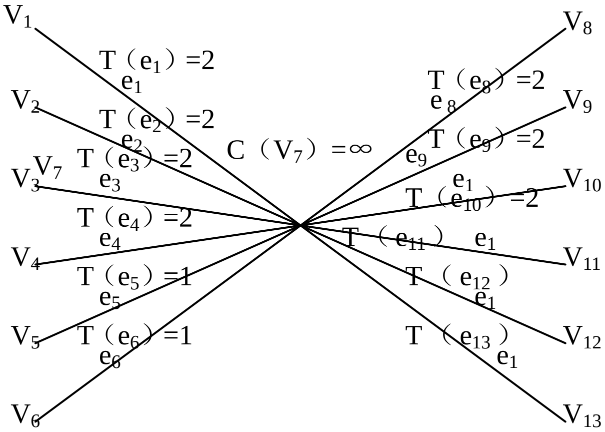


Fig.7. Model in grid computing environment

Using the matching algorithm in Graph Theory, the result is that the optimal competition time is 2 units in grid computing environment.

5.4 Further Model Analysis

In order to facilitate the modeling, this paper used the simplest structure of Supply Chain to establish the model. In fact, the real Supply Chain structure is more complicated. From the viewpoint of further model analysis, the saved time will become increasing as the Supply Chain structure is more and more complicated. And if all the global resources were in the Grid computing platform, the saved time will be shockingly huge.

Meanwhile, as the Supply Chain infrastructure is in the grid computing environment, the resources processing capacity is in the Global Repository (Fig.6). Therefore, the whole Supply Chain can be regarded as "tree" in Graph Theory and the Global Repository can be considered as the "root" of the "tree". Thus, calculating the optimal path of "tree" can get the "Optimal Competition Time". The key code to calculate the "Optimal Competition Time" is outlined as follows:

```
String strSQL;
strSQL = "select   son_id from parent_son where son_id not in(select parent_id from parent_son) ";
……
String cquery="select parent_id ,total_weight from parent_son where son_id=?";
……
String t_son_id="";
float Max=0;
outer1:for(int i=oleaf.length-1;i>=0;i--){
    float ftotal=0; float total_weght;
        t_son_id=oleaf[i].toString();
        int insert_sequence=0;
        ps=conn.prepareStatement(cquery);
         outer2: for(int j=0;j<2;j++){
            ps.setString(1,t_son_id);
            rs=ps.executeQuery();
            if(rs.next()){
                   t_son_id=rs.getString("parent_id");
                   ftotal=ftotal+rs.getFloat("total_weight");
                   j=0;
                   continue outer2;
              }else{};
                   ototal[i]=ftotal;
                   continue outer1;
          }
        }
……
for( int m=0;m<oleaf.length;m++){
      if(ototal[m]>Max)    Max=ototal[m];
}
```

Code.1: The key code

5.5 Model Comparison

Through the above analysis and comparison, it is concluded that the optimal competition time is 4 units in non-grid computing environment (Fig.5) while 2 units in grid computing environment (Fig.7), which means SCM based on KM in grid computing environment takes much less time than non-grid computing environment. As the information transmission and processing time can be reduced sharply in grid computing environment, the "bullwhip effect" in SCM can be solved more effectively.

Besides, the grid computing environment can integrate all the computing resources into a Global Repository so as to make full use of the idle resources, share more information and reduce the system costs.

6. CONCLUSIONS

This paper proposed the SCM based on KM in Grid computing environment from the viewpoints of both theory and model, especially, focused on research on the optimized model of information transmission for SCM based on KM in Grid computing environment. The advantages of SCM based on KM in Grid Computing environment includes efficient data-processing capabilities and fully use of idle resources, which is suitable for the development of SCM in the future..

REFERENCES

[1] Shihua MA, Yong LIN and Zhixiang CHEN, *Supply Chain Management*. Beijing: Machinery Industry Pub, 2000.(Chinese)

[2] Gunasekaran, A, "Supply chain management: Theory and applications", *European Journal of Operational*

Research Volume: 159, Issue: 2, December 1, 2004, P. 265-268.

[3] Wu yingliang,Xiao wancheng,Wang shujun,Qian jiannong, “self organization analysis of supply chain knowledge management system”, *System science journa*l，Vol.14 No.3,2006.07，P：83－88.

[4] Wang tao,Ru yihong, “knowledge management of logistics enterprises”, *China material circulation* , 2001.06 P:21-22.

[5] Lu dan, *research on implement of supply chain knowledge management*, Master’s degree paper.

[6] Zhang Mi, Xu Jianjun, Cheng Zunping, Li Yinsheng and Zang Binyu, “A Web service-based framework for supply chain management”[J],*Object-Oriented Real-Time Distributed Computing*, 2005. *Proc. ISORC 2005*, 18-20 May 2005, P. 316 – 319.

[7] Runtong ZHANG and Nin FAN, *Grid is Computing* Beijing: Tshinghua University. Pub. 2006. (Chinese)

[8] Zuomin LUO, Jing Zhang, Junhuai LI and Changsheng XIE, “A Survey of Grid Computing and Key Technologies” [J], *Computer Engineering and Application*s, Issue: 30, 2003, P. 18-22. (Chinese)

[9] Xinjun ZHAO, Guozhen TAN and Xunyu WANG, “Research on the Model of Supply Chain Management System Based on Grid Computing” [J], *Application Research of Computers*, Volume: 21, Issue: 4, 2004, P.82-84 (Chinese)

[10] Hui JIE, Peiqing HUANG and Cunlu ZHAN, “Supply Chain Modeling Method Based on SCOR Model” [J]. *Industrial Engineering and Management*, Volume: 2, Issue: 2, 2004, P.11-13, 22. (Chinese)

Tian Lan is a PH.D.student in the School of Economics and Management at the Beijing Jiaotong University in Beijing, China.,whose major is Management Science.

Dr. Runtong Zhang is a Professor in the Department of Information Management at the Beijing Jiaotong University in Beijing, China. Meanwhile, he is also an Invited Senior Researcher at the Group of Advanced Internet Technologies at the Nokia (China) R&D Center in Beijing. His research interests cover next generation Internet and communication networks, artificial intelligence, grid application study, electronic commerce and their applications. He has published more than 60 papers in refereed journals, like IEEE Transactions, and conferences. He is also an author/co-author of 6 books and a holder of three patents in next generation Internet.

Shengbo Shi is a graduated student in the School of Economics and Management at the Beijing Jiaotong University in Beijing, China, whose major is Management Science.

An Open Digital Library Grid Architecture

Jinbo Chao, Qiushuang Jing, Fuzhi Zhang, and Zhuang Liu
College of Information Science and Engineering, Yanshan University, Qinhuangdao, 066004, China
Email: jingqiushuang81@163.com

ABSTRACT

With the growth of the quantity of digital library, the problem of Interoperability is more and more prominent. This causes more and more appearance of solutions, but there are weak points more or less. After comparison, OAI-PMH is better, through gathering, collection, organization of metadata, it has achieved the goal of interoperability. Now the technology of grid is the most advanced technology, its introduction strengthened interoperability from the resources sharing angle, fundamentally has solved the question of seamless retrieval about the distributed, autonomously, isomerism digital library.

Keywords: Digital Library Interoperability OAI-PMH Grid Metadata

1. INTRODUCTION

Now the library has already successfully transited from the traditional library to the digital library (DL). The readers needn't to run between each library to search for several pages of materials again. You can only sit in front of your desk, glanced over library's website with rich numeral collections. But along with the digital library increases day by day, information-user also meets many questions when use digital library. Several aspects mainly manifest in the following[1]:

(1) Under open environment, different organizations separately use the different software and hardware platforms to establish respective digital library systems and the collection resources separately. Because the architecture of system and Technology are different, it is very difficult to correspond mutually and to share resource for these DLs.
(2) In order to obtain the material which is need, users have to submit identical request to many different DLs repeatedly.
(3) At present the majority of library information inquiries are based on the data-driven (In Inter OP project, Old Dominion university promote the framework of Lightweight Federated Digital Libraries-LFDL), the present existing search engines cannot to establish its content index. Regarding to the users who glance many DLs to search for materials, the search engine cannot satisfy their demands [2].

Therefore, for searching materials of a subject, the readers must adapt to different style search contact surfaces and different information expression forms. The identical inquiry request has to be submitted separately to each DL by different forms. They have got rid of the backward information retrieval ways, but still don't know how to proceed, in the voluminous information sea.

How to get the widespread distributed, autonomous, isomerous DLs conformity on internet and to shield their internal differences, to form a unity, transparent services to the users, to realize DLs' interoperability have become hot spots which DLs' constructor researched and practiced.

In DLs' domain, Interoperability usually is used for specifically describing the ability of exchanging and sharing documents, inquiring and service between each module in one DL or between different DLs[3]. Interoperability is the question essentially follows different systems and architectures appearance.

2. THE METHODS OF INTEROPERABILITY OF THE CURRENT DLS

The core question of interoperability lies in finding appropriate method to urge each independent DL to cooperate together.

At present, the methods of interoperability are mainly divided into three types according to the cooperation degree [4].
(1) Federation: Different DLs coordinate according to the common standard (such as Z39.50) when they are constructed, achieve mutual recognition depend on data sharing, all have special systems to realize interoperability. In this kind of federated system, the seamless is good, but the cost is big than others, so this method suits for massive participants.
(2) Harvesting: Each library carries on small transformation to be able to achieve the goal of interoperability according to the gathering protocol (such as OAI-PMH). It needs not to define completely the same standard, while community is loose. Therefore, this method suits massive DLs to participate.
(3) Gathering: Each library needs not to have any standard. Users may retrieve through the entire Internet so long as the information is public. The Web search engine is a best example. Gathering include the information members on the most wide range, even may contain the whole Internet, but it is obviously the grade of service makes us anxious.

The three interoperability models above are located in different spots in the functionality-to-cost curve [5] (fig.1). Federation can realize the formidable function of interoperability, but its cost is very high, only can be used in small scale federation. Gathering nearly not to make any limit to the members, its cost is low, but the function is very weak. It may be used by the all users in Internet. Gathering, located between both, is a compromise method. Although it did not have rich function compared with federation, but the cost reduced very much. Thus it is accepted quite easily by large-scale users. Although compared with Gathering, the cost enhanced, the grade of service also had been large improvement obviously.

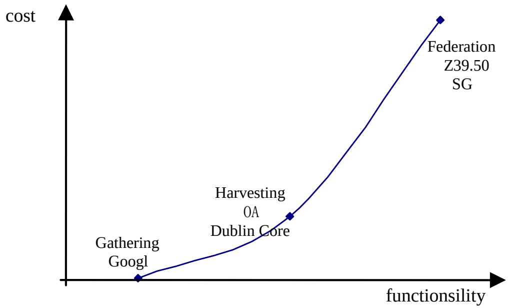


Fig.1. The functionality-to-cost curve of interoperability

In brief, it is cannot find a best spot forever in the functionality-to-cost curve. That is to say, we only can find one spot to suit some kinds of demands, but cannot find a spot to suit all demands. Our goal is to strengthen the function of our system, reduce cost, and achieve a higher level of interoperability through using new technology. And each kind of interoperability strategy we proposed must be face to reducing the cost and enhancing function. It displays the movement of downward and right in the curve, increases the bending of curve. OAI is one solution located in the middle of the curve. I place the key point in the research of OAI in the following part.

3. OAI-PMH PROTOCOL OVERVIEW

OAI (Open Archives Initiative) is a cooperative organization for the purpose of promoting development, issuance and share of network information resource. The website is Http://www.openarvhives.org. It was initiated by American Library and Information Resource Committee Members (CLIR), digital library alliance (DLF) and so on in October 1999. The plan of OAI tentative was proposed in a conference which was hold in Santa Fe in New Mexico for the first time. At first OAI is to solve question of the electronic periodical pre-printed book interoperability and metadata harvesting. After the development of more than one year by researchers, OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting, the first operable edition of OAI) has promoted in January 2001. The protocol of OAI-PMH1.0 version may be engaged in the network content publication for every kind of the organizations and any network servers. After more than one year's tests and feedbacks, the researchers in OAI organization promoted the second edition OAI-PMH2.0 in June 2002. The births of OAI-PMH1.0 and 2.0 have become two milestones in the OAI history [6]. At present the OAI we said refers to OAI-PMH2.0. The target of OAI is to support to the scholarly value metadata search. The goal of the OAI-PMH issued is to realize metadata interoperability between different information organizations on Web through metadata search, provides an interoperability frame which is nothing with the application.

4. GRID TECHNOLOGY OVERVIEW

The thought of grid originates from the electric power grid, the goal is transmitting the calculating ability and the information resource to the users conveniently as the power network through the network [7]. The grid function is making the high speed Internet, the high performance computer, the large-scale database, the sensor, the long-distance equipment and so on merge into one organic whole, to make sure computation resources, memory resources, information resources, knowledge resources and so on are shared comprehensively, then to eliminate the informational isolated island and resources' isolated island [8]. The goal of grid technology and the goal of the digital library technology is the same essentially. They all must solve the questions of the vast area, isomerism informational sharing, the interconnection and the interoperability. Therefore, the next generation of digital library research takes the application of grid seriously, attempts to construct the new integrated digital library system by using the grid technology, realizing the integration of a network information resources in time, space and content, to satisfy the demands of information from all walks of life in society [9].

According to the grid technology thought, packing the distributed, autonomous, isomerism digital library system into information node in the digital library grid, under special mechanism management, constituting the unified resources service system, making the users to use the information resources in the digital library group transparently through the grid gateway, it is no doubt that these will be the most charming and developing way to realize the digital library resources interoperability. The superiorities of grid technology solving digital library interoperability questions lie in:

(1) Taking massive mature, open technology and standard as the foundation, it is easy to realize and expanse, be accepted and promoted easily by multitudinous libraries.
(2) Surmounting platform irrelevant characteristic of isomerism system, it realizes thorough resources sharing.
(3) Possessing the consummate safety control mechanism, the security and reliability can be guaranteed.

The OAI-PMH frame realizes digital library interoperability through metadata interoperability, overcomes the scale question which the distributional search is unable to solve. But the differences of grid computation technology and traditional distributed computing lie in that, it pays attention to large-scale resources sharing and the cooperation usage among the multi-organizations, provides the essential method of resources sharing, has special advantages in the solution of isomerism platform compatibility and existent system integration [10]. We propose a new frame enhancing digital library interoperability – ODLG (Open Digital Library Grid), through unifying the grid technology with the metadata acquisition method. That is, we introduced the grid thought in the original OAI-PMH frame foundation to solve problems of library resources discovery, conformity, cross storehouse retrieval, and overcome the limitation of traditional digital library interoperability scheme, realize digital library group in the entire internet.

The key to understand the OAI technology is to understand two roles that "Service Provider (SP)" and "Data Provider (DP)". In fact, SP and DP may be the different organizations, also may belong to the identical organization, DP is a database possessing massive metadata, but SP extracts metadata from one DP or many DPs, and provides the increment service (Search, browse and so on). OAI-PMH connects DP (data providers) and SP (service providers) directly, simply realizes the metadata collection on http.(Fig.2) OAI-PMH protocol transmits data based on HTTP protocol, metadata is described by XML, this makes OAI-PMH protocol can be used extensively in the Web environment, and realizes information exchange among distributed isomerism resources effectively.

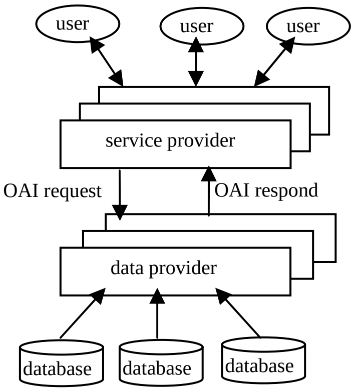


Fig.2. OAI protocol model

5. INTEROPERABILITY FRAME BASED ON GRID ARCHITETURE OAI-PMH

Open digital library grid frame is composed of three parts. The bottom is the data level, composed of many digital libraries in different regions, as the DP in the Interoperability digital library, provides the metadata stipulation by OAI-PMH, forms Metadata Repository follows the OAI-PMH; the uppermost layer is application layer of the hypothesized digital library, it provides the unification service connection for the user, provides services retrieval inquiry and so on; Between the data level and the application layer is the grid level, it has shielded distributed and isomerism characteristic of the bottom digital library, provides the unification the service connection for the application layer through metadata discovery, collection and the overall situation index

In order to strengthen the collection and the index dynamic performance, to speed up update speed of metadata, we can introduce three kinds of grid nodes in DL Grid architecture: Harvest Scheduler Service node, Metadata Harvest note and Metadata Collection Node. In the past, OAI-PMH directly connected DP and SP, simply realized metadata collection on http, but now the harvester node must collect metadata through Grid. Through uses the Grid computation node, it may increase the service provider's quality, support the high performance the search service of OAI federation. [11]

The work steps of ODLG (Fig.3): Through the open grid platform, the digital library grid has shielded the digital library distributed and isomerism, completed metadata discovery, metadata collection the overall situation index between the geographic distribution digital library, provided the consistent information service for the user [12]

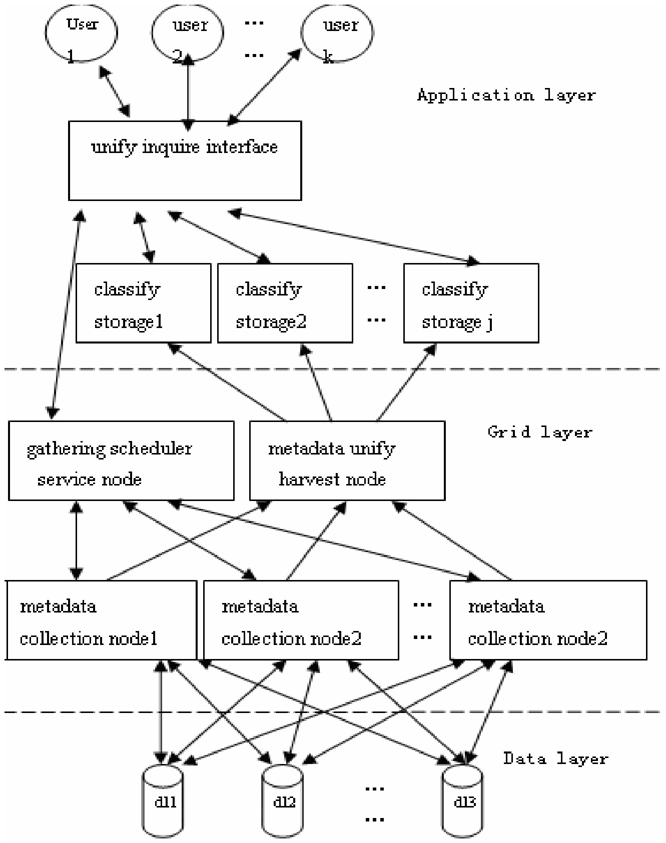


Fig.3. Interoperability model of digital library based on grid

a. Gathering scheduler service node assigns the task of gathering metadata to metadata collection nodes, and provides the status messages to the metadata unify harvest node.
b. Once a DL is assigned to a metadata collection node, the metadata collection node will use the OAI-PMH protocol to do the gathering task.
c. The metadata in different digital libraries will provide its messages to the metadata harvest node.
d. Metadata unify harvest node start gathering task, harvest all metadata which is gathered.
e. Metadata unify harvest node will organize all metadata according subjects, and sort to different classify storages.
f. After gathering, harvesting, storage tasks of metadata are finished, gathering scheduler service node assign the tasks to nodes again, and the next work is beginning.

Metadata is saved in a metadata storehouse, then it can provide the consistent search surface through the unification inquiry surface for user, transmits user's inquiry request to corresponding metadata storehouse, collects inquiry result, and submits to the end-user.

6. THE SUPERIORITY OF GRID AND OAI-PMH UNIFY

OAI-PMH achieves metadata interoperability, overcomes the problems of scale which the distributed search is unable to solve. And grid pays attention to large-scale resources sharing and the cooperation applications between the multi-organizations, has provided the essential method of the resources sharing, has superiority in solving heterogeneous platforms and integrating existing systems.

(1) In the process of metadata collection, if a collocation node failure, Task management and scheduling nodes will use grid FTP to dynamic pass collect code to this nodes code, take the collection task in this node, guarantee the normal collection.
(2) At present, sharing systems based on the OAI-PMH almost use centralism saves (such as NDLTD). We May use the concept of information reorganization under the grid environment, distributing save the metadata according to contents. Like this has oversized avoided the question of data oversized and long retrieval cycle which centralism saves brings, and Enhanced the retrieval efficiency.
(3) At present, OAI-PMH mainly adopts two methods to solve metadata origin's question: a. To imitate search engine, this method adopts search engine like reptile to dynamic seek metadata which is hided in the homepage. b. To establish enrollment and registration's system which is metadata's originate. To provide on own initiative conforms to the protocol standard metadata to carry on the enrollment and registration; and concentrates in the OAI official website, lists these data supplier's the website. These two methods not only have advantage but also have disadvantage; How provides the high grade metadata resources to the service provider will be our key work in the future.

Under the grid environment, the resources gain is realized through the resources registration. In Globus there is an important module MDS (Metacomputing Directory Service, MDS). It uses LDAP (lightweight directory access protocol) as the unified connection of information, manage each kind of resources and service in grid environment, provides the effective convenience directory service. Its function includes the information discovery, the registration, the inquiry, the revision, the cancels and so on. It provides a set of tools and the procedure connection to use in discovering, issuing and accessing each resource and information in computing grid. In MDS, the resources carry on the dynamic registration using grid resources registration protocol (GRRP), in high level, service may obtain the related information through grid resources inquiry protocol (GRIP), and carries on the conformity to the information.

7. CONCLUSIONS

At present, digital library faces the major issue and the essential technology－interoperability. Until now solutions although achieved some results but still have the limitation. The introduction of grid technology has happen to made up deficiencies. It has solved the high performance resources sharing and harmonious cooperation work in the network environment, eliminated the information isolated island and the resources isolated island, solved the problem of interoperability of digital library.

REFERENCES

[1] Luo Xueming. “Study on the Grid Architecture of Digital Library.” *Information Science*, 2006,(7):1045-1048.

[2] Xing Xiaochun, Zeng Chun, Li Chao, Zhou lizhun. “Research on magnanimity information manage architecture faced on digital library.” *Journal of Software,* 2004,(15):76-85.

[3] Blaze M, Feigenbaum J, Lacy J. “Decentralized Trust Management.” In: *Proc. of the IEEE Symposium on Research in Security and Privacy*, *Research in Security and Privacy*, Oakland, CA, May1996. IEEE Computer, Society, Technical Committee on Security and Privacy, IEEE Computer Society Press.

[4] Kurt Maly , Mohammad Zubair , Xiaoming Liu. Kepler – “An OAI Data/ Service Provider for the Individual.” *D-Lib Magazine*.April 2001. Volume 7 Number 4

[5] Yang deting, Yan Baoping. “Metadata Interoperability and the Open Archives Initiative.” *Application Research of Computers*, 2004(1):44-47

[6] Carl Lagoze and Herbert Van de Sompel . “The making of the Open Archives Initiative Protocol for Metadata Harvesting.” *Library Hi Tech*. Volume 21 Number 2. 2003

[7] Guo Hong. “Construct a digital-library intelligence service platform based on grid technology,” *Modern information*, 2006.(7):87-89.

[8] Foster I, Kesselman C. *The grid: Blueprint for a new computing infrastructure* [M].San Francisco: Morgan Kaufmann Publishers, 1998.

[9] Francine Berman,Geoffrey C.Fox,Tony *Hey.Grid Computing:Making the Global Infrastructure a Reality*,Wiley,New York, 2003

[10] Foster I, Kesselman C, Tuecke S. “The anatomy of the grid: Enabling scalable virtual organizations” [J]. *International Journal of Super Computer Applications*, 2001, 15(3):200-222.

[11] ZhengZhiyun, Bi Lepeng, Niu Zhendong, Song Hantao. “The Framework for Interoperability of Digital Library Grid.” *Computer Engineering and Applications*. 2005,(25)186-189.

[12] PanHao. “Architecture of the Digital Library Interoperability Based on the Grid Technology.” *Journal of information*. 2006, (7)10-15.

Study on Modeling of Multi-resource Base in Manufacturing Grid*

Yongfeng Li[1, 2], Buyun Sheng[1], Jianhua Jiang[1], Fei Tao[1]
[1.] Wuhan University of Technology, Wuhan 430070, China;
[2.]Taizhou University , Linhai, Zhejiang 317000, China
Email: [1]lyfisgerly@163.com

ABSTRACT

In order to realize the integration and sharing of resources such as services publishing center, legacy database, data market, data [1] warehouse in manufacturing grid, the concept of manufacturing grid multi-resource base was proposed. And the requirement of designing resource base under manufacturing grid was analyzed as well. A model of multi-resource base in manufacturing grid was established and some key technologies in the model were discussed.

Keywords: Manufacturing Grid, Multi-Resource Base, Modeling

1. INTRODUCTION

Manufacturing Grid (MGrid) as an advanced manufacturing technology is developed based on the development of the grid technology. Its goal is to realize the cooperation and sharing of resources, so as to maximize the benefits of both resources providers and users. At present, scholar’s research mainly focused on manufacturing grid concept, architecture, connotation, application prospect, and anticipated results etc.

In MGrid, all resources are described in the form of the service. With the in-depth study, each enterprise has established individual service publishing center. The centers are established and developed according to different protocol specifications, different developing platform, and different encapsulation pattern. Therefore, it is difficult for these services to be discovered, searched, and accessed etc. One of the current solutions is to assemble each service. However, due to lack of semantics and the inference support, the service assemblage still stays in the theoretical stage. We can take the question into account: how can we integrate and share these services in a higher level in the environment including services publishing center, traditional database, data market, data warehouse and other storage methods? Taking account of this idea, and combining the theories and methods of distributed database, multi-database, and data warehouse, the concept of MGrid multi-resource library (MGMRL was purposed. Finally, some key technologies in the model were investigated.

2. REQUIREMENT OF ESTABLISHMENT OF MGMRL

It is the core part of modern manufacture to establish the manufacturing resources database.

Based on the characteristics of the resources in MGrid, establishment of MGMRL should meet the following requirements:

(1) Distribution. These major libraries of storage resources in MGrid allocated in different enterprises, and dynamically add to the whole network as a node. The resources storehouse of various members is more likely to work in a distributed environment.

(2) Intelligence. In order to realize “demand-driven customization” and “real-time customization”, the management of resources has to be intelligent. The model of the MGrid resource should be analyzed and deduced instantly by the demand of user, and search the best resource and the shortest route to access in MGrid. This routing information will be recorded in the resource database to facilitate next visit. Meanwhile, the model may deal immediately with the errors in the processes of data access in order to maintain a smooth link and the robustness of the platform of manufacturing grid resource.

(3) Heterogeneity. The resources in MGrid are owned by different enterprises or individuals. The sources of the resources have some heterogeneity in environment and resources faces. Heterogeneous environment is mainly shown in the different platform that the sources of resources runs such as different hardware platforms, software system platform. Heterogeneous resources are mainly shown in different data model and the data semantics. The construction of the resources library in MGrid should hide these differences from the global.

(4) Autonomy. The provider of various sources of resources can manage their own data or resources independently, decide the degree of sharing and cooperate with other members independently, encapsulate their own data into services in order to publish them into the resources storehouse of the MGrid, update data and services dynamically, and change the state of data and services.

(5) Layer. The resources library in MGrid should be divided into different layers according to the demand of users, and establish different view. Because different types of data or resources published into the resources base lead to the complex types of services, it is necessary to construct a model of resource classification, and form a multi-level model of the resources services. This model should be able to reflect the establishment of resources library in MGrid.

3. CONSTRUCTIONAL MODEL OF MGMRL

The multi-resource base in the manufacturing grid has the characteristics of distribution, autonomy and heterogeneity, and the source of the data and resource in multi-resource base are both services. Therefore, a architecture of manufacturing grid multiply resource base is given in the manufacturing grid. The architecture is shown in Fig.1. The architecture referencing on the mode of B/S is divided into six layers: local resources layer, local resources library management layer, global resources layer, global resources library management layer, resources library view layer and resources application layer.

*Supported by the key technologies research and development program of Wuhan, China (20055102021); national natural sciences foundation of China (50620130441)

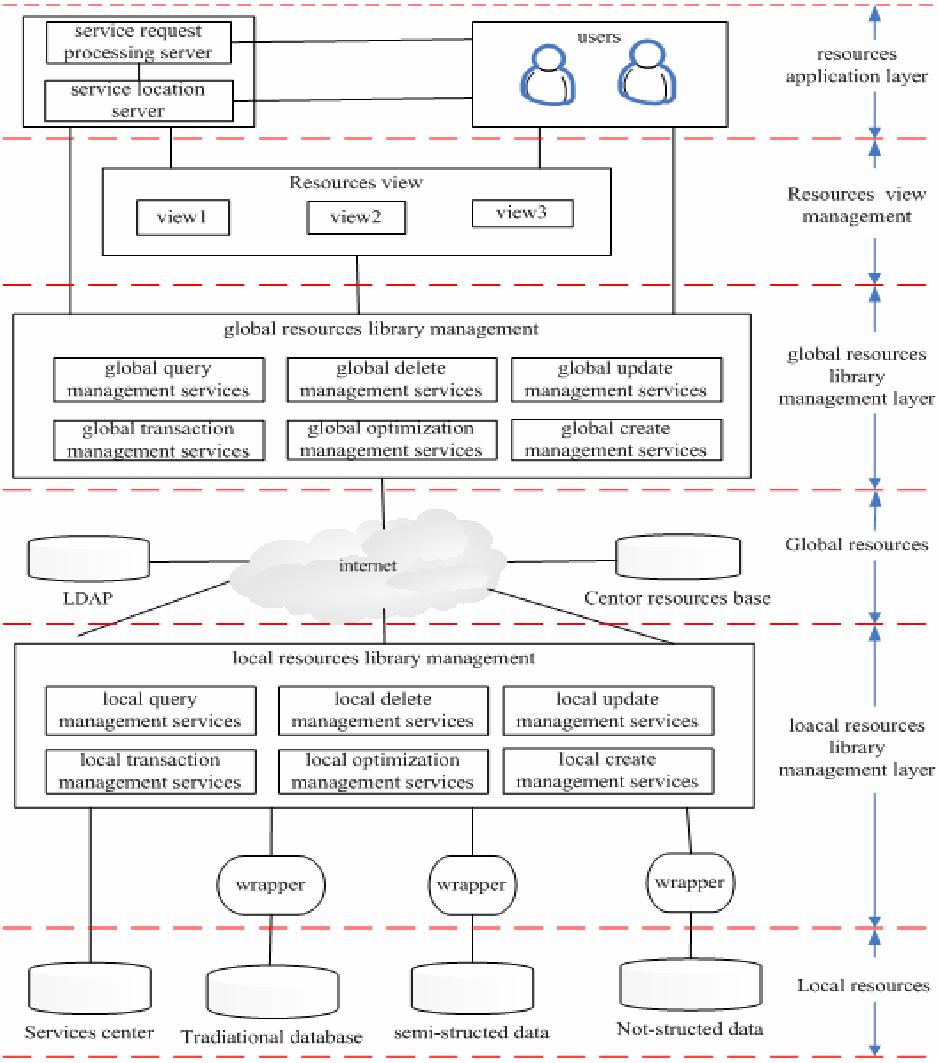

service request processing server
service location server
users
resources application layer
Resources view
view1
view2
view3
Resources view management
global resources library management
global query management services
global delete management services
global update management services
global transaction management services
global optimization management services
global create management services
global resources library management layer
LDAP
internet
Centor resources base
Global resources
local resources library management
local query management services
local delete management services
local update management services
local transaction management services
local optimization management services
local create management services
loacal resources library management layer
wrapper
wrapper
wrapper
Services center
Tradiational database
semi-structed data
Not-structed data
Local resources

Fig.1. the framework of manufacturing grid multiply resource library

(1) Resources application layer. It includes two components: service request processing server and service location server. The former accepts the request of user and deal with its requests. The latter accepts the result which the service request processing server carries out, and transports the locating information corresponding to operating on the resources library into global resources library management. In the meantime, the latter also accepts the view information downward and locates the user corresponding to the view.

(2) Resources base view management. It can made user access to data or services transparently, and meet the individual demand of different users. In order to improve the efficiency and performance of resources, the information queried by global resources library management is cached, and transport the information into services location server. The information fed back by the services location server is transferred into individual users.

(3) Global resources library services management. Its main function is to manage the global resources library. It includes global query management services, global delete management services, global transaction management services, global update services and global optimization management. Global query management server deals with the information transported by upper services location management, and optimizes the queried information, then the classified information is passed into location resources library services management through the network. Finally, the results of querying are collected and transported into resources library view management. The function of Global deleting management services and global update services are similar to the global query management server, therefore the details will not be discussed in this paper. Global transaction management is a supplementary of other services, which mainly deals with transaction resources and maintains the ACID（Atomicity Consistency Isolation Durability, ACID）of global resources library. Global optimization management server is also the supplementary of other server, maintaining the whole property handled the global resources base services management, the central repository and indexing resources LDAP(Light directory Access Protocol, LDAP).

(4) Global resources layer. It includes global center resources base and indexing resources base LADP. Global center resources stores the data of various resources library sharing. LDAP, which is to improve the efficiency of indexing, to store the information of the local resources management in the form of meta service in order to increase the rate of service location.

(5) Local resource library services manager. It includes local query management services, local delete management

services, local creation management services, local transaction management, local update services, local optimization management services and wrapper. The functions of these servers are similar to global resources services management. Wrapper is mainly to shield off the heterogeneous of lower database and integrate the legacy database system.

(6) Local resources layer. It defines a number of local resources or database, including the deposit of autonomy, heterogeneous database system, structure and semi-structured data. The wrapper of local resources services management layer defines the corresponding number of wrapper according to the number of the sources of local data. The latter, in the hand, will integrate the resources and model information of source of local resources into available resources for serving the upper services; in the other hand, the latter also accepts the information or requirement from the upper and analytic transmitting into the lower.

4. SEVERAL KEY TECHNOLOGIES OF THE MODEL

The aim of the construction of MGMRL is to shield off the differences of the various databases and services , and enables users to visit the required resources at any time. According to the concept, MGMRL is more distributed, more dynamic, more expansibility multi-database system. Therefore, we can discuss the key technologies of MGMRL based on the key technologies of multi-database. These key issues include unified data model, unified data manipulation language, conflicting resolution, distributed query and transaction management. However, the key technologies of MGMRL has its unique characteristics.

4.1 Unified Resources Pattern

The information of Multi-database usually comes from various database. Multi-database integrates the information in logic into a unified data pattern. This patter is established as a bottom-up approach by a specific application. However, in manufacturing grid, source of data is dynamic and it must support the different application. Meanwhile, sources of data themselves are an integrated framework of resources. Therefore, the methods of establishing the unified data pattern in multi-database are not suitable to establish the multi-resources library under MGrid. Nevertheless we can divide the members of MGMRL into different virtual resources library in accordance with certain rules. And then the members of virtual resources library establish unified framework of resources pattern in accordance with certain rules. Communication of these virtual resources is done by autonomy gateway. In the meantime, the global pattern frameworks of multi-resources library are established by the different or similar methods in virtual resources library. However, how the standards and methods is established poses to be a key issue to be solved

4.2 Distributed Query

Query of data of multi-database is usually inquired by key words. And the query is static in general. However, the data in MGrid are dynamic. They can join, suspend or withdraw from service because of certain reason at any time. Therefore, the data sources of MGMRL have the state themselves. These systems should be able to capture the state and be ready to react quickly in order to maintain the robustness of the systems, and the rapid, accurate, and high-quality information of services user querying are returned. Because of high autonomy of different source of data, a set of mechanism monitoring the change of resources is proposed when sources of data are wrapped into service, which can make high-level and real-time monitoring of the change of resources, can make faster and more accurate control, and can realize the continuity of inquiries resources. However, how to monitor the change of services, to schedule and optimize the services in MGMRL dynamically is a key issue to be researched.

4.3 Transaction Management

The objects which MGMRL manages are services. The services, including the resources that traditional data resources are encapsulated into services and the resources that are established by standard agreement of grid, are highly dynamic and autonomous. They can be in and out of the manufacturing grid platform freely abiding by certain right and lie in different local networks. Therefore, the conditions lead to many uncertain factors and complexity of transaction management in MGMRL. Serious ACID needs to meet the following requirements: services coordinator has the right to completely control the services participants and the established application system should be a tightly coupling. The requirements are not available in MGrid. Therefore, a possible conflict problem of transaction may take place. How do we remove these conflicts? All this has created a lot of difficulty in transaction management of MRMRL.

4.4 MetaService

Under the network environment, data are organized, saved and managed through metadata. But under the manufacturing grid environment, it is necessary to integrate many distributed service node and provide transparent services for users. Because in MGrid all the resources are services. Therefore the concept of MetaService is proposed in paper[11]. Namely metaservice is about service of the service. Resources service is the service which provides in the form of the grid. Metaservice is the service which locates between the grid application and the grid service and serves to the resources service or the meataservice itself. But in this article the metaservice is focused on the service dispatch in order to solve the width of network band and the bottleneck question of visiting. But the metaservice mentioned in this article includes not only the functions above, but also the functions of indexing, self-organization, convenience storage and searching. Therefore, the establishment of MGMRL must take the questions of organization, the storage and the management of metaservice into consideration.

5. CONCLUSIONS

Based on the present situation and tendency of the development of MGrid and the database, the paper, standing on the high level, outlined the form of the organization of the magnanimous resources storage (including the center of publishing service, traditional database, data market, data warehouse and so on). The model of MGMRS was proposed, and the meaning and the key content of the model were discussed

REFERENCES

[1] Fan Yushun, Zhang Liqing, Liu Bo, “Networked Manufacturing and Manufacturing Network,” *China Mechanical Engineering*, Vol.15, No.19, 2004, pp.1733~1738.

[2] Foster I, Kesselman C, Nick J, Tuecke S, “The

Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration," 2002.

[3] Wang Aimin, Fan Liya, Xiao Tianyuan, "Research on application platform and virtual enterprises modeling for manufacturing grid," *Chinese Journal of Mechanical Engineering*, Vol.41, No.2, 2005, pp. 176~181.

[4] Yan Bo, Huang Biqing, Zheng Li, Xiao Tianyuan, "Research on grid and its application in manufacturing industry," *Computer Integrated Manufacturing Systems*, Vol.10, No.9, 2004, pp. 1020~1030.

[5] Liu Lilan, Yu Tao, Shi Zhanbei, "Research on Rapid Manufacturing Grid and Its Service Nodes," *Machine Design and Research*, Vol.19, No.5, 2003, pp. 57~ 59

[6] Fabio Baroncelli1, Barbara Martini1, Luca Valcarenghi2, "Service Composition Model for Automatically Switched Transport Networks," *IEEE*, March, 2005.

[7] Minhong Wang, William K. Cheung, Jiming Liu, "Agent-Supported Web Service Composition for Supply Chain Management," *IEEE*, 2006, 3.

[8] Han Weihong; Jia Yan; Wang Zhiying, Yang Xiaodong, Key Techniques in Multidatabase Systems, computer engineer & science, Vol.21, No.6, 1999, pp. 49~52

[9] Liu zhao, "Research on the multi-database based on networked services," [Master paper].Wuhan: Huazhong University of Science and Technology, 2004.

[10] Ren Hao Li Zhigang Xiao Nong, "Database Grid:the Multi-database Built on the Grid," *Computer Engineering and Applications*, 2002, pp. 171~175.

[11] Zhihui Du,Francis C.M.Lau,Cho-Li Wang, "Design of an OGSA-Based MetaService Architecture Springer-Verlag Berlin Heidelber," 2004,pp. 167~174.

The Research and Application of Grid Information Service Based on Campus Data Grid

Fengying He, Xiufeng Jiang, Meiqing Wang
College of Mathematics & Computer Science, Fuzhou University, Fuzhou Fujian 350002, China
Email: fengyhe@163.com

ABSTRACT

The Information Service is one of the key components of grid system. In this paper, we design a system of CDG based on MDS4 and analyze the implementation of each modules of information service, finally the experiment results are shown.

Keywords: Index Service, Information Service, Monitoring and Discovery System, Quality of Services(Qos), Grids

1. INTRODUCTION

The grid system is an environment integrating various resources. Under the grid environment, the share and the coordination among resources and services can be obtained effectively. At present in the campus network, each kind of distributed computation resources and the education resources are mainly at dispersed state taking the institute as the unit, simultaneously in the campus network, there are many different systems which lack the unification standard, all these make it difficult to integrate resources and realize resources shared among different systems. In order to speed up the process of the campus network digitization construction, making unification deployment and unification management between shared data, computation resources, stored resources and application resources, we put forward the computation environment of campus grid platform based on the grid technology, through realizing the campus grid system that aims for service shared to solve the problems existing in campus network.

2. THE ARCHITECTURE OF THE CAMPUS DATA GRID

Based on the Globus Toolkit 4 (GT4), the Campus Data Grid(CDG) realize resource shared through combining various distributed information and computation resources into the single virtual environment. The architecture of the CDG system is made up of three layers: resource layer, grid middleware layer and resource layer, as shown in Fig. 1.

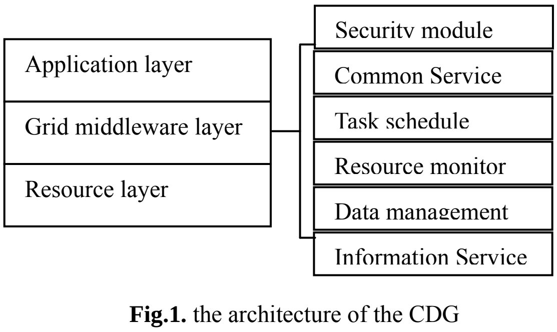


Fig.1. the architecture of the CDG

In Fig. 1, the resource layer is a set of all distributed resources in the CDG, mainly including of calculation resources, database resources and educational resources which distributed in each department of a university. Each grid node in the resource layer offers a unified interface which realized communication between resource layer and grid middleware layer, thus hiding the different construction of various resources.

The grid middleware layer is very important for users to use every kind of service resources in grid transparently. It includes of the following six service modules:

1) The module of information service: it is the key module of the grid middleware layer. The module of information service supports the find and submission of the data, resources and services in grid. It can provide a real-time reflection of the dynamic changes of resources and services. It is the foundation for distributed resources working coordinately.
2) The module of resource monitor: it is responsible for managing and monitoring the resources in the grid.
3) The module of task scheduling: it is responsible for the scheduling among different states of the tasks as well as the task migrations and decomposition of tasks.
4) The module of data management: it is responsible for data stored, data duplication management and reliable data transmission, providing the quality assurance for resource shared.
5) The module of public management: it provides the common services for whole platform, such as fault-tolerant ability, expansibility, user's management and accounting management.
6) The module of safety management: it guarantees the integrality of information adopting different safety measurements in different levels in the grid.

The application layer provides web interfaces and various grid applications for the users, simultaneously includes the tools of grid applications, like the performance evaluation tool, monitoring tool and so on.

In the whole architecture of the CDG, the module of information service is at the extremely important position. Most of modules in CDG such as resources monitoring, task scheduling and accounting management etc. are based on it. Due to this point, in this paper, we analyze this module with emphasis and realize it.

3. THE DESIGN OF THE MODULE OF INFORMATION SERVICE

The purpose of the module of information service is to hide the complexity of the grid platform and provide a uniform and friendly application interface to user. Users can use it to issue service and information, browse services and discover services satisfied with one's own necessary etc.

* Fund project: Province Development and Planning Commission fund project (SX2004-29)

3.1 Service Registration

The realization of service registration is based on the Index Service in MDS4. The Index Service is the central component of MDS4, which collects data from various sources and provides a query/subscription interface to the data.

The Index Service is a registry service similar to UDDI, but much more flexible. It has following characteristics:

(1) Indexes collect information and publish that information as resource properties. Clients use the standard WSRF resource property query and subscription/notification interfaces to retrieve information from an Index.
(2) Indexes can be registered to each other in a hierarchical fashion in order to aggregate data at several levels.
(3) Indexes are "self-cleaning"; each Index entry has a lifetime and will be removed from the Index if it is not refreshed before it expires.

In order for information to appear in the index, the source of that information must be registered to the Index Service. Information sources are registered using tools like mds-servicegroup-add. Mds-servicegroup-add creates a set of registrations in a WS-ServiceGroup and periodically renews those registrations. Registrations are defined in an XML configuration file, an example named registration.xml used in this paper is shown as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>
<ServiceGroupRegistrations    ...>
   <ServiceGroupRegistrationParameters    ...>
      <ServiceGroupEPR>
         https://222.192.73.63:8080/wsrf/services/DefaultIndex
Service
      </ServiceGroupEPR>
     <RegistrantEPR>
         https://222.192.73.63:8080/ogsa/services/tutorial/core/
notifcationspush/MathFactoryService
     </RegistrantEPR>
     <RefreshIntervalSecs>30</RefreshIntervalSecs>
     <Content ...>
      <agg:AggregatorConfig
xsi:type="agg:AggregatorConfig">
        <agg:GetMultipleResourcePropertiesPollType    ... >

          <agg:PollIntervalMillis>20000</agg:PollIntervalMill
          is>
          <agg:ResourcePropertyName>RP_VALUE</agg:Res
          ourcePropertyName>
          <agg:ResourcePropertyName>RP_LASTOP</agg:Re
          sourcePropertyNam>
        </agg:GetMultipleResourcePropertiesPollType>
      </agg:AggregatorConfig>
     </Content>
   </ServiceGroupRegistrationParameters>
   …
</ServiceGroupRegistrations>
```

Each ServiceGroupRegistrationParameters block specifies the parameters used to register a resource to a service group.

The ServiceGroupEPR parameter specifies the EPR of the service group to register to,the RegistrantEPR parameter specifies the EPR of the resource registered, the RefreshIntervalSecs parameter specifies the refresh interval of the registration and the Content parameter specifies the aggregator-source-specific registration parameters. The Index Service is built on the WS MDS Aggregator Framework and it can use any aggregator source to collect information. In common use, the Index Service uses the QueryAggregatorSource to gather resource property values from the registered resource. The GetMultipleResourcePropertiesPollType element contained within the AggregatorConfig element specifies that the DefaultIndexService should poll "RP_VALUE" resource property and "RP_LASTOP" resource property every (n) milliseconds, based on the value of the PollIntervalMillis parameter.

3.2 Service Browsing

The service named ContainerRegistryService is the default service center of GT4.All services once are deployed, they will become its member services. The module of service browsing displays the name and the EPR of all local services based on the service named ContainerRegistryService. This work is mainly completed by the class named GetServices.

The class named GetServices mainly finish the following work:

(1) use the class named WSResourcePropertiesServiceLocator to get the EPR of the service named ContainerRegistryService.
(2) call the function of getResourceProperty to query the value of entry attribute of service named ContainerRegistryService.
(3) use hashtable to save the names and the EPRs of all the member services.

The main codes of the class named GetServices are shown as follows:

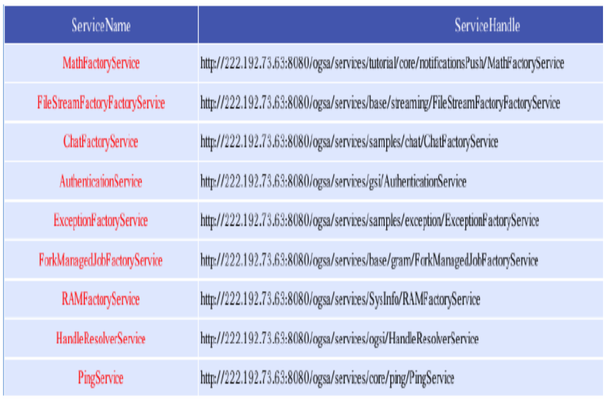
```
……
Hashtable servicelist = new Hashtable();
WSResourcePropertiesServiceLocator locator =
      new WSResourcePropertiesServiceLocator();
String   registryService=container.getURLString()
+ "ContainerRegistryService";
GetResourceProperty    port =
      locator.getGetResourcePropertyPort(new
URL(registryService));
GetResourcePropertyResponse                    response
=port.getResourceProperty(WSRFConstants.ENTRY);
Object[] entries =
           ObjectDeserializer.toObject(response,
EntryType.class);
     for (int i=0;i<entries.length;i++) {
       EntryType entry = (EntryType)entries[i];

       Servicelist.put(entry.getMemberServiceEPR().getAddres
       s(),
             entry.getServiceGroupEntryEPR().getServiceName(
));
                    }
     return Servicelist;
```

The running result of the module of service browsing is shown in Fig. 2.

3.3 Service Discovery

There are lots of services in grid environment, providing different functions. These services have different levels of QoS. It is obviously unrealistic that let user oneself look for service satisfied with the demand among boundless service ocean. So it is necessary to design a kind of mechanism to provide a service discovery function which can be used by users to find fastly the best service satisfied with users' own performance requirement.



ServiceName	ServiceHandle
MathFactoryService	http://222.192.73.63:8080/ogsa/services/tutorial/core/notificationsPush/MathFactoryService
FileStreamFactoryFactoryService	http://222.192.73.63:8080/ogsa/services/base/streaming/FileStreamFactoryFactoryService
ChatFactoryService	http://222.192.73.63:8080/ogsa/services/samples/chat/ChatFactoryService
AuthenticationService	http://222.192.73.63:8080/ogsa/services/gsi/AuthenticationService
ExceptionFactoryService	http://222.192.73.63:8080/ogsa/services/samples/exception/ExceptionFactoryService
ForkManagedJobFactoryService	http://222.192.73.63:8080/ogsa/services/base/gram/ForkManagedJobFactoryService
RAMFactoryService	http://222.192.73.63:8080/ogsa/services/SysInfo/RAMFactoryService
HandleResolverService	http://222.192.73.63:8080/ogsa/services/ogsi/HandleResolverService
PingService	http://222.192.73.63:8080/ogsa/services/core/ping/PingService

Fig.2. the running result of the module of service browsing

This paper proposed a kind of service discovery algorithm based on the QoS restrains. Reference [7] has carried on relevant research to QoS suitable for the grid service, proposing QoS properties including the response time, the reliability, the usability, the accuracy and the security, etc.. In order to choose and filter services conveniently, we sort out relevant QoS properties and design 4 indexes: Performance index, Load index, Available index and Average index.

(1) Performance index
This index reflects the disposition situation of machines. We mainly selected several representative parameters to calculate the performance index: CPU quantity, CPU frequency and the size of memory. Suppose there is a standard machine that can get “full marks”, its CPU quantity is 1, CPU frequency is 2GHZ and the size of memory is 256MB. In order to get the performance index of some machine, we shall first compare its parameters with the parameters of the standard machine to obtain the relative value of each parameter. Then the sum of all relative values, each multiplied by a particular factor is the performance index. The computing formula is as follows:

$$perindex = \left(\frac{cpucount}{def\_cpucount} * k_1 + \frac{cpuspeed}{def\_cpuspeed} * k_2 + \frac{memtotal}{def\_memtotal} * k_3\right) * 100$$

$$\sum_{i=1}^{3} k_i = 1$$

Among them, k1-k4 is used to represent proportion accounted for every parameter in the performance index

(2) Load index
This index reflects the load situation, mainly testing whether the machine is idle. The bigger the value of the load index is, the idler the machine is. We mainly selected several representative parameters to calculate the load index: the idle rate of CPU in 1 minute, the rate of available memory, the rate of available system size. Similar to the computing method of the performance index, the computing formula for load index is as follows:

$$loadindex = \left(\frac{cpufree}{\min} * t_1 + \frac{memfree}{memtotal} * t_2 + \frac{fsfree}{fstotal} * t_3\right) * 100$$

$$\left(\sum_{i=1}^{3} t_i = 1\right)$$

Among them, t_1~t_3 is used to represent proportion accounted for every parameter in the load index.

(3) Available index
This index reflects the accomplishment rate of task in a period of time, the bigger the value of the available index is, the more steady the machine is. This index is mainly based on the historical data, the monitoring module of the CDG system will monitor the implementation of each task, the information of the task will be recorded whether it is carried out successfully or not, some time later these information can be used to calculate the available index. The formula is as follows:

$$avaindex = \frac{\sum jobfinished}{\sum jobassigned} * 100$$

(4)Average index
This index reflects the comprehensive performance of the machines. While choosing a node to execute the task, we can judge which machine is the best according to the machines’ performance index, load index and available index synthetically. The bigger the value of the average index is, the more excellent the machine is. The formula is as follows:

$$avgindex = perindex * r_1 + loadindex * r_2 + avaindex * r_3 \qquad \left(\sum_{i=1}^{3} r_i = 1\right)$$

Among them, r_1~r_3 is used to represent proportion accounted for these three indexes in the average index.

The work of the service discovery is mainly completed by the class named CDGNode. It provides 4 methods to get 4 indexes described as above. The 4 methods respectively are getperindex(), getloadindex(), getavaindex(), getavgindex().

In the process of service discovery, first according to the functional requirements of users, the module of service discovery calls the module of service browsing to obtain all the services satisfied with user’s demand, then the module of service discovery call the algorithm of service discovery to calculate the average index of each machine node where services exit, at last return the best one to user.

4. SYSTEM TESTING

We construct a simple virtual organization imitating the grid environment by three PCs in the laboratory. The configuration of the experiment environment is as shown in Table 1. Among them, the machine of FDCP1 is the root-level member of the VO hierarchy.

We focused on the module of service discovery in the experiment. First we write service named MathFactoryService, the service provides simple mathematical computation functions (adds, reduces, multiplies and divides).Then we deploy the service in the container on the machine node of FDCP1and register it to the index service of local machine node, at last we create the instances of the service on three machine nodes in the VO. In order to imitate the use of pcs in the normal condition, we run some commonly used application program on each machine irregularly, such as text editing, Browsing webpage, etc., thus the idle rate of the memory and cpu is changed dynamically at random.

When testing , supposing the available indexes of the three machine nodes are all 100%, the value of r1、r2 and r3 respectively is 30%，40%,30% and the functional requirement of users is “Mathematical calculations”. As only "MathFactoryService" service provides the functions in the simulation environment, therefore, there are three nodes met the requirement in the whole VO. We chosen 3 group typical experimental data after lots of experiments shown in Table 2. From the experimental result we may draw the conclusion that the module of the service discovery can return the machine node having the best performance to user.

Table 1. the environmental configuration

Host Name	IP Address	Operating System	Configuration
FDCP1	222.192.73.63	RedHat Linux 9.0	P4 1.7G/256MB/40G
FDCP2	222.192.73.67	RedHat Linux 9.0	P4 1.7G/256MB/40G
FDCP3	222.192.73.68	RedHat Linux 9.0	P4 1.7G/256MB/40G

Table 2. the experimental data of service discovery

Group num	Node name	Performanc index	Load index	Available index	Average index	Node chosen	result
1	FDCP1	93.0	81.0	100.0	90.3	FDCP2	True
	FDCP2	93.0	94.0	100.0	95.5		
	FDCP3	93.0	88.0	100.0	93.1		
2	FDCP1	93.0	78.0	100.0	89.1	FDCP1	True
	FDCP2	93.0	54.0	100.0	79.5		
	FDCP3	93.0	49.0	100.0	77.5		
3	FDCP1	93.0	59.0	100.0	81.5	FDCP1	True
	FDCP2	93.0	51.0	100.0	78.3		
	FDCP3	93.0	50.0	100.0	77.9		

5. CONCLUSIONS

The essential characteristics of the grid is to combine existing resources as most as possible, realize resources shared and eliminate the isolated island of information. This paper design the whole architecture of the CDG platform, mainly analyze the technology involved in the module of the information service of CDG System and give the corresponding solution. At present the further developing work is still going on.

REFERENCES

[1] Li Xu, Ma Huadong, "A WSRF-based grid resource information service architecture", Huazhong Univ. of Sci. & Tech., 2006 ,9.

[2] Babu Sundaram, "Registered grid service in GT4", http://www.ibm.com/developerworks/cn/grid/gr-mdsgt4, 2005.6

[3] K. Czakowski, S. Fitzgerald, I. Foster .etc, "Grid Information Services for Distributed Resource Sharing", *Proceedings of the Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-l0)*, IEEE Press,August 2001.

[4] Official website of Globus, http://www.globus.org/toolkit/docs/4.0/.

[5] P. Dinda and B. Plate, "A unified relational approach to grid information services", *Grid Forum Working Draft GWDGIS-012-1,* February 2001. http://www.gridforum.org.

[6] Chamberlin, D, "Xquery 1.0: An XML Query Language", W3C Working Draft 07, 2001.

[7] Liu YT, Ngu AHH, Zeng LZ, "QoS computation and policing in dynamic Web service selection", In: *Feldman SI*, uretsky M, NajorkM, Wills CE, eds. Proc. of the 13th Int'l Conf. on World Wide Web (WWW 2004). New York: ACM Press, 2004. 66−73.

Research on Grid Portal in Manufacturing Grid Environment*

Yong Yin, Zude Zhou, Yihong Long
School of Information Engineering, Wuhan University of Technology, Wuhan City, 430070
Email: yiyng_hust@126.com

ABSTRACT

In traditional manufacturing system interface, users have to do complicate configurations and memorize various kinds of instructions. The intention to study Manufacturing Grid portal is to provide manufacturing resource sharers a friendly interface. In the paper, the architecture of Manufacturing Grid portal is brought forward. Key components of the portal, such as the user management module, Portlet middleware, Web Service server and LDAP server is researched on and realized. Via this portal, users can access diverse manufacturing resources through Web server, submit and then monitor manufacturing tasks efficiently and safely.

Keywords: Manufacturing Grid, Grid Portal, Portlet Middleware, Web Service Server

1. INTRODUCTION

Manufacturing Grid is the developing platform and supporting environment for modern integrated manufacturing system, which combines related design and manufacturing resources, processes and knowledge information during the whole life of the produces [1]. Manufacturing Grid is built on the ground of traditional TCP/IP infrastructures with various kinds of resources as its nodes [2]. Its ultimate objective is to realize resource sharing, collaborative design and manufacturing, to achieve lower costs, high efficiency of resource utilization and fast reaction to markets.

As Manufacturing Grid system is much complicated, some tasks need complex parameter configurations, and users have to remember all kinds of commands according to their programming interface. Moreover, when users share manufacturing resources over the Manufacturing Grid, traditional interface based on the Web needs tools of the third providers, and it also lacks the function of security such as single sign-on and authentication.

The purpose to research on Manufacturing Grid portal is to provide a friendly interface for manufacturing resource sharing. Through the portal, users can share the grid resource conveniently. The portal also provides functions of resource management, task submission, resource dispatching and monitoring, while hiding complex operations. Furthermore, it provides an integrated, point-to-point safety structure for the exploitation and management of manufacturing resources.

2. THE STRUCTURE OF THE MANUFACTURING GRID PORTAL

In Manufacturing Grid, the portal can integrate various kinds of grid service, and it is the only entry of these services. Users can assess the portal through Web browser, so as to approach each kind of grid service directly, or intercommunicate with the grid services through the grid middleware indirectly. The structure of the Manufacturing Grid portal is shown as Fig.1.

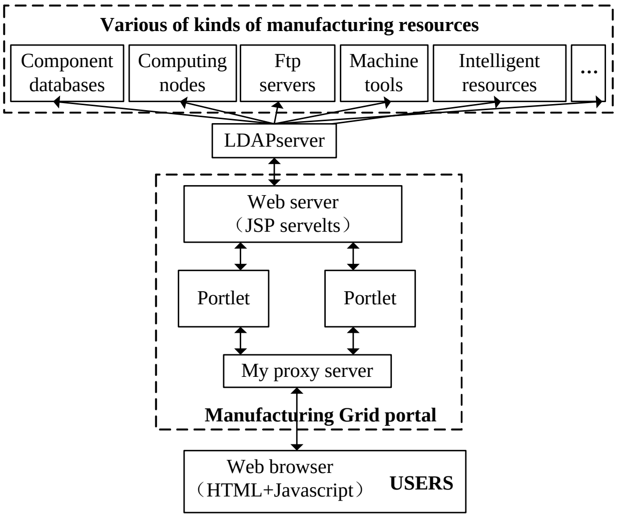


Fig.1. The structure of the Manufacturing Grid portal

As shown Fig.1, after being authenticated and then authorized by My proxy server, users can inquire or search each kind of manufacturing resources, or submit, operate or manage each kind of manufacturing tasks through Web server based on JSP by virtue of Portlet grid middleware. The results or error information of each manufacturing task can also be feedback to users through Web browser. In the structure of the portal, the manufacturing resource information management utilizes LDAP (Lightweight Directory Access Protocol). The Portlet is a kind of grid application, or called grid middleware. It is a java class independent of programming interface, so it can be loaded into Web server and then run on it. The difference between the Portlet and the Servlet is that, Servlet must intercommunicate with the Web browser, while Portlet only intercommunicates with user browser indirectly through the grid portal. From the viewpoint of the users, one grid portal page contains one or more Portlet windows, while the user interface is made up of several Portlets. Each Portlet provides one special access or a series of accesses of the grid services.

3. THE DESIGN OF THE MANUFACTURING GRID PORTAL

3.1 User Management

In the environment of Manufacturing Grid, if the user wants to access the manufacturing resources in the grid, he must be authenticated and then get his corresponding authorization. The research of user security management in Manufacturing Grid portal concentrates on this point. As shown in Fig.1, the user management is accomplished by My proxy server. The existed grid development environment, such as Globus Toolkits [3],

* Support by the Project Fund of the Hubei Province Key Laboratory of Mechanical Transmission and Manufacturing Engineering Wuhan University of Science and Technology(Grant No. 2007A18) and the Funding from the National Science Fund Committee (NSFC) of China (Grant No. 50335020).

doesn’t take user security management into account. In order to deal with user security issues, we adopt My proxy. My proxy can save temporary information of users on the server, and protect the system through certain defined rules or username /password. In this way, it is not necessary for users to show their digital certifications or to save delegates on each machine every time they access the grid system. In the meanwhile, when users want to access the grid system another time, they can access all the manufacturing resources just by obtaining the temporary delegates back from the My proxy server. This is called single sign-on.

After the user uses the username/password to log in the Grid portal through Web browser, Grid portal act as the user’s delegate and intercommunicate with the grid resources, so the grid portal server must get the user’s delegate certificate. In general, the user submits his delegate certificate to the My proxy server, while the portal server gets the user’s delegate certificate from the My proxy server according to the username/password provided by the user, and owns the delegate certificate during the whole process of the operation. The procedure of the user’s authorization and authentication on My proxy server is shown as Fig.2.

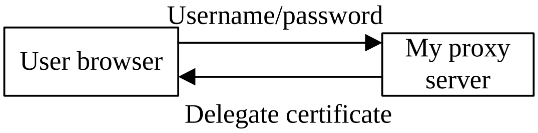


Fig.2. The process of user authentication and authorization

3.2 Portlet Middleware

In the design of Manufacturing Grid portal, the grid service requests submitted by users through Web browser is integrated into Portlet middlewares. According to different time, different roles or different environment of each user, different services can be selected and integrated by adding or removal the corresponding Portlet middlewares. Portlet middleware uses the standard interface to communicate with the grid service described with Web Services Description Language (WSDL). The inter-communication between the OGSA client and each grid service is achieved with standard protocols such as Simple Object Access Protocol (SOAP)、Hyper Text Transfer Protocol (HTTP) and Java Remote Procedure Call (RPC)[4]

As demonstrated in Fig.1, in the system of Manufacturing Grid portal, various kinds of Portlets are designed, such as the user’s log in Portlet、resource information Portlet、resource browsing Portlet、online user Portlet、user management Portlet、task dispatching Portlet and so on. Users access the grid portal through the Web browser, and to select different service Portlet according to different requests. The process is demonstrated in Fig.3.

3.3 Web Service Server

As shown in Fig.1., after being authenticated and then authorized by My proxy server, users submit their service request over the Manufacturing Grid portal to the Portlet middleware through Web browsers. In here, each submitted service is implemented by the Web server. The Web server is based on the JSP Servelts technology, and it is a universal platform unrelated with CPU architecture or programming language. Different applications depend on the Web server to interact and integrate with each other. In this way, the manufacturing resource information can be distributed on the Internet freely. The Web server adopts Web Service technology, which deals with issues confronted with Manufacturing Grid researchers, such as application integration、resources sharing、system intercourse operation and standardization[6]. The architecture of Web Service server itself is divided into four layers, namely network protocol layer, intercourse layer, service description layer and service discovery layer, as shown in Fig.4.

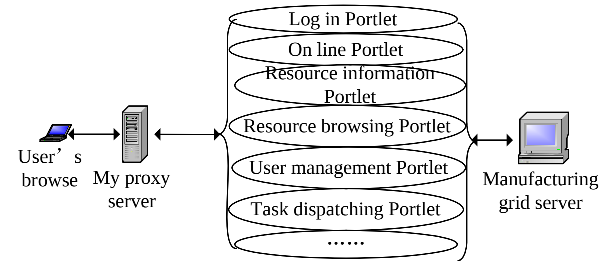


Fig.3. Portlet middleware

Web Service server

service discovery layer	UDDI
service description layer	WSDL
intercourse layer	SOAP
network protocol layer	HTTP、SMTP etc

Fig.4. Web Service server architecture

The layer functions and relations are described as following:

(1) Network protocol layer. The layer is based on the traditional TCP/IP protocol and built on the Internet infrastructure. The actual network protocol depends on specific program requirement.

(2) Intercourse layer. The function of the intercourse layer is to provide the standard protocol for the communication between the service consumer and the service provider. In the Web Service server, SOAP protocol is employed. SOAP is defined on the basis of XML, combining with the features of opening and extension of XML completely. SOAP is based on the TCP/IP’s application protocol such as the HTTP、SMTP、POP3 etc, which is compatible with the current network protocol to the most degree. SOAP doesn’t define any applicable semantics such as programming model or special semantics’ realization. It just defines a simple mechanism to express the application semantics.

(3) Service description layer. Service provider sends all the regulations for Web Service to the service applicant through the service description. In order to decrease the difficulty of the realization of service between the service provider and service applicant, the service description is critical. In this way, it is not necessary for the service provider and service applicant to know mutually the details of the bottom platform, programming language or distributed object model (if it exits). Service description is combined with the bottom SOAP basic structure to encapsulate the realization details of the applications of the service applicant and the Web service of the service provider. WSDL defines a series of syntax based on XML, and it can make the Web Service serve as the group of service access points that can alternate information.

(4) Service discovery layer. In the Web Service servers, the

utility of SOAP protocol and WSDL regulation accomplishes the integration of the information service, leaving alone the distinction in software or hardware platform and the programming language. In here, UDDI (Universal Description, Discovery and Integration) regulation is employed to accomplish the issuing of information service. UDDI is an efficient protocol which can describe, browse, find and integrate the service information based on Web Service, and it is a good mechanism that provides a information description for their interaction [7].

3.4 Ldap Server

In the Manufacturing Grid environment, various kinds of manufacturing resources can be added in or released from grid servers dynamically. The manufacturing information management is accomplished by the LDAP (Light Directory Access Protocol) server. In the system, the OpenLDAP is adopted as the protocol software, and its architecture is like a tree, as shown in Fig.5.

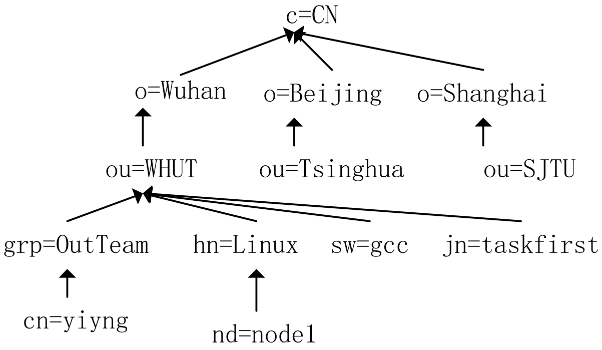
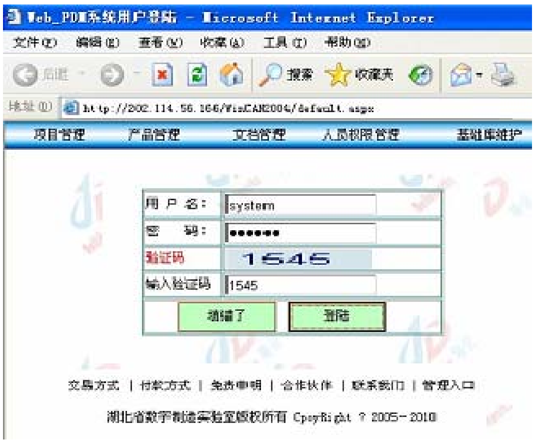


Fig.5. LDAP directory

Information is stored in the LDAP directory on the Manufacturing Grid server. The LDAP directory is divided into the OU (organization unit). OU contains the information items, so the system not only can be extended conveniently, but also can search or inquire information more quickly than traditional rational database. LDAP server can store all kinds of information, including the manufacturing resource information, such as Human Resource (HR), Manufacturing Equipment Resource (MER), Technology Resource (TR), Application System Resource (AR), Service Resource (SR), User Information Resource (UIR), Computer Resource (CR) and so on. Other Resource (OR) contains email address information, DNS information, NIS mapping, user list and computer name etc. When users access the information in LDAP server, the client Web browser sends information to LDAP server through the LDAP protocol. After being authenticated by the LDAP server, user then can perform corresponding operation on the database.

4. THE DESIGN OF THE MANUFACTURING GRID PORTAL

Supported by Key program of the National Nature Science funding of China and the funding of National 211 Project of China, the Manufacturing Grid system has been established with Dawning 4000 grid server. We are bending ourselves to apply for the access to China national grid system. On the base of this platform, Manufacturing Grid portal is designed, as shown in Fig.6 and Fig.7. Fig.6 is the user log in interface, and Fig.7 is the user task submission interface.

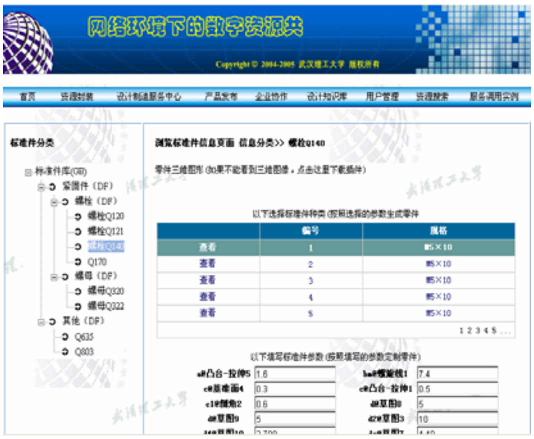

Web_PDM系统用户登陆 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D)
项目管理 产品管理 文档管理 人员权限管理 基础库维护
登陆
交易方式 | 付款方式 | 免责申明 | 合作伙伴 | 联系我们 | 管理入口

http://202.114.56.166/WisCAM2004/default.aspx

用 户 名：	system
密　　码：	●●●●●●
验证码	1545
输入验证码	1545

湖北省数字制造实验室版权所有 CpoyRight ? 2005－2010

Fig.6. User log in interface



浏览标准件信息页面 信息分类>> 螺栓Q140

零件三维图形 (如果不能看到三维图像，点击这里下载插件)

以下选择标准件种类 (按照选择的参数生成零件)

	编号	规格
查看	1	M5×10
查看	2	M5×10
查看	3	M5×10
查看	4	M5×10
查看	5	M5×10

1 2 3 4 5 ...

以下填写标准件参数 (按照填写的参数定制零件)

a#凸台-拉伸5	1.6	b#螺旋线1	7.4
c#基准面4	0.3	e#凸台-拉伸1	0.5
c1#倒角2	0.6	d#草图0	5
d#草图9	5	d2#草图3	10

Fig.7. User task submission interface

In Fig.7, the left frame of the interface displays the information of the manufacturing tasks, namely the information of the grid node. The information is organized with tree structure, with one node represents one task or one grid node. One task node contains several sub-nodes. When click one node, corresponding information of the node will display in the right frame of the interface.

5. CONCLUSIONS

The article elaborates the system structure of Manufacturing Grid portal, and provides the design and realization of the Portlet middleware, the Web Service server and the LDAP information management server. Through this portal, users can access the grid resources, submit their manufacturing tasks and monitor the result of the task. The next step of our research is to design the users' customized Web interface, to integrate more functions into the portal, such as the graphic display of the topological structure of the grid, the statistical information the manufacturing tasks, and to enhance functions of task scheduling and its real-time monitoring etc.

REFERENCES

[1] Foster I, Kesselman C, Tueche S, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *International Journal of High Performance Computing*

Applications, 2001, 15(3): 200-222.

[2] Cao J W, Jarvis S A, Sain S, et al, “ARMS: An Agent-based Resource Management System for Grid Computing,” *Scientific Programming*, 2002: 135-148

[3] Zheng J, Jiang X. F, “The Summarization of the Design Tools for Grid Portal,” *Computerization and Modernization*, 2006,129(5): 101-104

[4] Foster. I, “Globus: A Metacomputing Infrastructure Toolkit,” *International Journal of Supercomputing of Supercomputer Application*, 1998, 11(2): 115-129

[5] Chen B, Chi X. B, Wu. H, “The Design and Realization of the Computer Grid Portal,” *Microelectronics and Computer*, 2004,21(9): 15-18

[6] Novotny. J, “An online credential repository for the grid: MyProxy,” *High Performance Distributed Computing*, USA: IEEE Press, 2001. 104 -111.

[7] Anand. N, “The legion grid portal. Concurrency and Computation: Practice and Experience, Grid Computing environments,” *Special Issue* 13-14, 2002.14: 1365-1394.

Yong Yin is an Associate Professor in School of Information Engineering, Wuhan University of Technology. He graduated from Wuhan Institute of Science and Technology in 1997; from Huazhong University of Science and Technology in 2000 with specialty of mechanical engineering and automation. His research interests are in intelligent manufacturing, embedded control and grid computing.

Comparing Open Source MPI Implementations Performances in a New Grid Computing Environment

Zhixiang Zhao[1], Jianguo Wang[1], Haiwu He[2]
[1]Computer Science and Engineering College Xi'an Technological University
[2]Computer and Information Engineering College, HOHAI University, Nanjing, JiangSu, China
INRIA Grand-Large/LRI, Paris-Sud University Orsay, France, Corresponding author
E-mail : xinxiangshine2006@126.com,wjg_xit@126.com, hehaiwu@hhu.edu.cn, haiwu.he@inria.fr

ABSTRACT

We compare the performances of the 3 most popular MPI implementations-LAM/MPI, MPICH1,MPICH2 utilizing a totally new GRID infrastructure –Grid'5000 in INRIA. This new GRID infrastructure Grid'5000, a 5000 CPUs nation-wide infrastructure for research in Grid computing will be explained. Some basic communications tests including both point-to-point and collective MPI communications which represent communications patterns in common use are well presented. In the end of this paper, with the analysis of results of tests, we give some practical suggestions to choose appropriate MPI implementation for scientific computing MPI users.

Keywords: MPI, LAM/MPI, MPICH, Grid'5000, Performance

1. INTRODUCTION

Nowadays, the Message Passing Interface. (MPI) [1] is widely used of scientific engineering computing in High Performance Computing (HPC) machines. And some High Performance Clusters play a more and more important role in HPC. It is very common to install the Open Source, free (GPL) Message Passing Interface (MPI) implementation [2]to facilitate parallel computing on such systems. A tradeoff is often made between commercial MPI implementations and the free alternatives. Additional binary packages or specific installation instructions are provided in these commercial implementations. They are expected as well to be customized for particular computer architectures and even supply user support. So there is an interest to find out the overheads of a free implementations and whether this lack of techinical support is really fatal. With more and more clusters home-made are commonly used for our scientific computing, these free MPI implementations are often installed. So the investigation of Open Source MPI implementations becomes very meaningful.

Open Source free (GPL) implementations have been developed for many years. Following the new MPI-2 standard, recently the "next-generation" of MPI implementations has come out, With MPI-2, LAM/MPI [4]continues to build new features into the previous implementation. Now LAM/MPI has stopped development to commence the new collaborative project Open MPI, which is completely redesigned from MPI-2 standard. However MPICH2 which is a totally new implementation shared no MPICH1 codes.

Grid'5000[5] is not a Grid. It is a highly configurable, controllable and monitorable Grid computing environment that can be reconfigured to work as real Grids.

The main aim of this paper is to compare and contrast three Open Source MPI implementations: LAM/MPI, MPICH-1,MPI-CH2 on Grid'5000, identifying their strengths and weaknesses on common HPC clusters to give some recommendations to choose appropriate MPI implementation for scientific computing MPI users.

2. GRID'5000

Grid'5000 (www.grid5000.org) is a nation wide Grid computing environment. Grid'5000 is not an emulator. It aims at providing a strong reconfiguration, control and monitoring infrastructure, transforming the full system into a scientific instrument as a Grid computing component.

In fact, GRID'5000's strong reconfiguration ability enables a lot of configurations. The system software can be totally reconfigured on all processors, at a nation wide scale. Like physics instruments, Grid'5000 can be used and shared by many researchers. Its control infrastructure makes researchers run both interactive tasks and batch experiments. Experiments in batch mode follow a complex sequence of operation including the following steps: reservation, reconfiguration, run preparation, run and post run resource liberation. Since researchers can fully reconfigure the software of all Grid'5000 processors, a fully restricted environment is imposed for the security reasons. Basically, no communication packets can exit or enter Grid'5000 except by highly secured gates. This security architecture has be designed to avoid the possibility of using Grid'5000 as a platform for massive DOS attacks. During the experiments several probes capture and record nodes and networks measurements.

The implementation of Grid'5000 is realized by a nation wide cluster of clusters. 9 sites in France are equipped with clusters with powerful CPUs whose number varies from 100 to 1000. All the sites are interconnected by RENATER through VLANS implemented by MPLS at level 2. These clusters are isolated from the Internet. From the Internet, all accesses to clusters are firstly forwarded by firewall host of laboratory. Strong authentication and authorization verification procedures check the access rights of users. Once a user is logged in Grid'5000, he can access to all the Grid'5000 processors without restriction. A reservation tool called OAR permits to reserve nodes on several sites. After the reservation, the user can deploy all specific softwares and reboot the reserved nodes (Kadeploy is used as deployment and multi-site reboot system). After the reboot, the user can start the experiment.

3. IMPLEMENTATIONS

The clusters we used are mainly in the site of Orsay in Grid'5000. The configuration is described in table1, and table 2.. We have to customize our computing environment: Firstly, we install all 3 MPI implementations and IMB in one computing nodes, make a system image. Secondly we reserve the nodes which we want to do our computing. Thirdly we clone the same system to all the computing nodes by using the deploying tool "Kadeploy" from the system image we just created. Finally, we created our own computing cluster including all reserved nodes in Grid'5000. When we do our computing in Grid'5000, the system is transparent after establishment of our own computing environment. That is to say, when we use the "cluster" just created, we can not feel the existence of Grid'5000. And there is no additional communications and others costs inside our "cluster" due to Grid'5000. It acts just like a "cluster" alone. Once the time of reservation expires, all computing nodes will be rebooted, this virtual "cluster" will not exist.

Table 1. Clusters of Orsay in Grid'5000

Number of CPUs	CPU	Memory
216	Dual AMD Opteron 246,1.95 GHz	1.96G
121	Dual AMD Opteron 250,1.95 GHz	1.96G

Table 2. Network of Orsay Clusters

Bandwidth	latency
941Mb/s	0.09ms

A freely available benchmark, timings and throughput of a particular implementation can be found easily. This benchmark should be popular, long-standing and peer-approval in order to be confident as to accuracy and possibly compared in the future with different machines results. Benchmarks can automate and organize tests for the functions defined by the MPI-1 standard adapted to the Open Source MPI implementations to test. This automation can be run on varying CPU numbers and interconnects, with different sizes and types of data.

The Benchmarks fulfill our demand are Intel MPI Benchmarks which are used for comparing our MPI implementations performances. The Intel MPI Benchmarks (IMB), formerly known as Pallas MPI Benchmarks, is a "concise set of benchmarks targeted at measuring the most important MPI functions" [3] reporting timings and throughput data for latency and bandwidth comparisons. The benchmarks focus on recording timings across all involved processes, with the aim to producing an average recorded time. This is implemented using the following system:

```
for ( i=0; i<N_BARR; i++ ) MPI_Barrier(MY_COMM)
  time = MPI_Wtime()
  for ( i=0; i<n_sample; i++ )
    execute MPI pattern
time = (MPI_Wtime()-time)/n_sample
```

Processes are synchronized, and the MPI benchmark begins its repetitions of the testing formula. After all the repetitions are completed, the time is divided by the number of repetitions to form a recorded local time to that process. The number of repetitions is initially 1000, but for the largest message sizes is progressively smaller. IMB then reports the maximum and minimum recorded time and calculates the mean time from the result on each process, in order to provide the approximate time of one operation across the system. For runs on large CPU counts, further benchmarks are taken with smaller CPU counts, while other CPUs wait using MPI Barrier. For example benchmarking 16 CPUs also produces benchmarks for 8, 4 and 2 CPUs, where such a benchmark is meaningful.

The Intel MPI Benchmarks User Guide [3] organizes the benchmarks into three subsets, we give explanations of only transfer benchmarks tested.

Single Transfer Benchmarks

This group benchmarks a single message, sent between two processes. There are two variants in this group

• Ping-Pong

Using MPI Send and MPI Recv, a message is sent from processor A to B, and then a similar message back from B to A. The time reported by the benchmark is half this timing.

• Ping-Ping

With PingPing, processes A and B simultaneously execute MPI Isend, a non-blocking send, followed by MPI Recv, a blocking receive, and MPI Wait, waiting on the earlier MPI Isend. The time reported is the entirety of this sequence.

Parallel Transfer Benchmarks

These benchmarks measure performance with all processors executing similar routines concurrently.

• SendRecv

With SendRecv, the processes form a periodic chain, with each individual process sending data to it's right, and receiving from it's left. This is achieved using the MPI SendRecv function.

Collective Benchmarks

These benchmarks are best suited for analyzing the implementation and algorithms used in MPI collective operation, i.e. those where information from all processes is operated upon. While the benchmark suite is very thorough, essentially with a benchmark for each of the MPI-1 standard's defined operations, this investigation focuses on a subset of commonly used functions. Furthermore the speed of each message transferring is not the only concern, the mathematical operations that manipulate the collected data have a small effect. The MPI library may have optimized operations on data types but the core numerical operations are the same, having used the same compilers. The data from collective operations is reported in raw latency timings.

• AllToAll

Benchmark for the MPI_AllToAll function. Every process inputs X*(#processes) bytes (X for each process) and receives X*(#processes) bytes (X from each process).

4. BENCHMARKS RESULTS

To make sure results comparable, the same version of each MPI implementation has been installed and used on Grid'5000 clusters. We use the latest stable release of each library. This makes a better tested implementation which could give a more accurate evaluation of that implementations ability. These versions are:

- MPICH1 1.2.7p1, Released in June, 2005
- MPICH2 1.0.3, Released in November, 2005
- LAM/MPI 7.1.2, Released in March, 2006

The executables were compiled from source in all instances. Compilers vary across platforms and implementations. Different compilers could not affect the ability to compare performance across machines, thanks to the relatively small amount of computation that the library actually performs. Because most of the time will be spent in Operating System (OS) system calls, where the compiler has little impact. These were combined with the latest release of the Intel MPI Benchmarks, version 3.0. This was compiled from source on all nodes, using the compilers and MPI implementations listed above.

For our tests, we just do the typical benchmarks of communications which are commonly used in scientific engineering computing: Ping-Pong, Ping-Ping, SendRecv, and AllToAll. The sizes of messages transferred are following: 8 bytes (small size), 10k bytes (middle size), 1M bytes (large size, in the case of "AllToAll" is 100M bytes). The time in tables and figures are measured in milliseconds. To obtain better results, we tried transfers as many as possible. The time of repetitions of MPI transfer benchmarks for small and middle size is 1000, for large size is 400. The values of time shown are all means. The amount of processors used for benchmarks ranges from 2 to 64, and is increases by twice every time.

Table 3. Ping-Pong Test

	MPICH1	MPICH2	LAM/MPI
8bytes	56.64	53.59	46.32
10Kb	193.36	185.31	169.14
1Mb	11767.58	9159.86	9105.07

Table 4. Ping-Ping Test

	MPICH1	MPICH2	LAM/MPI
8bytes	62.5	60.75	53.51
10Kbytes	195.31	181.19	162.82
1Mbytes	40429.69	14018.52	12680.77

In Table 3 and Table 4, we notice that for our configurations, LAM/MPI outperforms everything in point-to-point communications. MPICH2 is better than MPICH1. MPICH1 has the worst performance.

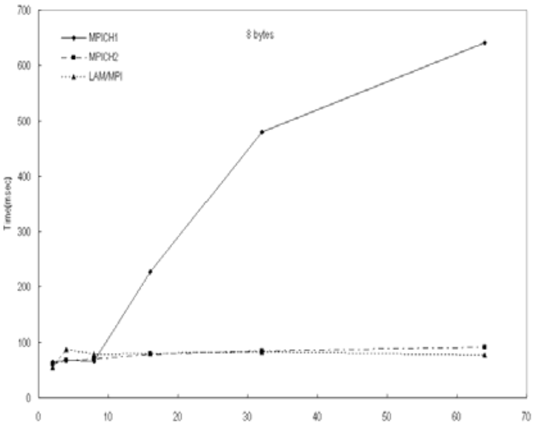


Fig. 1. MPI Performances of SendRecv 8 bytes

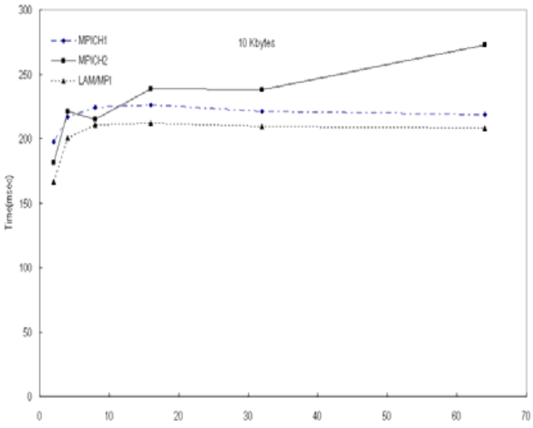


Fig. 2. MPI Performances of SendRecv 10k bytes

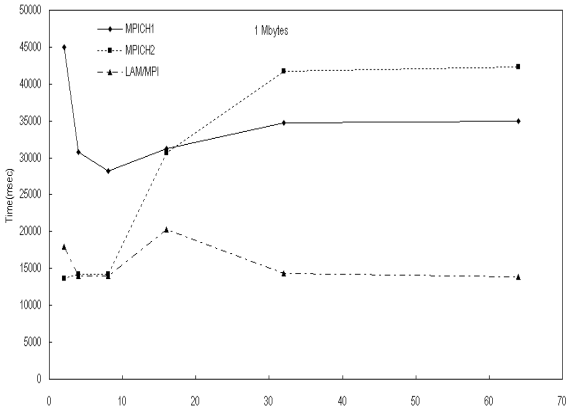


Fig. 3. MPI Performances of SendRecv 1M bytes

In Fig.2, Fig.3 and Fig.4, we can remark that totally for SendRecv this parallel transfer, LAM/MPI has the best performance, than MPICH2, the last one is still MPICH1.

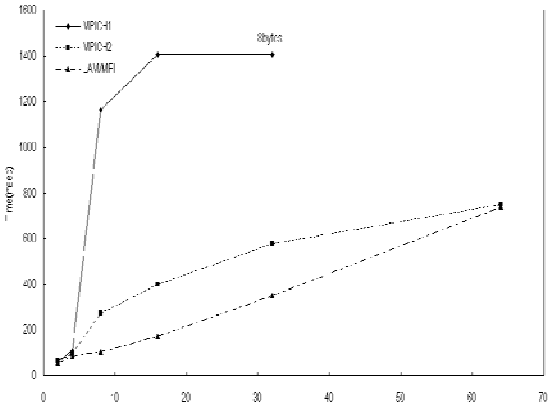


Fig. 4. MPI Performances of AllToAll 8 bytes

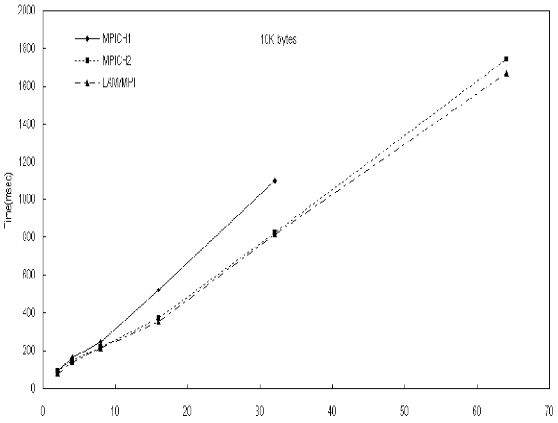


Fig. 5 . MPI Performances of AllToAll 10k bytes

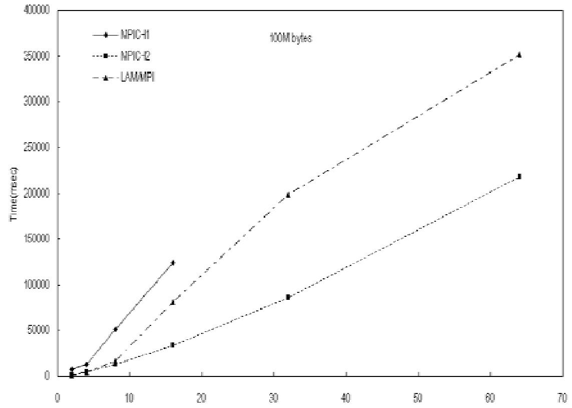


Fig.6. MPI Performances of AllToAll 100M bytes

In Fig. *4*, Fig. *5* and Fig.6, we show the most talkative communication of MPI: AllToAll communication. For MPICH1, in Fig. 4 and Fig. 5, benchmarks can not run on more than 32 processors, and for large size message of 100M, in Fig.6, when number of processors is more than 16, the benchmark will block. For LAM/MPI, when size of message is not large, the performance of LAM/MPI outperforms in Fig. 4 and Fig. 5. In Fig.6, when message transferred is large with a size of 100M. MPICH2 has the best performance.

5. CONCLUSIONS

We made many benchmarks of 3 most popular MPI implementations: MPICH1, MPICH2 and LAM/MPI on a new Grid computing environment Grid’5000. When the virtual clusters are established inside of Grid’5000, we have a very typical scientific computing cluster environment for benchmarks.

MPICH1 implements many collective operations with advanced algorithms, often choosing one based on the MPI function’s passed variables. Performance results shows that the core MPICH1 system does not handle multiple simultaneous large messages very well, with performance of collective operations degrading as message size increases.

MPICH2 uses the same algorithms as MPICH1, yet outperforms it. The major change between MPICH1 and MPICH2 is the complete reworking of the library’s architecture. This is apparent as with MPICH2 a daemon is now run on each host before execution of the MPI program, whereas before the hosts would be connected at runtime of the program.

With the consideration which kind of communication is the majority in our scientific computing program, we can choose a right Open Source MPI implementation. When point-to-point and parallel transfer communications are the most important, LAM/MPI seems a better choice. When collective transfers are used a lot in the program, MPICH1 is the first one to give up. When large messages are sent in this case, MPICH2 is more efficient. MPICH1 is not on the list of choices because of the upgraded MPICH2, and less performing than LAM/MPI.

REFERENCES

[1] MPI Standard Web Site. http://www-unix.mcs.anl.gov/mpi/index.html

[2] Snir M, Otto SW, Huss-Lederman S. Walker DW, *Dongarra J. MPI.*

[3] Intel Corporation Intel MPI Benchmarks: “Users Guide and Methodology Description. IMB v3.0”

[4] The LAM/MPI Team, “Open Systems Lab”. LAM/MPI User’s Guide Version 7.1.2

[5] C.Frank,C.Eddy et al “Grid'5000: A Large Scale Highly Reconfigurable Experimental Grid Testbed”, *IJHPCA* (International Journal on High Peerformance Computing and Applications), 20(4)481-494,2006

A Model of Resource Reservation in Grid Based on Computing Economics

Jing Li [1,2] Shuyu Chen[3]
[1]College of Computer Science, Chongqing University, Chongqing, China
[2]Department of Computer and Modern Education Technology, Chongqing Education College, Chongqing, China
3. College of Software Engineering, Chongqing University, Chongqing, China
Email:li_jing1@163.com

ABSTRACT

Grid resource management plays an important role while enabling the sharing and coordinating of resources in Grid computing environments. Resource reservation is an important part of the Grid resource management. An advance reservation is a scheduling object which reserves a group of resources for a particular timeframe for access only by a specified entity or group of entities. Combining resource reservation with computing economics, we propose a mathematic model which can determine resource booking quantity. In this model, taking economic interest and reputation index of service supplier into account, an optimization problem of two targets under hypothesis is simplified in reason. Through numeric calculation and analysis, we show there is an optimal point for the advance booking of resource.

Keywords: Grid, Resource Reservation,
Computing Economics, QoS

1. INTRODUCTION

In Grid, resource has such main features as decentralization, heterogeneity, and autonomy, and resource management is a challenging technology of Grid. One of resource management model is based on marked economics which focus on resource management and allocation under marked competition. To maximize the profit of both sides, it utilizes supply & demand philosophy to adjust resource provider and resource consumer. At present Distributed computational architecture based on marked economy is proposed in some literatures[1][2], namely, Grid Architecture for Computational Economy.

In order to guarantee the quality of service(QoS) in Grid, resource reservation is a very effective approach[3]. In general, two types of resource reservations in computer networks can be distinguished: immediate reservations which are made in a just-in-time manner and advance reservations which allow to reserve resource a long time before they are actually used. Advance reservations are especially useful for Grid computing but also for a variety of other applications. In advance reservation environments, additional new services can be served such as malleable reservations which can lead to an increased QOS of the network.

Combining resource reservation mechanism with computing economics, we propose a new model of resource reservation in Grid. In this scheme, taking economic interest and reputation index of service supplier into account, we simplify an optimization problem of two objectives under hypothesis in reason, to get an optimal point for the ordered resource number.

2. RELATED WORK

Four strategies for scheduling reservations are presented and compared in literature[4]. The results of the comparisons show that some strategies increase the resource fragmentation and are therefore unsuitable in the considered environment while others lead to a significantly better performance of the network. Besides discussing the performance issue, in this paper the software architecture of a management system for advance reservations is presented.

The Globus Architecture for Reservation and Allocation(GARA)[5] enables the construction of application-level co-reservation and coallocation libraries that applications can use to dynamically assemble collections of resources, guided by both application QoS requirements and the local administration policy of individual resources.

Existing resource reservation approaches were designed in terms of the stability of resource demand in general. Nevertheless, in Grid environments, resources and demands are both highly variable, which is the inherent characteristics of Grid and results in limitation of existing approaches.

3. PROBLEM ANALYSIS AND MODEL ASSUMPTION

Several roles are defined as follows:

- Grid Currency (GC). It is an universal equivalent in Grid environment, and all the Grid resource measurement can be converted proportional to the measurement of GC[6].
- Grid Resource Provider (GRP). Providing service by selling resources, it obtains GC.
- Grid Resource Consumer (GRC). Buying resources by paying GC, it obtains service.

If the Grid is thought of as a market, resource reservation business will be started because GRP wants to improve the quality of service to attract more GRC. GRP promises that GRC ordering resource do not prepay deposit, and if one could not come to make use of resource, he can use it preferentially in the next round. Making resource reservations would not need any additional cost.

When starting resource reservation business, for one service on offer, if GRP restricts the number of reserved resource to be his upper limit, then there will always be the case that some consumers having ordered service can not come and resources are not fully utilized, resulting in decrease of profit. In contrast, if not confining the number of reserved resource, when consumers ordering service come in time, exceeding the load burden capability, the complaint will arise, leading to reputation damage and economic loss such as guests decreasing, the compensation paid by provider, and squeezing out the consumers in next round. Hence we need to thinking about the economic interest together with Reputation Index to determine the optimal point of resource number.

The economic interest of GRP can be measured by revenue taking out cost and compensation, while reputation index by confinement of the number of GRC who has ordered service but can not get service because of the GRP loading down. Note that the critical factor, GRC booking service will come on time or not, is random. Accordingly the economic interest and reputation index should be weighted on average, which is an optimization issue of two objectives.

Before presenting the resource reservation model, we make following assumptions:

Assumption 1. A service is served as a work flow, consumers can not join in midway; only when this round of service is over, the next round will start.

Assumption 2. The number of consumer that one round of service can accept is constant n, and the price of service is constant w. The service cost, denoted as c, is independent of consumer quantity. The unit price of service is $w = c / \lambda n$, where λ is a profit adjusting factor.

Assumption 3. The number of reserved service is m, m>n, and the probability of every consumer not coming to consume is p. It is independent whether each consumer comes to accept the service.

Assumption 4. The loss brought by each consumer who ordered the service but can not use, namely, the compensation given to GRC, is denoted as constant b.

4. MODEL OF RESOURCE RESERVATION

The economic interest of GRP can be measured by average profit S. Profit s of each round service is represented by price detracting cost incurred by service and compensation possibly created. When there are k GRCs can not come to use the resource in time in m consumers, we have the following equations:

$$s=\begin{cases}(m-k)w-c, & m-k\leqslant n\\ nw-c-(m-k-n)b, & m-k>n\end{cases}\qquad(1)$$

According to assumption 2, the invalid replica quantum K obeys binomial distribution, and then:

$$p_k = P(K=k) = C_m^k p^k q^{m-k},\ q = 1-p \qquad (2)$$

The average interest S, namely the mathematical expectation of s, is:

$$S(m) = \sum_{k=0}^{m-n-1}[(nw-c)-(m-k-n)b]p_k + \sum_{k=m-n}^{m}[(m-k)w-c]p_k \qquad (3)$$

Simplifying Equation (3), and note $\sum_{k=0}^{m} k p_k = mp$, we will get:

$$S(m) = qmw - c - (w+b)\sum_{k=0}^{m-n-1}(m-k-n)p_k \qquad (4)$$

We would try to achieve m for given n，w，c，p to maximize S(m).

Considering the two aspects, economic interest and reputation index, GRP should guarantee that the number of consumers squeezed out is not too many, and hence we can use the probability of consumers squeezed out not more than a certain number to be metric indicator. The probability of consumers squeezed out more than i is $P_i(m)$, equal to there are not more than m-n-i-1consumers unable to come on time in m consumers booking service, so we have:

$$P_i(m)=\sum_{k=i}^{m} p_k \qquad (5)$$

For given n, i, obviously $P_i(m)$ increases monotonously with m.

To sum up, although S (m) and $P_i(m)$ are the two objectives of this optimizing issue, we can give a constraint that $P_i(m)$ will not exceeding a fixed value, try to find a solution of the function whose single objective is S (m).

To reduce the parameters in S(m), choosing H (m) as new target function by dividing S (m) by cost, which means average unit interest. Note that in Assumption 1,there is $w = c / \lambda n$, and from Equation (4), we can get:

$$H(m)=S(m)/c=\frac{1}{\lambda n}[qm-(1+b/w)\sum_{k=0}^{m-n-1}(m-k-n)p_k]-1 \qquad (6)$$

In (6), b /w is the ratio of compensation in service price, so that the issue will be trying to achieve m maximizing H (m) for given λ，n，p，b /w, and then the constraint is:

$$P_i(m) = \sum_{k=i}^{m} p_k \le \alpha \qquad (7)$$

α is a plus quantity less than 1.

5. A SOLUTION OF THE MODEL

Because the model (6) (7) can not analytically solved, we set several groups of data to be calculated by software MATLAB.

We set n = 400，λ= 0.6，p = 0.05，when m is changed from 400 to 430, b /w uses 0.2 and 0.4, respectively, the result is shown in Fig.1.

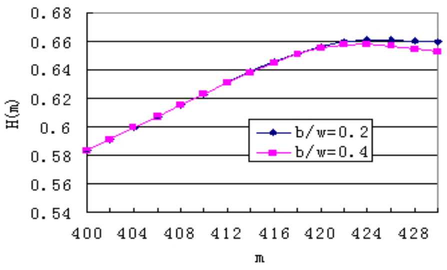


Fig.1. H (m) variation with m when n = 400

When m changes from 400 to 430, and b /w gets 0.2, we compute the probability of consumers squeezed out from 5 to 10. The result is described in Fig.2.

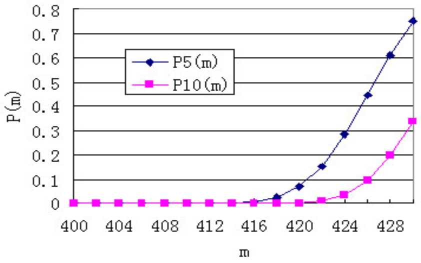


Fig.2. P (m) variation with m when n = 400

In addition, we set up n = 300, not changing other parameters, the variation of H (m) with m and the variation of P (m) with m are similar. We have not figured them for brevity.

We analyze the experiment result as follows:

1) For every n, p, b/w，The variation of H (m), the average profit, with m, are firstly increase, and then decrease. The change is small near maximal value, while the probability of consumers who visit invalid data replication exceeding 5 and 10, P_5 (m) and P_{10} (m) increase more rapidly, so we should reference the maximal value of H (m), giving an acceptable α in the constraint which is Equation (7) and deciding appropriate m.
2) For given n and p, when b / w increases from 0.2 to 0.4, the decrease of H (m) does not exceed 2%. And so in order to improve reputation index, we might as well pay relatively more compensation to the consumers squeezed out.
3) Taking economic interest and reputation index of GRP into account, we would like to confine P_5 (m)<0.3 and P_{10} (m)<0. 1. From Fig.1 and Fig.2, when n = 400，if p is estimated as 0.05，we will get m = 424.

6. CONCLUSIONS

Existing resource reservation mechanism is not flexible in Grid. We propose a mathematic model which can determine the quantity of consumers who order the service in advance. In this model, taking economic interest and reputation index of service supplier into account, we simplify an optimization problem of two objectives under hypothesis in reason. Through numeric calculation and analysis, we show, when the consumers of service are confined in a range, there is an optimized point for the consumer quantity which will maximize the average profit. Because there is no punishment factor, some hackers may order resource with evil intent, leading to real consumers not achieve service in time. In the further research, consumers' faith will be taken into account, and different resource reservation strategies will be established according to different guests.

REFERENCES

[1] R Buyya, D Abramson, and J Giddy. "A Case for Economy Grid Architecture for Service-Oriented Grid Computing." *Proceedings of the International Parallel and Distributed Processing Symposium*: 10th IEEE International Heterogeneous Computing Workshop (HCW 2001), San Francisco, USA, 2001.

[2] R Buyya, D Abramson, and J Giddy. "An Economy Driven Resource Management Architecture for Global Computational Power Grids." In *proceedings of the 2000 International Conference on Parallel and Distributed Processing Techniques and Applications* (PDPTA 2000), Las Vegas, USA, CSREA Press, USA, 2000.

[3] I Foster, C Kesselman, C Lee, R Lindell. "A Distributed Resource Management Architecture that Supports Advance Reservations and Co-Allocation." *Proceedings of the International Workshop on Quality of Service*, Brighton, 1999.

[4] L O Burchard, H U Heiss, C De Rose. "Performance Issues of Bandwidth Reservations for Grid Computing." *Proceedings of 15th Symposium on Computer Architecture and High Performance Computing*. Los Alamjtors: IEEE Computer Society, 2003:82-90.

[5] http://www.globus.org.

[6] G Medvinsky, C Neuman. "NetCash: A Design for Practical Electronic Currency on the Internet." *Proceedings of 1st the ACM Conference on Computer and Communication Security,* Berlin, 1993.

Jing Li is a doctor student of Computer College in Chongqing University, an instructor in Chongqing Education College. She graduated from Chongqing University in 2005 and got the master's degree with specialty of computer software and theory.

Shuyu Chen is a Full Professor and dean of Software Engineering College in Chongqing University. He graduated from Chongqing University in 2001 and got the doctor's degree with specialty of computer software and theory.

A Bisection Scheduling Algorithm Using Run-time Prediction Based on Grid Computing

Gongli Li, Li Chen, Dan Li, Guangwei Wang
The Department of computer science of Hua Zhong Normal University
Wuhan, Hubei, 430079, China
E-Mail: cleanyaa@163.com

ABSTRACT

Grid computing is a major research area with increasing impact on the scientific and business field. Since a grid connects large-scale geographically distributed computers, scheduling in such environment is a complex undertaking. In order to simply the scheduling work, we propose a resource filter policy to choose appropriate resource set for the task firstly. Then we present a bisection scheduling policy that uses predicted mean and variance resource capacity information to make task-mapping decisions. This can not only simplify the large-scale resource scheduling issues but also reduce the task execution time

Keywords: Grid Computing, Resource Filter, Bisection Scheduling Algorithm, Run-time Prediction

1. INTRODUCTION

Grid computing [1,2] aims to integrate idle computational power over the Internet and provide powerful computation capability for users all over the world [3]. Grid computing gives users access to widely distributed networks of computing resource to solve large-scale tasks such as scientific computation. It is anticipated that grid computing will develop to provide high performance computing capabilities. But the gap between the high expectations and available real-life grid applications is still remarkable, so the substantial work has been spent on improving the grid performance.

In order to take full advantage of grid computing resource, it is important to employ efficient task scheduling algorithms. The scheduling problem on grids has been widely studied [4, 5, 6]. In this paper, we focus on the divisible task [7,8], that the submitted tasks can be divided into arbitrary chunks, and base on this, we propose the bisection scheduling algorithm, that uses predicted mean and variance resource capacity information to make task-mapping decisions.

In most cases, to guarantee the system performance, the scheduler is required to predict the performance of tasks on the various resources. Effective use of heterogeneous and dynamic grid systems requires new approaches to performance prediction.

However, accurate predictions of run times are difficult to achieve for parallel applications running in shared environments where resource capacities can change dynamically over time. The PACE toolkit[9] we used in this paper can combine application and resource models at run time to produce a high level accuracy performance data.

The rest of the paper is organized as follows: In section 2, we describe the grid computing model we will concern. In Section 3, the scheduling policy is described on the model. In Sectin4, the d the efficiency of scheduling policy is discussed. Finally, we conclude our paper and discuss our future work is Section5.

2. SCHEDULING MODEL

The model architecture our scheduling algorithm based on is shown inFig.1.
We use the PACE toolkit to estimate execution times for a divisible task on a given set of architecture types prior to run-time. And it also contains others submoduels, such as task management, task execution, and resource monitoring

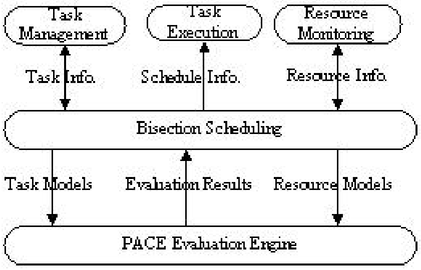


Fig.1. The Bisection Scheduler Based on PACE Toolkit

Task management receives the task and the information of the task from user, gives a unique identification number to each task and then awaits the attention of the scheduler. Task management also adds, deletes or inserts tasks in the task queue.

Task execution is responsible for executing the program which associated with a task on a scheduled list of processors.
Resource monitoring is responsible for gathering statistics of the process nodes. These statistics may include availability, load average and idle time. In this system the resource monitor queries each node every five minutes. And then this is provided to scheduler.

The main components of the PACE toolkit include application tools, resource tools, and an evaluation engine. The PACE evaluation engine can combine application and resource models at run time to produce performance data (such as total execution time).While PACE evaluations complete relatively quickly (usually in the order of a few tenths of a second) [10].

PACE models are modular and are constructed by a set of hierarchical layer. At the uppermost level, applications are captured as a sequence of parallel subtasks; where each subtask is subsequently described using control flow and resource usage information. The parallel characteristics of the subtasks are described using parallelization templates which allow the computation communication interactions between processors to be described. At the lowest level in the model the target hardware is characterized. Fig.2 illustrates the main components of the PACE toolkit [11].

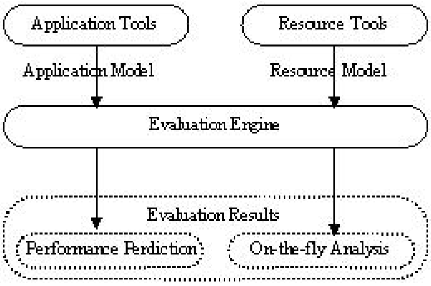


Fig.2. The PACE Toolkit

3. SCHEDULING ALGORITHM

In this work, our scheduling algorithm based on the PACE performance predictions.

Consider a grid resource set P with n computing nodes. Suppose that w_i is the performance information of each computing node p_i and that represents the number of unit tasks the p_i can process per minute. The parameter τ_i represents the start time of computing node p_i. If the node is p_i idle, $\tau_i = 0$, and it also means that the p_i can start the task at once and needn't to wait.

$P = \{p_i \mid i=1, 2, \ldots\ldots, n\}$ (1)

$w = \{w_i \mid i=1, 2, \ldots\ldots, n\}$ (2)

$\tau = \{\tau_i \mid i=1, 2, \ldots\ldots, n\}$ (3)

The task T is considered to be run on P, and the task we discussed here is divisible. It means that the task can be divided into arbitrary subtasks. The number of subtasks lies on the system state and the task scale. Assume that the task T is divided into m subtasks.

$T = \{T_j \mid j=1, 2, \ldots\ldots, m\}$ (4)

When the task is submitted, the user has to specify with the related information, such as task scale ε and the execution deadline π from the user.

The execution time should contain the computing time and the communicating time, so the execution time T_{exe} can be defined as:

$T_{exe} = T_{comp} + T_{comm}$ (5)

T_{comp} is related to the scale of the subtask and the performance of the executing node and T_{comm} is related to the traffic of the subtask and the network condition.

The goal of schedule is to minimize the execution time and the user's deadline should also be satisfied as far as possible. Before we schedule the task, we use the resource filter policy to filter some resources that doesn't meet the user's qualification and select a possible set of resources as the schedule object. This eliminates resources from being considered, so it can simplify the schedule work. The resource filter algorithm is described as below:

```
Resource filter:
    For i=1 to n
        Compute $\tau_i$ with PACE;
        If $\tau_i$ =0    Add $P_i$  To PE1;
        Else   If $\tau_i < \pi$
      Add $P_i$  To PE2;
        Endif
    Endif
  Endfor
End
```

After we select the appropriate resources set, we can begin to schedule the task.

Just as we discussed previously, the task execution time should contain the computing time and the communicating time, if schedule the task to all of the possible resources, the computing time can be reduced, but the communicating time among the executing nodes may increase dramatically. So we propose a bisection scheduling algorithm to find an optimal resource set to minimize the total execution time

```
  Bisection schedule:
    Compute $w_i$ for PE1with PACE;
    Queue { $w_1, w_2, \ldots, w_k$ } ;
      t1=schedule T in { $p_1, p_2 \ ,\ldots, p_k$ };
      t2= schedule T in { $p_1, p_2 \ ,\ldots, p_{k/2}$ };
      If    t2≤t1
        P= { $p_1, p_2 \ ,\ldots, p_{k/2}$ };
        Schedule T via bisection until t2>t1;
        Process T with P;
    Else
      Compute $\tau_i$, $w_i$ for PE2 with PACE;
      Queue { $\tau_1, \tau_2, \ldots, \tau_l$ } ;
      P=P1+P2 ;
      Schedule T in P via bisection;
  Endif
End
```

4. EXPERIMENTAL VAILDATION

We compare our scheduling approach, bisection algorithm with other two methods, one is that schedules the task on all of the possible resources, and another one is that just uses the idle resources ($\tau_i = 0$) to process and never waits for any resources.

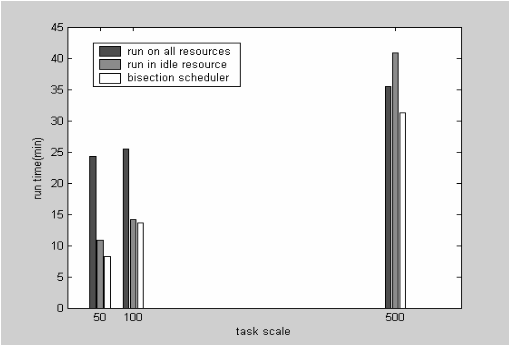


Fig.3. Performance comparison between different scheduling policies

Fig.3 shows the performance of our bisection scheduling with other two policies we introduced previously. Fig.3 presents that, whatever the task scale is large or small, the bisection scheduling policies gains the shortest running time.

5. CONCLUSIONS AND EUTURE WORK

In this paper, we proposed a bisection scheduling policy which based on the performance prediction of the system. Our algorithm aims to minimize the execution time, while satisfy the user's require as far as possible. In order to simplify the scheduling procedure, we use the resource filter to select possible resources set, and then the scheduling work just needs to take into account of this set.

This paper has studied a task distribution environment where tasks are arbitrarily divisible. How to extend our approach to the more general environments is an interesting issue. And thinking of more tasks qualifications, such as needing several kinds of resources, existing conflicts of resource, is a nature extension of this work

REFERENCES

[1] I. Foster and C. Kesselman,“The Grid:Blueprint for New Computing Infrastructure,” Morgan Kaufmannn, 1999.

[2] I. Foster , C. Kesselman, et al., “The Anatomy of the Grid: Enabling Scalable Virtual Organizations,”in *International Journal of High Performance Computing Applications*, vol.15, 2001,pp.200-222.

[3] *Grid computing: Making the Global Infrastructure a Reality*,F.Berman,G. Fox,and T.HEY, EDS.John Wiley and Sons,2003.

[4] A. Galstyan, K. Czajkowski, and K. Lerman,“ Resource Allocation in the Grid Using Reinforcement Learning,” http://www.isi.edu/lerman/papers/papers.html , 2003.

[5] H. Casanova, A. Legrand, D. Zagorodnov, and F. Berman, “Heuristics for Scheduling Parameter Sweep Applications in Grid Environments,”in P*roc. Ninth Heterogeneous Computing Workshop (HCW 2000)*, May 2000, pp.349-363.

[6] H. Casanova, J. Hayes, and Y. Yang, “Algorithms and Software to Schedule and Deploy Independent Tasks in Grid Environments,”in *Proc. Workshop Distributed Computing, Meta computing, and Resource Globalization*, Dec. 2002.

[7] Y. Yang and H. Casanova, “Umr: A Multi-Round Algorithm for Scheduling Divisible Workloads,”in P*roc. Int'l Parallel and Distributed Processing Symp. (IPDPS '03)*,Apr. 2003.

[8] “Rumr: Robust Scheduling for Divisible Workloads,” in *Proc. 12th IEEE Symp. High Performance and Distributed Computing (HPDC-12)*, June 2003.

[9] G. R. Nudd, D. J. Kerbyson, E. Papaefstathiou, S. C. Perry, J. S. Harper, and D. V. Wilcox, “PACE-A Toolset for the Performance Prediction of Parallel and Distributed System”,*Int. J. High Performance Computing Applications, Special Issues on Performance Modeling Part I*,Vol.14,No.3,2000,pp. 228-251.

[10] [10] Junwei Cao, Daniel P. Spooner, Stephen A. Jarvis, Subhash Saini and Graham R.Nudd, “Agent-Based Grid Load Balancing Using Performance-Driven Task Scheduling” www.dcs.warwick.ac.uk/research/hpsg/documents/CaoJ.ipdps03.pdf

[11] Junwei Cao, Stephen A. Jarvis, Subhash Saini and Graham R. Nuddrd “GridFlow:Workflow Management for Grid Computing”in *Proceedings of 3 IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2003)*,2003,pp.198-205.

Gongli Li is the graduate student of the department of computer science of Hua Zhong Normal University. The research field is distributed computing and computing grid.

Li Chen is the processor of the department of computer science of Hua Zhong Normal University. The research field is software engineering and software architecture.

The Access and Integration of Database Grid Based on DMMS

Jianshe Dong, Jingrong Li, Qiuyu Zhang, Sanjun Sui, Zhi Wang
College of Computer and Communication, Lanzhou University of Technology
Lanzhou, GanSu Province, China
Email: dongjs@lut.cn, sunny2006@mail2.lut.cn, zhangqy@lut.cn

ABSTRACT

The information ages bring about the need of lots of information, while the Grid developments in recent years have provided enhanced support for the exchange and share of information of heterogeneous platforms and databases in large-scale domains, but how to get efficiently the information needed by users is still a problem. The access and integration of database Grid is a key issue in a Data Grid. This paper introduces a novel scheme for accessing and integrating of database Grid based on Double Middleware Model System (DMMS). This model achieves secure resource share with high-performance based on GridFTP by utilizing flexibly extensive language XML.

Key words: Grid, Middleware, DAI, OGSA, DMMS

1. INTRODUCTION

Data resource plays an important role in many fields. How to efficiently integrate all kinds resources of hardware and software which distribute in different fields and platforms, computers and databases, and how to provide uniform and flexible access are urgent problems of the development of the whole Internet. Grid was defined as flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources—what were referred to virtual organizations [1]. From above, the characteristics of the Grid meet the needs very well.

The present DBMSs are autonomic, heterogeneous and dispersed, and can not satisfy the needs of providing uniform interfaces to access flexibly very well. Using Grid technology to integrate and to access database resource will help us to utilize the present database resources efficiently. Grid based database technology has been researched in many scientist institutions. The most distinguished of such researches are spitfire project [2] that affiliates with European Data Grid, and OGSA-DAI[3] that affiliates with UK e-Science program.

In this paper, we present a useful and high-performance middleware, which is not only for scientific research and for industry, but for commerce and civilian combining with OGSA and Globus Toolkits 4.0.4[4]. This middleware utilizes flexibly extensive language XML which has high ability in describing data, and communicates with terminal client through HTTP, FTP, SMTP and SOAP, and communicates and transfers data with Grid environment by GridFTP in order to achieve high-performance, secure, convenient resource share.

2. PRESENT RESEARCH STATUS

Since the Grid technology came into being, the research on combining Grid and database has been a hotspot. The following two directions have been formed:

1) How to integrate the present databases into Grid almost without modifying anything, in order to protect and utilize the present data resource. Many researchers and experts are working in this field, thence formed the nomenclature database Grid. So-called database Grid is a Grid environment in which databases are the main resources, to screen the heterogeneous platforms and databases by middleware which provides clients with uniform interface.
2) How to utilize database technology in Grid environment to establish Grid database. In this field, some large companies have developed databases which support Grid.

For the later, IBM has developed DB2 database which support Grid, and Oracle has developed Grid enabled database Oracle10g. For the former, most of the present databases do not support Grid, but how to utilize the present database efficiently? Some research institutions have made some progresses in this domain.

3. THE DEFICIENCIES OF THE PRESENT DATABASE GRID

By researching into the specialties of the database Grid and analyzing on the present items, and synthesizing the research articles of David Berry, Allen Leniewski and Malcolm P Atkinson [5][6] and so on, referring to the latest standards [7][8][9] of the GGF (Global Grid Forum) DAI Work Group, the problems of the database Grid are as followed:

There are different kinds of database productions at present, such as mesh database, relation database and so on. Even for the same kind of database, there are different database productions from different company. They have their own data descriptions and manipulation modes, and have not uniform access interfaces, either. The job of integrating these different kinds of database productions into Grid, keeping the function of the present database, and providing uniform interfaces to clients is an imperative duty of the research.

Vitalizing data resource, drawing the needed data and assembling virtual resources are also problems. When data resources have been drawn, because different databases have different data description and manipulation mode, we should standardize these data, aggregate these data, assemble virtual data source.

Distributed databases can not coordinate very well. Guaranteeing the ACID characteristics of the transaction in the Grid environment is also very important. Because of the complexity and dynamic of the Grid environment, it may cause database inconsistent state and out of integrality without ensuring the ACID characteristics of the transaction. If a database is in inconsistent state, we should take appropriate strategy to restore it into consistent.

4. PROBLEM SOLUTION

For the sake of achieve uniform interface, we introduce two approaches to solve the above problems:

Firstly, developing uniform database interfaces based on ODBC/JDBC. Since the technology of ODBC/JDBC is very mature, it is used as a main approach for open application programming interface. And it provides core functions of almost all the relational database. From above, utilizing middleware to connect database with Grid environment based on ODBC/JDBC is feasible. For instance, spitfire project manipulated data based on the following structure, as shown in Fig.1. It implements the communication between clients and middleware by the protocol of HTTS+GIS, and transfers data based on the XML mode, and achieves middleware services by Java Servlet. The work flow of spitfire is that middleware receives the XML request form clients, and then invokes JDBC after some proper transform, and drives database fulfill clients' requests operations. When JDBC return results, middleware transfers it into XML document and sends to the clients. Although this approach costs less, and can satisfy the requirement of almost all the application programs, it has some defects. At first, it may destroy the uniformity of the data interface. JDBC only supports core functions of most relational database, if it happened that functions that JDBC do not support, we need to develop private interfaces of those database, which will destroy the transplant ability of the interfaces which across the databases. Secondly, this approach may destroy virtual data resources. Spitfire uses URL to locate the data resource, and then it will destroy virtual data resources. That is feasible in science research. But in commerce and civilian, the exposed data resource would be easily been attacked by the hackers, and would not be convenient for common clients, either. Thirdly, database integration, which is based on OGSA by packing databases into service, utilizes middleware to provide uniform interfaces. DAI-WG has released web services data access and Integration specification – the relational realization (WS-DAIR) specification, version 1.0, and web services data access and integration specification – the XML realization (WS-DAIX) specification, version 1.0. We should take the advantage of WS-DAIX and WS-DAIR, but not to utilize the limit databases interfaces of WS-DAI Core. Although the latest Globus Tookit4.0.4 has contained WS-DAI to file, XML databases and relational databases, these kinds of implementation only use WS-DAI Core, which not only makes the developing and debugging procedure complex, but also has not provided all the interfaces of XML databases and relational databases. From above discussions, we present the access and integration of database Grid based on double middleware model system as shown in Fig.2.

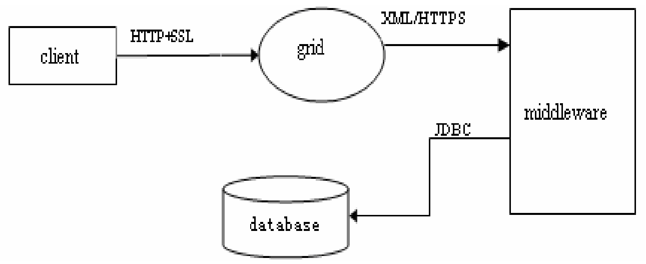


Fig.1. Database integration model of spitfire

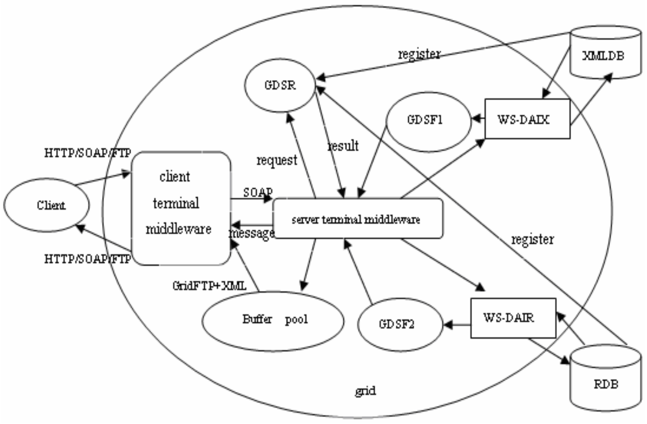


Fig.2. Double Middleware Model System

First of all, for the sake of screening the heterogeneity of the interior Grid environment and without altering the application programs of clients to provide uniform manipulation interfaces to clients, the clients could utilize data resources efficiently and conveniently without training. The client middleware grants client to apply for resources by protocols HTTP, FTP, SOAP, and so on. The client middleware transfers and packs these requirements into uniform XML format，and then sends it to the server terminal middleware by SOAP protocol. After receiving these requirements through uniform interfaces, the server terminal middleware queries GDSR to select the type of database, then sends request to GDSF1 and GDSF2. Considering different database types, GDSF establishes GDS1 and GDS2 to deal with different databases operations to utilize WS-DAIX interface and WS-DAIR interface in the business logic layer to submit applications to XML databases and relational databases. The results will be returned to different GDS by WS-DAIX and WS-DAIR. At the same time, this server terminal middleware is informed of transferring data by many times, not by one time. So server terminal could transfer data once or more than once to integrate into logic data aggregate, and then, sends the data to a data buffer pool to achieve virtual data resource. When the number of data is large, we have not to wait, for all of the data has been prepared, but could make clients get parts of the data at first to improve the network traffic and to reduce the waiting time of the clients. Server terminal middleware simultaneously sends a message to that client middleware to inform the client middleware of fetching data from buffer pool. In that case, it can help server terminal middleware to improve concurrency of transaction. Until server terminal middleware had integrated all the data, it would not send a finish message to clients to inform the clients of that the data has been disposed. Then client middleware sends the data to clients by client's protocols, so clients could access Grid resources without changing clients application programs. Client middleware could ask server middleware to transfer data again if there is something wrong with transferring data to clients. To sum up, in this way, we realized the clients middleware and server terminal middleware coordinated to access and integrate database resources securely and effectively.

5. CONCLUSIONS

From a long-term perspective, it is more meaningful in the Grid development to access and integrate database resources based on DMMS, not on ODBC/JDBC. But because the Grid environment is very complexity and the speed of the network is not so high, some technologies still need research. At present, the research on Grid is still in going. The direction of the grid is to satisfy the needs of the scientific research, industry, commerce and civilian as a general platform. Our double middleware model is introduced just for these requirements and can utilize data share on hybrid platform.

REFERENCES

[1] Foster I, Kesselman C, Tsudik G et al. “The Anatomy of the Grid,” *International Journal of Supercomputer Applications*, March 2001,15(3): 200-222.

[2] https://webafs3.cern.ch/hep-proj-spitfire/.

[3] http://www.nesc.ac.uk/.

[4] http://www.globus.org/toolkit/.

[5] *OGSA Data Architecture Service*, GGF18, 2006.

[6] *Access Service for RDF Data Resource*, GGF18, 2006

[7] Web Services Data Access and Integration-The Core (WS-DAI) Specification, Version1.0, http://www.ggf.org/ documents/ GFD.74.pdf.

[8] Web Services Data Access and Integration-The XML Realization (WS-DAIX) Specification, Version1.0. http://www. ggf.org/documnts/GFD.75.pdf.

[9] Web Services Data Access and Integration-The Relational Realization (WS- DAIR) Specification, Version1.0. http://www.ggf.org/doc-uments/GFD.76.pdf.

Jianshe Dong received M.Sc. in Communication and Information System in 2003 at Lanzhou University of Technology. From 2004 to now he is a Ph.D. student at Lanzhou University of Technology. His main research interests include anti-spam technology, distribute system and Image analyses.

Jingrong Li, student of master. Born in Baiyin, Gansu province in 1983, she has published some academic papers .Her research interests include: grid middleware, grid database and parallel computing.

Function-level Virtual-Machine-based E-Lab *

Lihua Ai, Siwei Luo
School of Computer and Information Technology, Beijing Jiaotong University
Beijing, 100044, China
Email: {lhai, swluo}@bjtu.edu.cn

ABSTRACT

The paper proposes a concept of function-level virtual machine which is ultimate of grid computing. Further an E-Lab model constructed from function-level virtual machine is presented, which plays an important role in compensating for the lack of device investment and making rare or expensive equipment maximum utilities. The architecture framework is detailed for the E-Lab model, which instantiate the function-level virtual machine extended from process level and system level virtual machines. The E-Lab can be reached just from a browser. However, the learner can utilize kinds of equipment.

Keywords: Grid Computing, Virtual Machine, E-Lab, Model, E-Education

1. INTRODUCTION

With the increasing performance of network infrastructures and the development of grid computing [1], it is possible to construct virtual machine providing and managing heterogeneous and distributed resources. On the one hand, the research of the virtual machine would create new opportunities to novel computer architectures so as to overcome the bottleneck of existing ones. On the other hand, virtual machine provides a prospective platform for distant electronic laboratory (short for e-Lab).

As the popularity of e-learning, people have adapted to the convenience of network electronic educating, this lays good foundation of e-Lab. What is more that e-Lab can compensate for the lack of device investment and make rare or expensive equipment maximum utilities.

All those show that e-lab has become a demand at current era. It is necessary to find an appropriate implementation approach. We propose an e-Lab model from the perspective of organization-level virtual machine which is based on grid.

2. RELATED WORK

Leleve L et al. [2] propose a start of reflection on how E-Labs could comply with emerging e-content standards. They exploit the idea out of TIPY, which is a web-based training platform. Bardeen M. et al. [3] describes a case study that uses grid computing techniques to support the collaborative learning of high school students investigating cosmic rays.

Heinz [4] details that the difference between Grid and Web computing has become less evident with the introduction of the Open Grid Service Architecture and the definition of a Grid service. In OGSA as well as in the Web Services Resource Framework, a Grid service is basically a Web service with some additions to make it persistent (that is, it's able to store state information persistently rather than transiently at the server beyond the lifetime of a single request). Grid technology frequently uses Internet as well as Web service technology. It provides service add-ons and therefore extends the field of web computing.

We think virtual machine is the ultimate evolution of grid computing. So we propose E-Lab implementation model based on virtual machine technology.

3. VIRTUAL MACHINE TECHNOLOGY

In general terms, a virtual machine in computer science is software that creates an environment between the computer platform and the end user in which the end user can operate software [5].

3.1 Existing Virtual Machine

James E.Smith and Ravi Nair [6] indicate that there are process-level and system-level virtual machines. Fig.1 shows a process virtual machine. Where, virtualizing software translates a set of OS (Operating System) and user-level instructions composing one platform to another, forming a process virtual machine capable of executing programs developed for a different OS and a different ISA (Instruction Set Architecture). Fig.2 shows a system virtual machine. Here, virtualizing software translates the ISA used by one hardware platform to another, forming a system virtual machine, capable of executing a system software environment developed for a different set of hardware.

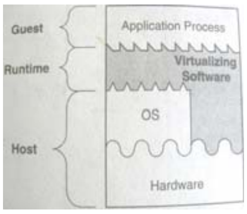


Fig.1. A Process Virtual Machine

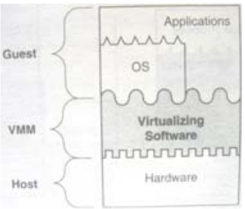


Fig.2. A System Virtual Machine

The existing virtual machines care for the virtualization of ISA and OS. We are interested in upper level applications and functions of the platform. Therefore, we propose function-level virtual machine.

3.2 Function-Level Virtual Machine

Fig.3 presents a function-level virtual machine. In this architecture, virtualizing software translates operation requests from other platform to local, or on the contrary, forming a function-level virtual machine, capable of providing function executing and operation response.

The function-level virtual machine technology is concerned with providing useful building blocks for the construction of software components that can work with one another in the federation of function-level virtual machines. It raises the level of the communication activities of application programs

* Beijing Key Subject Construction Project (XK100040415).

through the support of abstractions such as remote method invocation, communication between a group of processes, notification of events, the partitioning, placement and retrieval of shared data objects and the transmission of multimedia data in real time.

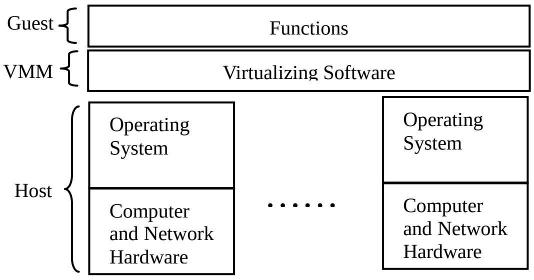


Fig.3. Function-level Virtual Machine

The function-level virtual machine consists of the real hosts providing the original functions. This function integration is versatile, reconfigurable and dynamic in service lifetime. The virtualizing software defines service interfaces exposing functionality as a web service.

3.3 Virtualizing Software Stack

The virtualizing software components characterize the function-level virtual machine. The utilization of the function-level virtual machine is application specific. However, the infrastructure is common for any function-level virtual machine.

The function-level virtual machine is decentralized controlled. However, there should be a trustable federation among the hosts consisting of the function-level virtual machine. The virtualizing software components must address four basic functionalities: execution service, data moving service, information issuing service and security. Fig.4 illustrates the virtualizing software stack.

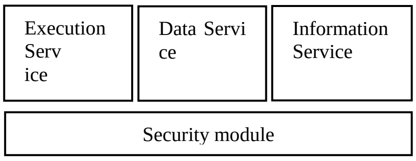


Fig.4. Virtualizing Software Stack

The execution service is concerned with job submission and coordination of remote computations. The data service implement data transfer among hosts in virtual machine. While information service provides index and trigger services.

4. E-Lab Model

E-Lab is a reconfigurable laboratory. All of experiment items are made up of the required function integration of the virtual machine.

4.1 E-Lab Architecture

E-Lab architecture is detailed in view of protocol and service as Fig.5.

The HTTP request information from a client or E-Lab learner is provided to a servlet. The servlet container creates a ServletRequest object and passes it as an argument to the portal service method.

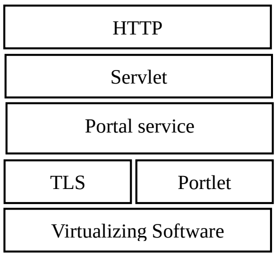


Fig.5. E-Lab Architecture

The portal server providing portal service is an entry point to a set of resources that the virtual machine wants to make available to the portal's users. Besides, the portal also provides personalization, presentation layer of different lab systems, and single sign-on.

The portal server invokes pluggable components, portlets, to produce content that becomes a fragment of a document. The portal server wraps the fragments with the necessary markup and combines the result into a single document which is sent as a response to the access mechanism.

However, before invoking the portlets to utilize the virtual machine resources, the portal should first invoke proxy manager portlet to retrieve the learner's proxy credential. This authentication to the credential proxy server can prove whether the learner in E-Lab is authorized or not. The authentication is processed over TLS (Transport Layer Security) channel. Because the obtained credential is stored in the session, those portlets utilizing the virtual machine resources will automatically pick it up so that the utilization could be implemented via the virtualizing software.

4.2 E-Lab Topology

E-Lab topology is as Fig.6.

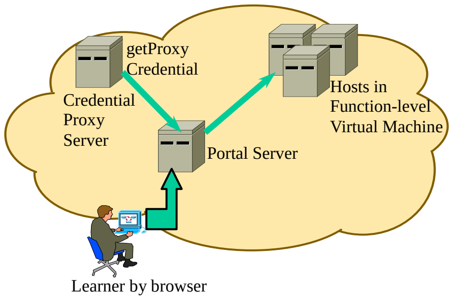


Fig.6. E-Lab Topology

Hosts consisting of the function-level virtual machine could be anywhere in internet on condition that they are installed with the virtualizing software and join the dedicated federation of this specific application. The portal server as the gateway of the virtual machine should also be installed with the virtualizing software and join the federation.

The E-Lab learner can login the function-level virtual machine by browser. After his proxy credential authentication, the learner could utilize the resources of the virtual machine.

The credential proxy server and the security service of the virtualizing software refer to the same authentication information for utilizing resources of the virtual machine. The setup purpose of the credential proxy server is only for the users, such as E-Lab learners, through the portal server.

4.3 E-Lab Implementation Tools and Example Use

We implement E-Lab model by Globus Toolkit4.0 [7] as the virtualization software, Tomcat5.0.28 [8] as the servlet container, GridSphere2.2.8 [9] as the portal and the portlets container, gridportlets1.3.2 [9] as example portlets, and Myproxy [10] as the credential proxy server. We also develop the SPARC portlet to the portal.

We apply the E-Lab to our computer architecture course for Sun SPARC environment. It is well known that Sun SPARC is a typical RISC CPU, which ISA (Instruction Set Architecture) and pipeline techniques can reflect the RISC CPU featured design thoroughly. However this kind of computers is rare in campus for their prices. Therefore, the only four Sun Ultral10 of our department are added to the function-level virtual machine federation. The E-Lab learner can submit the high level language program, such as C source code, to get the compiled assembly code by Sun Ultra10 so that to observe the delay slot fill by compiler schedule policy. The learner could logon E-Lab portal and use the SPARC portlet to submit job, such as Solaris ‘as’, ‘ld’, ‘cc’ etc. command. Then use the SPARC portlet to download the compiled assembly code.

5. CONCLUSIONS

E-Lab can compensate for reality laboratory environment. It is promising for distant education with laboratory. Our E-Lab model is constructed on an open architecture, therefore, has scalability in further. It is necessary to aggregate more lab content portlets in next step.

REFERENCES

[1] I.Foster and C.Kesselman, eds., “The Grid: Blueprint for a New Computing Infrastructure”, Morgan Kaufmann Publisher, 1998

[2] Leleve A., Prevot P., Benmohamed H., et al. “Generic E-LAB Platforms and Elearning Standards,”in *[C]. International Conference on Computer Aided Learning in Engineering Education (CALIE2004)*, Grenoble, France 2004

[3] Bardeen, M. Gilbert, E. Jordan, T. Nepywoda, et al. “The QuarkNet/grid collaborative learning e-Lab,”in [*J*]. *Future Generation Computer Systems*,May 2006, vol22(6),pp700-708

[4] Heinz Stockinger, “Grid Computing: A Critical Discussion on Business Applicability,”in I*EEE Distributed Systems Online*,2006,vol7(6).

[5] http://en.wikipedia.org/wiki/Virtual_machine [EB/OL]. 2007

[6] James E. Smith, Ravi Nair,*Virtual Machines: Versatile Platforms for Systems and Processes [M]*,Elsevier, 2005

[7] http://www.globus.org/toolkit/ [EB/OL] 2007

[8] http://tomcat.apache.org/ [EB/OL] 2007

[9] http://www.gridsphere.org/gridsphere/gridsphere [EB/OL] 2007

[10] http://grid.ncsa.uiuc.edu/myproxy/ [EB/OL] 2007

Lihua Ai is an on-job doctor candidate, lecturer in school of computer and information technology, Beijing Jiaotong University. She is responsible for computer architecture course for college student. Her research interest is grid computing, parallel and distributed processing.

Siwei Luo is a supervisor, professor in school of computer and information technology. He is a vice chairman member of Professional Committee of computer architecture of China Computer Federation. His research fields are grid computing, parallel and distributed processing, and artificial neural networks.

Checkpointing Algorithm in Computation Grid Service System Based on Mobile Agent*

Zhirou Zhang, Xiaohua Zhang, Dawei Dong
Network and Information Center, North China Electric Power University, Beijing 102206, China
Email: zhangzr@ncepu.edu.cn

ABSTRACT

A non-close, non-block [1] and low-overhead checkpointing algorithm is proposed for computational grid service system which provides computation function for mobile agents. This algorithm can make global consistent checkpoints under the condition that mobile agents communicate out of computational grid service system and require fewer control messages. In addition, Synchronous Garbage Collection Mechanism of this algorithm can avoid inconsistent state between local checkpoints which are caused by different commitment time of local checkpoints.

Keywords: Grid, Computational Services, Mobile Agent, Checkpointing Algorithm

1. INTRODUCTION

Grid is referred as the next generation of Internet [1]. Grid computing Technologies enable widespread sharing and coordinated use of networked resources [2]. Grid is changing the ideas of networks, sharing, and collaboration. Academic communities, together with enterprises, are endeavoring to have grid used more widely. And Open Grid Services Architecture (OGSA) [3] has been proposed to standardize Grid architecture. But OGSA only standardize elementary functions and behaviors of grid services for grid' multiformity and flexibility. So how to implement various services and how to access services conveniently to promote efficiency is an important research direction.

Mobile Agents have characters of autonomy, mobility, intelligence, collaboration, and security [4]. They can run and control their own states and actions without users' direction. So, in complicated and changeful grids, Mobile Agents carrying tasks of user or application can decide their paths according to their tasks and network environments at that time, which can get over the disadvantage of process migration that need to be controlled artificially in grid. Therefore, Mobile Agents can be used as carriers of tasks to reduce the load of programmers and users and increase application's adaptability to complex grid environments.

According to the investigation, utilization ratio of computer resources is low in daily routine of most organizations, even if computers are being used. In statistics, processor utilization by PC systems is less than 5% and Servers' CPU utilization is less than 15%[5], which means a great waste of computer's process ability. If these idle resources constitute a computational grid service system that can provide computation services for computation Mobile Agents which carry computation tasks, expense for High Performance Computer can be saved. Of course, because of variableness of networks and nodes, Fault-Tolerance mechanism for computation grid service system that consists of idle computers in an organization must be researched to guarantee computation tasks can be executed normally even if there are faults in networks and nodes and to reduce Fault-Tolerance overhead. So in this paper, a non-close, non-block and low-overhead checkpointing algorithm is proposed. The rest of this paper is organized as follows. Section 2 presents the architecture of computation grid service system which consists of idle computers along with network architecture in an organization. Section 3 proposes this new checkpointing algorithm in detail. Section 4 present experiments and data. Section 5 is a conclusion of this paper.

2. ARCHITECTURE OF COMPUTATION GRID SERVICE SYSTEM CONSISTING OF IDLE COMPUTERS IN AN ORGANIZATION

Fig.1 is the architecture of Computation grid Service System which consists of idle computers in an organization to provide computation services for computation tasks of Mobile Agents and whose management hierarchy is formed with network hierarchy. In this architecture, computation service nodes (idle computers) which are in the same LAN compose a super-computation service and managed by management node of this super-service. And this super-service and other super-services which are in the same layer of networks also compose a service which lies in a higher layer. That management hierarchy is formed with network hierarchy will have computation tasks, which belong to a computation application, assigned in a smaller scope of networks to reduce communication overhead between tasks.

In Fig.1, "Assign Task (1)~(5) is the order of task assignment. Tasks of an application are assigned from top to bottom and then from bottom to top. Firstly these tasks are assigned by Top Management Node to management node of its sub-service, then are assigned to sub-service of this sub-service, as far as to the bottom management node. The bottom management node assign task to computation nodes it manages. If there are remaining tasks unassigned, the super-management node of this bottom node assign them to another bottom management node under it, as far as all tasks are assigned to computation nodes from bottom networks to higher layer networks. In this way, tasks of an application can run in a smaller scope of networks, which reduce communication overhead between them.

3. NON-CLOSE, NON-BLOCK, AND LOW-OVERHEAD CHECKPOINTING ALGORITHM

In the foregoing computation grid service system, networks and nodes may fail. So Fault-Tolerance mechanism must be researched for it. Checkpointing and Rollback algorithm is a classic Fault-Tolerance method. Checkpointing saves states of program periodically or unperiodically to have program recovered to a checkpoint by Rollback mechanism and continue execution on the condition of networks or node

* This work was supported by Research Funds of North China Electric Power University (No. 200622001)

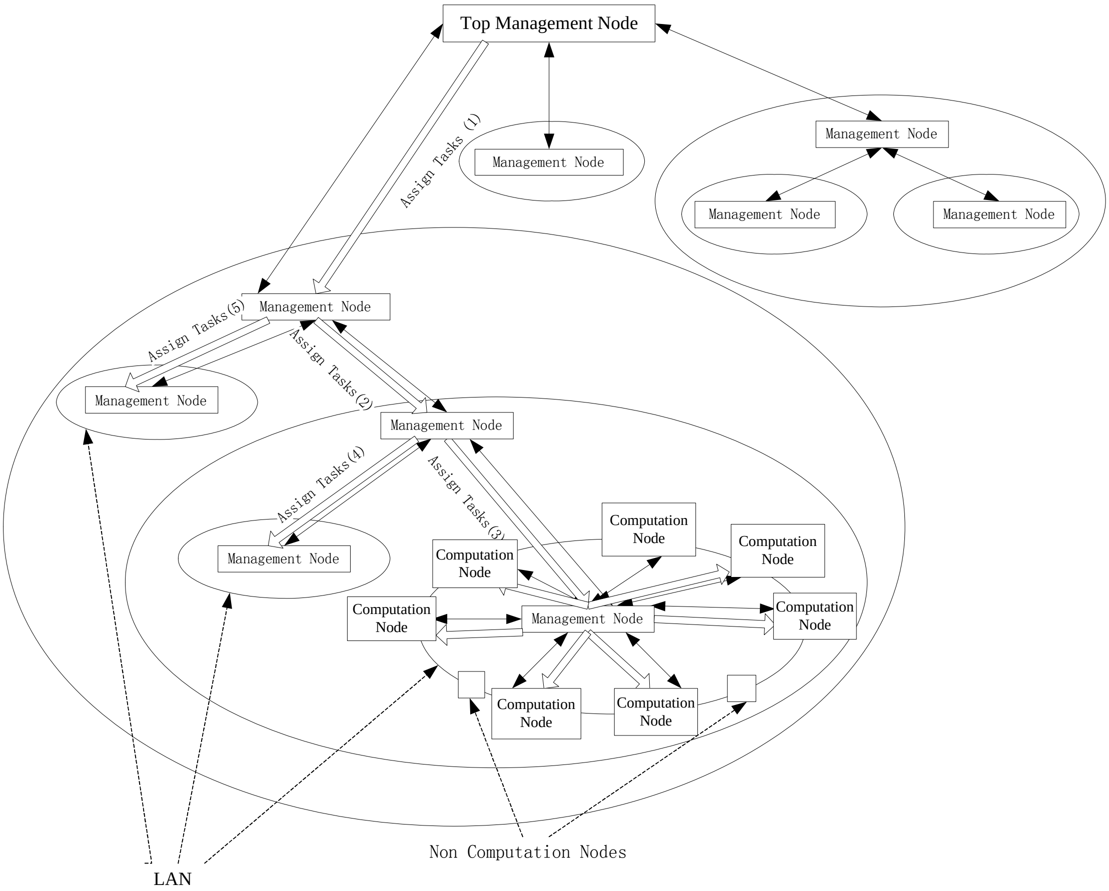


Fig.1. Architecture of Computation grid service system According to Network Hierarchy

failing. Processes of distributed programs need communicate with each other, so when processes are rolled back, messages which are sent by sender aren't be received by receiver, or messages which are received by receiver are thought as unsent by sender.

All of these conditions maybe produce faults for processes. So checkpoints of processes of a program must be in a global consistent state [6], that is, when these local checkpoints are established, messages that were sent have also been received by its destination. A global consistent checkpoint is composed of local checkpoints that are in a global consistent state. When a system rolls back to a global consistent checkpoint, it can continue execution without faults.

According to the architecture of foregoing computation grid service system, checkpoints for it are also constructed with network hierarchy. Checkpointing and rollback services of computation nodes are managed by Fault-Tolerance services that lie on the management nodes of the LAN that these computation nodes lie in. And these Fault-Tolerance services are also managed by Fault-Tolerance services on the higher layer. To manage messages sent by nodes, all messages are attached with checkpoint number (CN) of the latest checkpoint established before messages are sent on behalf of sending time of messages.

Traditional "block" coordinated checkpointing algorithm suspends processes (block) to wait for all messages arriving their destination. So with the expanding of system, communication delays increase, waiting time of processes increase, the overhead of checkpointing increase, and running efficiency of system reduce. The algorithm proposed in this paper is non-block, that is, processes needn't to be suspended. Every processes of program record the number of messages that are sent or received by itself in the interval of current checkpoint (from latest checkpoint to that time) with Message Sending and Receiving Table (MSRT). MSRT records process ID that send or receive messages and the number of messages. In the course of checkpoint establishing, these MSRTs of processes are sent to Top Fault-Tolerance service which computes how many messages should be received by these processes respectively with MSRTs to make sure all messages sent are received. With the results of computation, all processes can't establish local checkpoints until all messages that should be received arrive. During the course of waiting for messages, processes can continue to run, that is they need to be suspended. So this algorithm is "non-block".

Computation Mobile Agent that visit computation service may communicate with application or other Mobile Agents that lie out of computation grid service system, and there isn't Fault-Tolerance mechanism or there are different Fault-Tolerance mechanism out of the system, so the checkpointing algorithm must guarantee communication in and out of grid system don't bring faults when processes need to be recovered with rollback algorithm, that is, this checkpointing algorithm is "non-close". It records messages received or sent to message log of latest checkpoint and subsquent checkpoints, then after process roll back, process can enquire message log to avoid resending message that have been sent by mistake, and process also don't miss messages that have been received with message log that records messages from outside.

To reduce the number of control messages of checkpointing algorithm to reduce communication overhead, Fault-Tolerance services aren't notified at the time of establishing checkpoint, but are notified with checkpoint interval by Top Fault-Tolerance services when the system is initialized. All Fault-Tolerance services begin to establish a checkpoint every checkpoint interval from the time of initialization. When a service node restart or a new node join in the system, it get relative information which includes checkpoint interval, the time from latest established checkpoint to the time of node's joining, the number of next checkpoint that will be established, from its upper management node. The node compute the time of its checkpointing with the 2 fore information and the number of checkpoint it will establish with the last information. If checkpoint interval needs to be changed, Top Fault-Tolerance service notifies all lower Fault-Tolerance service. With this method, the checkpointing algorithm use n control messages (n is the number of process) fewer than traditional checkpointing algorithm, which reduce overhead of algorithm.

The non-close, non-block and low-overhead checkpointing algorithm is following:

1) When a checkpoint needs to be established, Checkpoint and Rollback service check whether this checkpoint has been established or not. If it hasn't, temporary checkpoint is established and a message of checkpoint establishing (Cpestab) is sent to upper Fault-Tolerance service, Cpestab attach a MSRT of the processes that lie on the same node with Checkpoint and Rollback service; if it has, only Cpestab is sent.
2) When Fault-Tolerance services receive Cpestab messages from all lower Fault-Tolerance services and Checkpoint and Rollback services, it compute its MSRT and send Cpestab with this MSRT to its upper Fault-Tolerance service.
3) Top Fault-Tolerance service collects all Cpestab and computes a total MRST, and then notify the number of messages that should be received to processes.
4) If process which receives notice finds that there isn't messages need to receive and recent checkpoint have committed, go to Step 6); If it need to waiting for messages or recent checkpoint haven't committed, go to Step 5)
5) The process of checkpoint and Rollback service wait for messages, and other process can continue execution (computation, communication, and so on). When a message received by a process, if the CN of this message isn't more than current CN (the number of checkpoint latest established) of the process, this message is added to temporary checkpoint which has the same CN with it and subsequent temporary checkpoints and reduce the number of messages that should be received of this process. If the number of messages that should be received of this process is equal to 0 and the latest checkpoint has been committed, turn to Step 6). If the CN of this message is more than current CN， n temporary checkpoints ($n = CN_m - CN_{current} - 1$) are established, where CN_m is CN of this message and $CN_{current}$ is the number of latest temporary checkpoint; and the establishment time of subsequent checkpoint is computed from this time to reduce difference between the establishment time of checkpoint of processes.
6) This temporary checkpoint is committed as formal checkpoint, and Cpcommit message is sent to upper Fault-Tolerance service which send Cpcommit to its upper Fault-Tolerance service when it receives Cpcommit from all lower services.
7) When Top Fault-Tolerance service has received Cpcommit from all lower services, it updates the CN of current committed checkpoint and sends garbage clear message (Gbclear) to lower services. Then checkpoint before current checkpoint is eliminated.

Step 6) and 7) are necessary, which are called Synchronous Garbage Clear mechanism. This mechanism means that a global checkpoint can't be deleted until next global checkpoint is committed (that is, all local checkpoints of this checkpoint are committed). On the contrary, if every process deletes elder local checkpoint immediately when its new local checkpoint has been committed, because local checkpoints which are in the same global checkpoint commit in different time, there will be a fault when processes need roll back: the processes whose new local checkpoint haven't been committed must roll back to elder checkpoint and the processes whose new local checkpoint have been committed and their elder checkpoint have been deleted can't roll back to them. In Garbage Clear mechanism, elder checkpoints of all processes can't be deleted until all new checkpoints are committed, so processes can roll back to them whether their new checkpoints are committed or not. The problem coming from that local checkpoint of processes commit asynchronously hadn't been researched in traditional checkpointing algorithms, so Garbage Clear mechanism is a advantage of this algorithm and a contribution of this paper.

There are 2 figures to show this algorithm clearly. Fig.2 is State Transition Diagram of it. And in Fig.3, 1)~7) are steps in this algorithm, arrows mean transfer of control messages, Checkpoint and Rollback service means all Checkpoint and Rollback services in the nodes, and Fault-Tolerance service means all Fault-Tolerance services on different layers. With this figure, it concludes that this algorithm need 4n (n is the number of processes) control messages, whose message complexity is O(n), where 2n messages are used for checkpoints' global consistent state and other 2n messages are used for Garbage Clear. Traditional checkpointing algorithms need 3n control messages at least. Although our algorithm need 4n messages, only 2n messages are needed to have local checkpoints turn

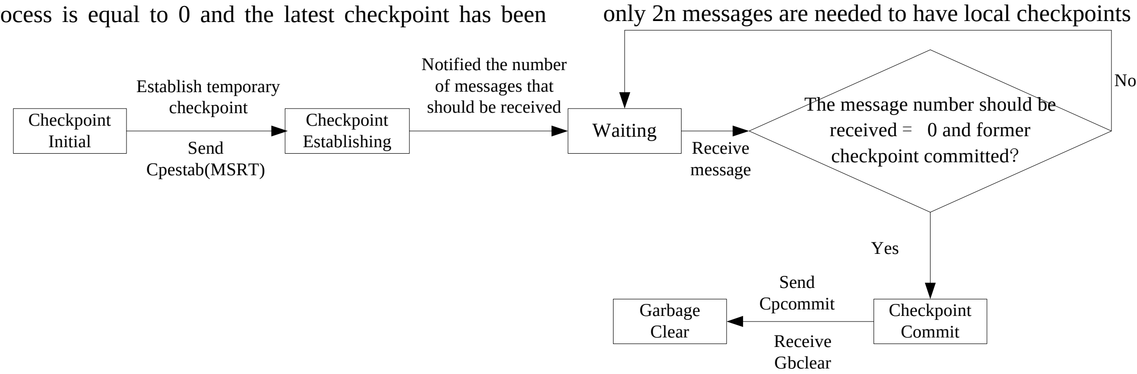

Checkpoint Initial
Establish temporary checkpoint
Send Cpestab(MSRT)
Checkpoint Establishing
Notified the number of messages that should be received
Waiting
Receive message
The message number should be received = 0 and former checkpoint committed?
No
Yes
Checkpoint Commit
Send Cpcommit
Receive Gbclear
Garbage Clear

Fig.2. State Transition Diagram of Checkpoint and Rollback Service

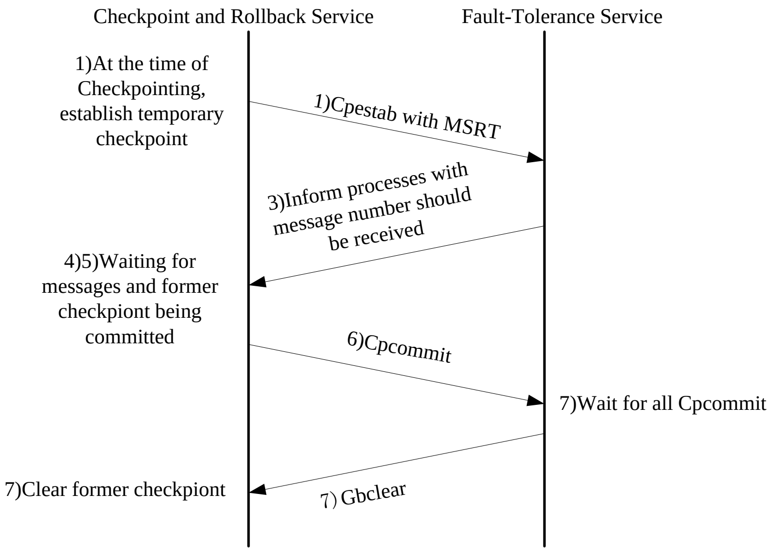

Checkpoint and Rollback Service
Fault-Tolerance Service
1)At the time of Checkpointing, establish temporary checkpoint
1)Cpestab with MSRT
3)Inform processes with message number should be received
4)5)Waiting for messages and former checkpiont being committed
6)Cpcommit
7)Wait for all Cpcommit
7)Clear former checkpiont
7) Gbclear

Fig.3. Steps of the Checkpointing Algorithm

into global consist state, which is fewer than that needed by traditional algorithms and other 2n messages are need for Clear which isn't covered by those algorithms. In this sense, the new algorithm is low-overhead.

4. EXPERIMENT AND DATA ANALYSIS

Experiment on overhead of the new checkpointing algorithm, is respectively done with 2, 4, 6, 8 computers which have the same configuration in 100M Ethernet. The interval of checkpoint is 30s. Parallel computation application used for experiment is power of matrix which is strong-calculation. Figure 4 is average overhead of this algorithm for a global consistent checkpoint, which shows there is near linear relationship between checkpoint overhead and node quantity. So checkpoint overhead of this algorithm don't increase largely with expand of scale of distributed system, which means this algorithm is a good distributed checkpointing algorithm.

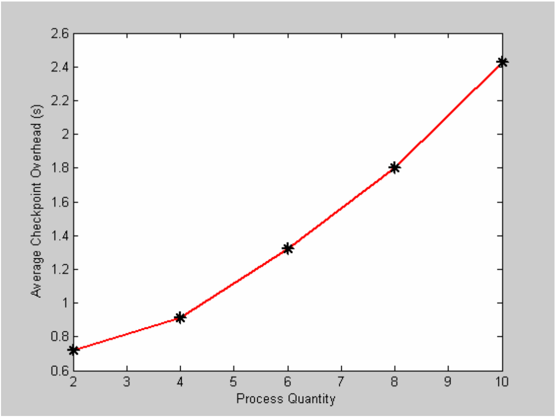


Fig.4. Average Overhead of Global Consistent Checkpoint on the condition of Different Number of Processes

5. CONCLUSIONS

In this paper, a non-close, non-block and low-overhead checkpointing algorithm is proposed for fault-tolerance in computational grid service system which consists of idle computation resources in organization. And this algorithm has Garbage Clear mechanism that guarantee global checkpoint is in consistent state on the condition of local checkpoints being committed asynchronously. The features of it increase efficiency and reduce overhead for the system. Experiment shows the overhead of checkpoint is near linear relationship with the number of processes, so it is a good algorithm.

REFERENCES

[1] Ian Foster, "Internet Computing and the Emerging Grid," http://www.nature.com/nature/webmatters/grid/grid.html, 2000-12-07.

[2] I. Foster and C. Kesselman, editors, *The Grid: Blueprint for a Future Computing Infrastructure*, 1999.

[3] Du Zhihui, Chen Yu, Liu Peng, "OGSA--A Service Based Grid Architecture," http://www.chinagrid.net/grid/talksanddocs.htm, 2002-09

[4] Wang Rong-bo, Zhou Chang-le, "The Study of Mogent (Mobile Agent): An Overview, Application Research of Computers," 2001. Vol 6, 9~11.

[5] Irving Wladawsky-Berger, "The Future Utility Of IT [J]," *Optimize Magazine*, 2002, Issue 13：59.

[6] Rajkumar Buyya (ed.), "High Performance Cluster Computing," Architectures *and Systems*, Volume 1, ISBN 0-13-013784-7, Prentice Hall PTR, NJ, USA, 1999.

Research and Design of Campus Service Grid Application System Based on OGSA*

Lingfu Kong, Guoqing Liu
Department of Computer Science and Engineering，Yanshan University
Qinhuangdao, Hebei, P.R.China
Email: liuguoqingbb@sina.com

ABSTRACT

Service Grid based on OGSA can improve the sharing ability of resource. In order to take full advantage of effective resource of LANs in campus, on the cluster system, the Campus Service Grid Application (CSGA) system is designed based on the Globus. Using the mature grid middleware technique, the system implements the sharing of computing resources and services of LANs in campus. This paper designs the Campus Service Grid Application (CSGA) based on OGSA-GT3, and illustrates the implement scheme, execution flow and implement result. At the end, it discusses the direction of service grid system.

Keywords: Grid Computing, OGSA, Globus, Service Grid, Virtual Organization.

1. INTRODUCTION

Grid technique is the next generation high performance distributed computing technique, it tries to break the restrict of the geographical position and the heterogeneous of platforms, to implement all the resources including computing resources, storage resources, communications resources, software resources and information resources of internet to share completely. It aims to support the resources of virtual organization sharing and cooperating [1]. In simple words, the “Grid Computing” is that it assembles kinds of autonomy resources and systems to utilize the instrument of communications to make kinds of resources join to be seamless, compositive and organic and in order to implement sharing, cooperating and combination computing of resources to offer kinds of services based on grid for user. The character of grid concept is that it can build a computing and management of data platform in the environment of virtual organization which is dynamic and composed by many frameworks [2]. The character of grid resources sharing, cooperating and the management ability of the grid technique to heterogeneous systems and cross organizations systems are very suitable to offer support for grid services system on the basic structure of bottom middleware.

At present, the implemented system of grid environment offers more mature middleware, such as Globus Toolkit, Legion and so on. Using these middleware can joint a set of cooperating resources and build application grid system which meets to special requirement conveniently [3]. At the forepart of grid application, it aimed mostly to the increasing computing demands, especially at the high performance computing problem of the forecast of weather, the modeling of great DNA structure, the movement of celestial bodies, etc., so the data grid and computing grid which can solve high performance computing and science computing have taken the lead in high performance computing. Comparatively, the development of service grid and the business application was weaker. As the third generation application platform of Internet, unless offering general and large-scale services for the domain of industry and business, grid can’t obtain better development, and that the service grid is the only way to apply in industry. Therefore, the standard of grid technique which was Five-Level Sandglass Architecture based on protocol initially developed to be OGSA (Open Grid Services Architecture), which combines with Web Services and takes service as center.

In order to improve the ability of sharing of resource and take full advantage of the computing resources and services existed, in the environment of LANs campus, the Campus Service Grid Application (CSGA) system is designed based on the Globus on the cluster system to meet to the application requirement. This system can develop application programs based on the grid middleware technique to run in the campus service grid according to the patulous application requirement. On the platform of CSGA, the providers of service can deploy and release computing service, also can find the services met to its requirement and use the effective resources on the grid storage array. This paper discusses the development environment of CSGA, the range of application and function character. The second section introduces the Globus Toolkit 3 and OGSA, the third section designs the Architecture of CSGA based on OGSA-GT3 middleware system and discusses its function and character; the forth section brings forward the implement scheme and illustrates the computing service example executed in the CSGA; the last section sums up this paper and points out the research work in the future.

2. GLOBUS TOOLKITS 3 AND OGSA

The Globus project is the most influential project related to grid computing internationally at present. It is a research project of Argonne national Library of USA, initiated at the middle of 1990s. Globus researches broadly at pivotal theory and technology of grid computing, it includes information security, resource management, information service, data management and the environment of application development. And the grid computing grid toolkit which can run at multi-platform is developed to assist at designing and constructing large grid experiment and application platform and developing great application programs suited to large grid system.

At the beginning of 2003, Globus project group associated with IBM together brought forward the Globus Toolkits 3.0. It takes the kernel basic infrastructure as basic and is one of reference implements of OGSI (Open Grid Service Infrastructure). The aim of OGSA is that defining a universal and standard architecture based on grid. OGSA which collects all merits of Web Service regards all of the grid resources as grid services--one special Web Service. OGSA is established based on the SOAP, WSDL, and WSI Web Service technology to

* The research work is supported by Natural Science Fund of Hebei Province in China (Grant No.F2006000281), Research and Development Program of Hebei Province in China (Grant No.062135127).

sustain distributing status management, lightweight examination and discovery, and asynchronous notification. All of the external modules are opened trough the WSDL interface description which derives from the Global service criterion. Because of GT3 providing the software library sustaining grid application programs, this toolkit solves the problems of security, information discovery, resource management, data management, communication, failure examination and transplant. Through adopting the interface definition language of industry standard and Web Service tools, it makes OGSA-GT3 fit into researching the service grid system. At present, many scholar according to different service requirements advanced different service grid hierarchies [4-7], however, their characters were rather accordant.

The OGSA-GT3 programming keeps to the distributed computing schema—proxy-stub model. It includes server and client programming. Server and client is weak coupling, the association between them get across the WSDL service description file. When services are developed, the provider of service must provide the WSDL service description file correspond to the service. The file describes specifically the service interface, method of invoking service and the relation between service invoking and bottom communication protocol. After receiving the WSDL service description file, the client user may build the stub of service invoking, through which to accomplish invoking service.

3. ARCHITECTURE OF CAMPUS SERVICE GRID APPLICATION SYSTEM

According to the requirements of campus services, and using the laying method, we designed the architecture of CSGA, showed in Fig.1.

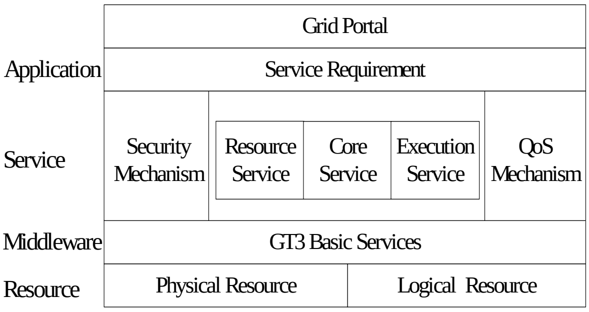


Fig.1. Architecture of OSGA based on OGSA--GT3

The CSGA architecture based on OGSA-GT3 includes Resource, Middleware, Service, Application four layers from bottom to top, their main functions and characters as follows:

(1)Resource: it is the foundation of grid application, constituted by kinds of resources, including physical resources clusters, storage equipments and great instruments, and logical resources net bandwidth, software programs and application services. These resources probably belong to different virtual organizations and having their own accessing policy, So Resource has heterogeneous, distributed and autonomy characters of grid application environment.

(2) Middleware: Being Similar with OGSA in function, this layer offers basic service of GT3, including GRAM (Grid Resource Allocation Manager), which is responsible to solve the execution of computing service requirement using remote resources, allocate requirement resource and monitor services, GASS, which offers support for file accessing and storage of service grid and GridFTP. This layer using the basic services of OGSA-GT3 offers foundation for Service, as well as offering main protocols for communication of heterogeneous resources.

(3) Service: In order to support the development, running, deployment and debugging of grid application which based on service, this layer solves sharing and cooperation of kinds of services. Its core includes three kinds of services—resource service, core service and execution service, and at the same time it offers grid security and QoS (Quality of Service) mechanism. We deploy the service programs written using Java on the GT3 platform to offer services for terminal users. The main characters and differences lie on the encapsulated functions and service objects.

① Resource service: MDS (Monitoring and Discovery Service) is the primary, in which the GIIS (Grid Information Index Service), resource service and special resource interact, coordinate many kinds of resources shared, and collect dynamic and static information of system, support users to look over the status of service running.

② Core service: It offers correlative functions for kinds of services to manage the service deployed in the distributed grid and maintenance of account information of paid service. Moreover, it offers a group of existed service list for virtual organization and fixes on the service kind according the application requirement in the practical application.

③ Execution service: It is used to execute the requirements submitted by user to estimate the system resource is satisfied or not, the loading is balance or not and job schedule policy, it is also used to manage the users and resource certificates.

④ Security mechanism: It is important supplement for resource management protocol of virtual organization; it can offer the security basis support such as identification, authorization and access control for kinds of services. It chooses the GSI (Grid Security Infrastructure) at the aspect of grid security technique. The GSI brings forward many GSP (Grid Security Protocols) to join different virtual organizations and uses security protocol of transport layer to control the security further.

⑤ QoS mechanism: It offers QoS guarantee for Application and Service. Moreover, it carries through the QoS quality arrangement between user and schedule service, offers the ability of QoS end to end guarantee for kinds of services, and can ensures the QoS of many service cooperation application problem.

(4) Application: The function of this layer is that applying kinds of grids and sending service requirements to other layers. After Application sending the service requirement, Service is responsible to find the corresponding service, invoke the remote resources through Service and Middleware, and deal with the corresponding requirement to implement the applied function. Both the service of provider and applicant can send out service requirement. In the CSGA system based OGSA, because of the resource is dynamic and instantaneous, the service requirement may be too.

Grid Portal made of user management server offers uniform operation interface and global operation view.

4. THE IMPLEMENT SCHEME AND EXPERIMENT OF CSGA

Based on the architecture of CSGA and middleware OGSA-GT3 we designed the implement schema of CSGA system met to the requirement, showed in Fig.2.

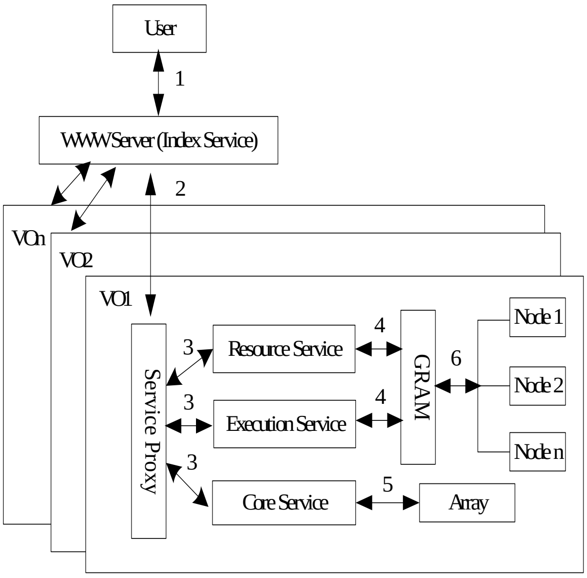


Fig.2. Implement scheme of OSGA

(1) Terminal user runs the HTTP client program, accesses the Web Portal, storages the index service consisted of GIIS of MDS in the WWW server. The index service includes directory services been responsible to static information and database services been responsible to dynamic information. The static information of directory services is consisted of user information including management of user proxy security certification, and resource information, it uses LDAP directory information tree to store. Moreover, the user uses GSI of Globus to certificate the user when the user logs on the system.

(2) After receiving the requirement from user, Web Portal submits it to proxy module, which stored the data list between grid service and registration information of service. When received the requirement of user, service proxy searches the service data list, find out the service registration center, and processes the service requirement and response between applicant and provider of service;

(3) According the information of service response center generally is IP addresses list, the server which executed response service was found out to execute resource service, execution service and core service separately;

(4) If the requirement of user is related with resource and execution service, the basic GRAM of GT3 will be started to collect dynamic and static information, monitor the status of system running, submit the computing task and manage the users. GRAM is responsible to parse and deal with the description of service requirement, and decides to execute or refuse the operation of service requirement, manages the remote monitoring task and updates MDS;

(5) If user brings forward requirement of deploying service, the core service module and the tools GT3 offered will be started to deploy the service in disk list of array, and update the registration information, services list and MDS in corresponding service proxy layer;

(6) Using GRAM of OGSA-GT3 and own allocation algorithm program sends the result of allocation to nodes using GridFTP to execute service requirement, and returns the result to user from one layer to another.

At the period between user submitting the service requirement and receiving the service result, it can find out the execution status using MDS, and also execute operations such as pause, terminate and hanging-up according to the running status of service.

At present, the services running on the CSGA mostly are computing service--aimed at parallel robot emulation system analysis, numerical value computed of finite element and high performance computing, resource monitoring service and the storage and access service of numeric document [8-10]. Next, we give out the result of used the BLAS (Basic Linear Algebra Subrouting) library to test HPL (High Performance Linpack) on the CSGA platform [11-12]. This experiment submits the computing tasks to 8 computing nodes to test the system performance, in the configuration files, N denotes that the computing array is N*N. The result is showed in Fig.3.

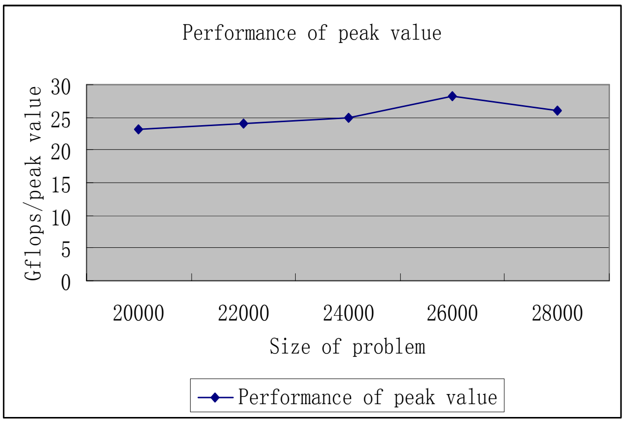


Fig.3. System performance of different problem sizes

5. CONCLUSIONS

Service grid technique is the beginning of that grid is applied in many domains. This paper develops the CSGA faced to service according to campus requirement. According to the present service requirement, we configured resource service, execution service and core service 3 kinds of service, and using OGSA-GT3 middleware discussed the architecture of system and implement schema. Our next research is that adding multimedia service, offering QoS guarantee mechanism policy, improving the performance of system, solving the connection among grids and expanding our services based on OGSA-GT3 middleware.

REFERENCES

[1] I.Foster, C.Kesselman, The Grid 2: *Blueprint for a new Computing Infrastructure,* Morgan Kaufmann Publishers, Inc. 1999

[2] I.Foster, C.Kesselman, S.Tuecke, “The anatomy of the grid: enabling scalable virtual organizations”, *Supercomputer Application,* Vol.15, No.3, 2001, pp.200~222.

[3] Wang bin, Yu huashan, Li Hongzhong, Xu zhuoqun, “Design and Implementation of a Scientific Computing Service Grid Application System”, *MINI-MICRO SYSTEM*, Vol.26, No.28, 2005, pp.1318~1321.

[4] Hu Chunming, Huai Jinpeng, Sun Hailong, “Web Service-Based Grid Architecture and Its Supporting

Environment”, Journal of Software, Vol.15, No.7, 2004, pp.1064 ~1073.

[5] Hu Likai, R&D of the Grid Service Oriented Middleware for the Integration of Designing Software, Master thesis, Zhejiang University, 2004.

[6] I.Foster, C.Kesselman, S.Tuecke, “The anatomy of the grid: Enabling Scalable Virtual Organizations”, Journal of Supercomputer Applications, Vol.15, No.3, 2001, pp.1-10.

[7] F.Berman, G.Fox, T.Hey, Grid Computing: Making the Global Infrastructure a Reality, John Wiley & Sons, Ltd., New York, 2003, 9~21.

[8] Li hao, Kong lingfu, Chen Jing, “LINPACK Performance Analysis of SMP Cluster System”, High Technology Letters, Vol.14, 2004, pp.13~16.

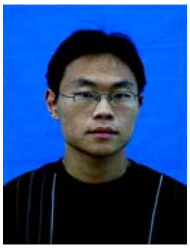
[9] Chen Jing, Kong Lingfu, Wang Xuan, “Study on Optimizing Data Access Strategies in the Environment of Data Grid,” *Proceedings of the 8th International Conference for Young Computer Scientists*, Bijing, 2005.

[10] Kong Lingfu, Jin Jing, Chen Jing, Zhang Xiaoyan, “System of Grid Resource Monitoring Service”, *Communication and Computer*, Vol.1, 2006, pp.28~31.

[11] HPL. http://www.netlib.org/benchmark/hpl/ [EB/OL].

[12] K. Goto BLAS Library. http://www.cs.utexas.edu/users/kgoto/ [EB/OL].

Lingfu Kong is a Full Professor, Doctor Advisor and a president assistant of Yanshan University. He graduated from Harbin Institute of Technology and received his Ph.D. degree in 1995. His main research interests include research of parallel robot and parallel machine tool, research and development of great database system based on C/S and B/S, research of parallel/distributed computing system and research of intelligence information processing. At the aspect of parallel robot and parallel machine tool researching, he has held four projects such as “863” national high technology research and development project, and wrote one monograph which obtained the first award of advancement of science and technology in He Bei Province. At other aspects, he also acquired important achievements and breakthroughs. Now he is one of the directors of CAAI and assistant manager of Computer Academy of He Bei Province. In recent years, he has published about 60 papers on international and national academic publications and important international academic conferences.



Guoqing Liu is a post graduated; his main research interests include gird computing and parallel computing.

Resource Management and Scheduling Model in Grid Computing Based on an Advanced Genetic Algorithm *

Hao Tian[1], Lijun Duan[2]
[1]School of Computer Science and Technology, Hubei University of Economics
Wuhan, Hubei 430205, China
[2] Department of Computer and Information Engineering, Hubei Institute of Economics Management
Wuhan, Hubei 430079, China
Email: haotian@mail.whut.edu.cn

ABSTRACT

Grid computing is one of the research hotspots in high performance computing area. Efficient scheduling of complex applications in a grid environment reveals several challenges due to its high heterogeneity, dynamic behavior, and space shared utilization. We studied the characteristics of gird tasks, analyzed several main grid resource management models, and built a common model of grid tasks, put forward a hierarchy grid resource management and scheduling model, expounded the ideas of designing the model and described the details of it. Moreover, we proposed an advanced genetic algorithm (GA) in its scheduling strategy, particularized the principle and the function of this algorithm as well as giving the concrete plan to put every step of the scheduling strategy into practice. Finally, the algorithm was simulated with the aid of SimGrid toolkit. The results indicate that the model could more efficiently realize global scheduling and managing. It could be an effective measurement for grid scheduling.

Keywords: Resource Management, Scheduling, Strategy, Model, GA

1. INTRODUCTION

Grid computing is a parallel computing environment including many heterogeneous and remote computing resources. Its object is to build a universal and mass computing process virtual system consisted by several distributed resources including computation hosts, network bandwidth and data centers [1]. Grid computing has broad prospect of application in many fields such as commerce, transportation, meteorology and education.

Resource management is one of the key technologies in grid computing. It couples grid resources logically as a single integrated resource for users. Users can communicate with the grid system directly without considering the complexity of grid resources and grid architecture. Generally speaking, grid resource management system has three kinds of basic services: resource distribution, resource detection and resource scheduling.

Scheduling is an important part of grid computing. The efficiency and acceptability of resource management mainly depend on its scheduling strategy. Scheduling program allocates needed resources to the corresponding requests, including cooperation allocation through different systems. However, resource scheduling is becoming a complicated problem because of the dynamic and heterogeneous characteristics of grid system as well as the different needs for the resources of the applications applied in grid system. Scheduling in gird computing is a non-deterministic polynomial complete problem.

Based on our previous work[2], we bring forward a hierarchy grid resource management scheduling model and improve the advanced genetic algorithm (GA) in its scheduling strategy.

2. MAIN GRID RESOURCE MANAGEMENT SYSTEMS AND SCHEDULING STRATEGIES

2.1 Main Resource Management Systems

Architecture of the model of resource management mainly depends on the number of the resources that needed to be managed and the tasks that needed to be scheduled, and it also rests with whether the resources are in a single area or in several areas. Several grid resource management systems have been proposed in the last few years. There are three main kinds of them: centralized management model, distributed management model and hierarchy management model. Hierarchy management model is a mixed model (combination of centralized and distributed models), it synthesizes the characters of the two models, it not only can obey the local scheduling policy of resource owner, but also can manage whole system with the best scheduling method, so it’s suitable for grid system. Agent technology is a hot topic in the distributed object-oriented systems. Agent can provide a useful abstraction on the grid environment. By their ability to adapt to the prevailing circumstance, agents will provide services that are very dynamic and robust, and it is suitable for a grid environment. Agents also can be used to extend existing computational infrastructures.

2.2 Main Scheduling Strategies

Grid scheduling strategy is the core of grid scheduling. There are so many special grid scheduling strategies such as the scheduling strategy Using Legacy Codes [3], the scheduling strategy based on LDS [4], the scheduling strategy Using Predicted Variance [5], Optimal Job Scheduling [6], PUNCH [7], XtremWeb [8] and so on. Their difference mainly lies in their scheduling algorithms. Most of them adopt a conventional strategy where a scheduling component decides which jobs are to be executed at which resource based on functions driven by system-centric parameters. GA is widely used in grid scheduling, and there have been many modifications for it. We have adopted the scheduling strategy based on an advanced GA in our previous work. In this paper, we also use the scheduling strategy in our model and we improve the algorithm.

3. GRID TASKS ANALYSIS

3.1 Characters and Model of Grid Tasks

* This work is Supported by the Educational Ministry of Hubei Province, China (Grand No.D200619006)

Different grid resource schedulers and different scheduling strategies have different views about grid task. Under the condition of taking complete time as optimization target, in the process of grid scheduling, we regard every user's request as a meta-task, and partition a meta-task into several independent tasks, viz. regard a meta-task as a union of independent tasks and all tasks can be scheduled, but we don't exclude the dependence between any two tasks. We regard a task as a combination of a data transfer subtask and a computing subtask, and we only consider the two subtasks using correlated resource to execute tasks. The model of meta-task is showed in Fig. 1.

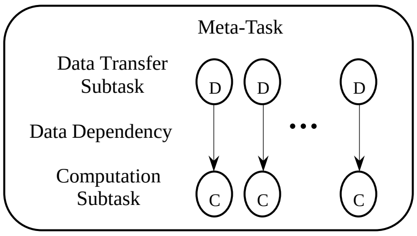


Fig.1. Model of a meta-task

In Fig.1, data transfer subtask means the communication needed in computation subtask, viz. the part that needs expending store time and communication time. Computation subtask is the part of expending computing time. In fact, it's reasonable to partition task in current grid system and computing module. In the process of grid scheduling, a group of computers or mainframe computers are usually regarded as a computation resource (agent), because they are autonomous systems based on SSI (Single System Image), they have sound scheduling measures. After an application (meta-task) is submitted to grid, scheduling program locates the data the meta-task needs and assigns every task to its most suitable agent, and the agent executes the task after it gets the data. Changing the scheduling policy of a computer group or a mainframe is impossible and unnecessary [9], so we do not care whether there are any other subtasks.

3.2 Grid Task and Grid Resource

From what we have analyzed above, it is clear that every task relates to a data center and a computation agent (a group of grid resources) after scheduling. Data center first transfers the data that computation subtask needs to computation agent, and computation agent begins to execute the computation subtask as soon as the data transfer has finished, so the complete time of a task can be regarded as the sum of the complete time of its data transfer subtask and the complete time of its computation subtask.

4. MODEL OF RESOURCE MANAGEMENT AND SCHEDULING BASED ON AN ADVANCED GA

Fig.2 shows the frame of our model, which includes four parts, viz. task partition module, scheduling decision module, information collection module and grid resource module.

4.1 Task Partition Module

Task partition module partitions a grid application (a meta-task) into several independent tasks. There are many useful methods to do this. In this paper, it is assumed that any meta-task can be partitioned into several independent tasks efficiently.

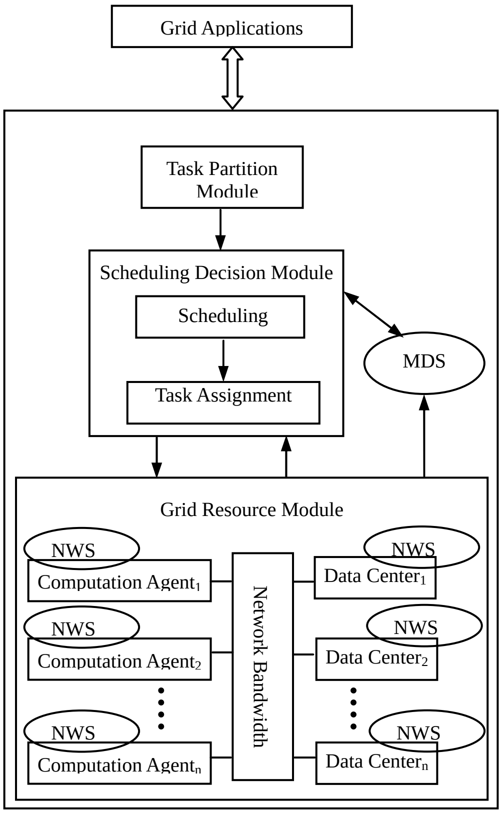


Fig.2. Model of grid resource management and scheduling

4.2 Scheduling decision Module

Scheduling decision module uses the scheduling strategy based on the advanced GA to distribute every task to a corresponding group of computation agent and data center. It is the key component of the model.

4.3 Information collection Module

Information collection module is composed of MDS [10] and NWS [10]. MDS is a grid information management system, it is used to collect and issue the state information of a system, we can gain much information from it: union of available resource agents, attribute of every agent, such as type of processor, speed of processor, amount of available processors etc. NWS is a distributed monitor system, it is specially designed to monitor resources in existence and network state, it can provide short-term network capability forecast, and it works on every agent so as to provide real time monitoring. We can get such data by using NWS: *availableCPU*, *currentCPU*, *bandwidthTcp*, *latencyTcp* and *connectTimeTcp* etc. Information collection module collects information about grid resources, and feedback it to the scheduling decision module, so as to provide evidence for scheduling strategy.

4.4 Gird resource Module

Gird resource module includes several heterogeneous computation agents, data centers and network bandwidth. They are the hardware layer of a grid, and we suppose that the module can provide all the resources that a meta-task needs.

5. SCHEDULING STRATEGY OF THE MODEL

We use the scheduling strategy based on an advanced GA. GA is a method of stochastic optimization and search, and it has the self-adapted search ability which has potential ability of learning. It expresses solution of problem as chromosome, before executing algorithm, it produces a group of chromosomes, viz. tentative solution, and puts the tentative solution into the environment of problem, then chooses the chromosomes can fit the environment from the group on the principle of survival of the fittest, produces a new generation group of chromosomes fit the environment better through crossover and mutation. After several generations anagenesis, there will be a new group of chromosomes fittest the environment, viz. the best solution of problem.

5.1 Description of the Problem

Based on previous discussion, it is assumed that a meta-task is *T*, and it can be partitioned into *l* independent tasks, grid computation system is composed of *n* heterogeneous computation agents *C* $(C=\{C_0\ ,\ C_1,\ldots\ C_{n-1}\})$ and *m* heterogeneous data centers *D* $(D=\{D_0\ ,\ D_1,\ldots\ D_{m-1}\})$, every computation agent can get data it needs from any data center, then there are $m \times n$ communication lines (network bandwidths) in whole grid system, viz. $m\times n$ groups of grid resource. We express these groups as a $m\times n$ matrix *R*; express the transfer time from a data center to a computation agent as a $m\times n$ matrix *DT*; express the computing time of a task on a computation agent as a $m\times l$ matrix *CT*; express the start time of a task on a computation agent as a $l\times n$ matrix *ST*, its value decided by the executing order of a task and the idle degree of the related computation agent.

This shows, the essential of grid resource scheduling is distributing *l* independent tasks to $m\times n$ groups of grid resource in order to minimize complete time of a meta-task and use grid resources sufficiently.

As the model shows, the values of the matrix *DT* and the matrix *ST* can be provided by NWS directly. The value of the matrix *CT* can be calculated by Eq. (1).

$$CT(T_i,R(j,k)) = \frac{load(C_j)}{speed(C_j)\times processors(C_j)+1} \qquad (1)$$
$$(0 \le i \le l-1, 0 \le j \le n-1, 0 \le k \le m-1)$$

where *speed*(C_j) is the average processor speed of agent *j*; *Processors*(C_j) is the amount of processors of agent *j*; *load*(C_j) is the task load assigned to agent *j*. The three indexes can be gotten by NWS and MDS.

Complete time of a task can be calculated by Eq. (2).

$$T_iT = DT(j,k) + ST(i,j) + CT(T_i,R(j,k)) \qquad (2)$$
$$(0 \le i \le l-1, 0 \le j \le n-1, 0 \le k \le m-1)$$

Complete time of a meta-task can be calculated by Eq. (3).

$$MTT = \max\{T_iT \mid 0 \le i \le l-1\} \qquad (3)$$

So the task of scheduling is to realize the best assignation of grid tasks on groups of grid resource to minimize *MTT*.

5.2 Initialization

The chromosomes coding technology in our strategy is subsection according to the amount of tasks, a section expresses a task, it forms from a task mark T_i and a resource group mark *R(j,k)*, it means that task T_i is executed by resource group *R(j,k)*. In order to create an initialization population, tasks are distributed to resource groups equally and stochastically.

5.3 Fitness Function

Choosing a proper fitness function can evaluate every iterative solution well. In this paper, the fitness function can be calculated by Eq. (4).

$$f = T_iT_c/(T_iT_m+1) \qquad (4)$$

where T_iT_c is the complete time of task T_i with current scheduling strategy, and T_iT_m is the complete time of task T_i with the best scheduling strategy.

5.4 Selection and Anagenesis

We calculate the fitness function of task T_i of every generation, keep it if it fits the demand of convergence, then choose other tasks into choice union and realign their orders in chromosome.

Anagenesis is used to generate next generation chromosome. In the chosen sections of chromosome, it first tries different combinations of tasks within changeable limits, viz. tries to distribute the tasks to random grid resource groups once again, then calculates their complete time, and chooses the best scheduling combination, consequently finishes a time of anagenesis, the new chromosome will be the origination of next anagenesis, in this way, the global optimum solution will be found.

6. EVALUATION

We used the SimGrid [11] toolkit to evaluate our scheduling algorithm. The platform used for simulation is an example of grid model included in the SimGrid package.

In the evaluation we used this platform to simulate applications with different number of tasks (1000 and 10000 tasks) and quantity of computation per task (100,500, 1000 and 2000 MFlop/s) using deployments with 64 and 90 nodes (groups of grid resource). In our experiments we assumed that communication costs to send one task to an agent is fixed (0.001 Mbyte/s) and to receive the result is irrelevant. We also assumed that the maximum start time of all tasks is 10 seconds.

Fig.3 illustrates the measurements obtained for an application containing 1000 tasks begin executed in 64 nodes of the platform, with computation amount per task varying from 100 to 2000 MFlop/s. Using 64 nodes and computation quantity ranging from 500 to approximately 1000 MFlop/s, the model presented the best results.

In Fig.4 the same measurements are presented for an application with 10000 tasks. Using tasks with no more than 400 MFlop/s the model presented the best results. It behaved better from 200 to approximately 500 MFlop/s than after 500 MFlop/s.

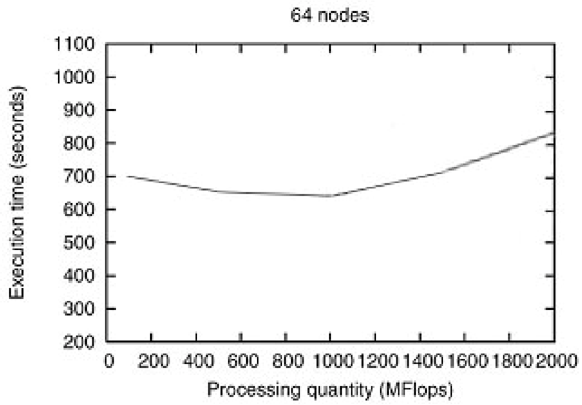


Fig.3. Measurements scheduling 1000 tasks using 64 nodes

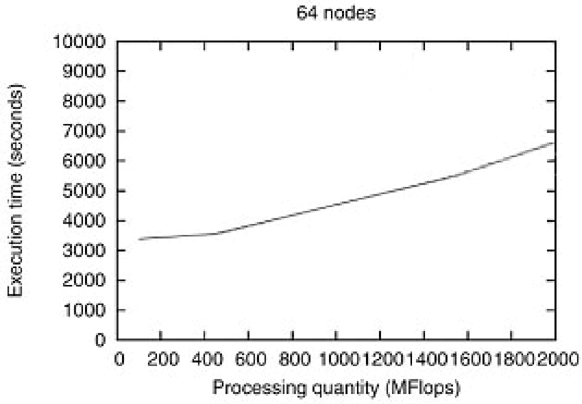


Fig.4. Measurements scheduling 10000 tasks using 64 nodes

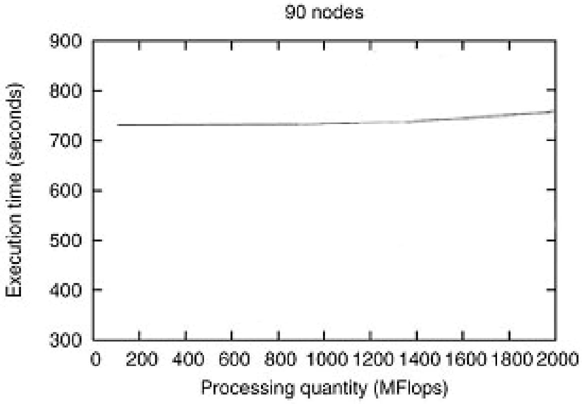


Fig.5. Measurements scheduling 1000 tasks using 90 nodes

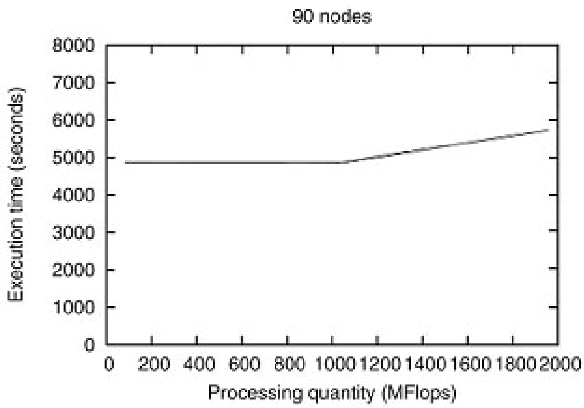


Fig.6. Measurements scheduling 10000 tasks using 90 nodes

Fig.5 illustrates the measurements obtained for an application containing 1000 tasks begin executed in 90 nodes of the platform, with computation amount per task varying from 100 to 2000 MFlop/s. In Fig.5 the line is smoother than others in fig.3 and fig.4. Using 90 nodes and computation quantity ranging from 100 to 2000 MFlop/s, the model behaved stable.

In Fig. 6 the measurements are presented for an application with 10000 tasks executed in 90 nodes. Using tasks with no more than 1100 MFlop/s the model presented the best results.

These figures show that the model behaves better in the grid environment with more tasks when there are same level quantities of nodes, and it also behaves more stable in the grid environment with more nodes when there are similarly many tasks, so it is clear that the model performs better with a higher number of tasks and groups of grid resource.

7. CONCLUSIONS

On the base of current research, we take the hierarchy management model as prototype, build a mended grid resource management and scheduling model, propose an advanced GA in its scheduling strategy, and simulate it by using SimGrid toolkit. The result shows that the model has good expandability, can realize globally optimum scheduling, it is an efficient plan of grid resource management scheduling. Its main limitation currently lies in the modeling of the grid capacities to treat the grid tasks. What we need to do is to improve the data modeling and the possibility to extrapolate the model for values of the workload that could be less regular. Nevertheless we believe this novel approach is promising and already a very good alternative to be considered when a scheduling algorithm is needed for scheduling applications in grid computing.

REFERENCES

[1] I. Foster, C. Kesselman, *The Grid2: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Press, 2003, pp.20-25.

[2] H. Tian and L. Duan, “Resource Management and Scheduling Model in Grid Computing Based on An Advanced Genetic Algorithm”, *Proceeding of the 5th International Conference on Distributed Computing and Applications for business, engineering and sciences* (DCABES'2006), Hangzhou, China, October 2006, Shanghai University Press, Vol.1, pp. 238-242.

[3] P. Kacsuk, A. Goyeneche et al, “High-Level Grid Application Environment to Use Legacy Codes as OGSA Grid Services”, in the *Proc. of the 5th International Workshop on Grid Computing* (GRID 2004), 8 November 2004, Pittsburgh, PA, USA, IEEE Computer Society, 2004, pp. 428-435.

[4] Fabrício A.B. da Silva, Sílvia Carvalho et al, “A Scheduling Algorithm for Running Bag-of-Tasks Data Mining Applications on the Grid”, in *Parallel and Distributed Processing and Applications: Second International Symposium*, ISPA 2004, Hong Kong, China, December 13-15, 2004, Lecture Notes in Computer Science, vol. 314, Sept. 2004, pp. 254-262.

[5] L. Yang, J.M. Schopf et al, “Conservative Scheduling: Using Predicted Variance to Improve Scheduling Decisions in Dynamic Environments”, in The

Proceedings of the ACM/IEEE Supercomputing 2003 Conference, November 2003, p.31-36.

[6] J. Soldatos, E. Vayias et al, “Grid Donors Resources Utilization Analysis towards Optimal Job Scheduling”, in the *Proc. of the DPSN '04, Workshop*, held in the Scope of IFIP Networking, 2004, Athens, May 14th 2004.

[7] R. Buyya, D. Abramson et al, “Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid”, *International Conference on High Performance Computing in Asia-pacific Region* (HPC Asia 2000), Beijing, China. IEEE Computer Society Press, USA, 2000.

[8] G. Fedak, C. Germain et al, “XtremWeb: A Generic Global Computing System”, Proceedings of *the 1st IEEE/ACM International Symposium on Cluster Computing and the Grid* (CCGrid 2001), May 15-18, 2001, Brisbane, Australia, IEEE CS Press, USA, 2001.

[9] L. Zha, Z. Xu et al, "Grid Task Scheduling Simulations Based on Simgrid", *Computer Engineering & Application*, vol. 14, pp. 90-92, 2003.

[10] Y. Du, Y. Chen et al, Grid Computing, Tsinghua University Press, 2002, pp.86-90.

[11] H. Casanova, “Simgrid: A toolkit for the simulation of application scheduling”, in the *Proceedings of the IEEE Symposium on Cluster Computing and the Grid* (CCGrid’01), 2001.

Hao Tian is an instructor of the School of Computer Science and Technology, Hubei University of Economics. He graduated from Wuhan University of Technology and received his Master degree in 2005. His research interests are in distributed parallel processing, grid computing and communication technology.

Lijun Duan is an instructor of the Department of Computer and Information Engineering, Hubei Institute of Economics Management. She graduated from Wuhan University of Technology in 2003. Her research interests are in computer applications, soft engineering and grid computing.

A Resource Scheduling Algorithm in the Grid Environment

Xiangchun Han, Tao Zhang
Department of Computer Science and Engineering，Yanshan University
Qinhuangdao, Hebei, P.R.China
Email: 2028087@163.com

ABSTRACT

Satisfying quality of service is one of the most aims of the grid resource scheduling. After researching many algorithms based QoS, we improve the Min-min algorithm and introduce a scheduling algorithm based two dimensions of QoS. The algorithm fully considers not only the tasks' demand towards resource's QoS, but also users' QoS restrictions to the tasks.It make resources allocated reasonable.

Keywords: Qos, Resource Scheduling, Grid Environment, Min-Min, Grid Resource

1. INTRODUCTION

Resource scheduling is a integral part of parallel and distributed computing. Extensive research has been conducted in this area leading to significant theoretical and practical results. However, with the emergence of the computational grid, new scheduling algorithms are in demand for addressing concerns originating from the Grid infrastructure.

With the extending of grid research, quality of service becoming a new factor in research of resource scheduling. In the grid environment, applications may have some requirements to computing resource's QoS. Thus it needs to filtrate resources for the algorithm in order to allocate tasks to the appropriate resources. However, many conventional scheduling algorithms such as Min-min, Max-min, sufferage and so on, don't consider the QoS[10], and the algorithm based on QoS only stay on one dimension QoS[2][3].Aim at solving problem, we introduce a kind of grid resource scheduling algorithm guided by two dimensions of QoS based on the Min-min algorithm.

2. ISSUE DESCRIBING

2.1 Scheduling Mode

Scheduling modes can be divided into two kinds, one is on-line mode and the other is batch mode. In the on-line mode the users' tasks are scheduled to the resource as soon as arrive. It does well in scheduling effect but real time. In the batch mode all the tasks are collected in a set called Meta-task, tasks are scheduled to the resources when scheduling period arrive or the tasks are accumulated up to some amount. Although the batch mode does worse than the on-line mode in real time, it could gain better scheduling effect because of collecting more information of the tasks and the resources. In our text we adopt the batch mode, and suppose all tasks are independent.

2.2 Definition of QoS

Grid is that multi-user share the resources, and tasks submitted by users are various. These tasks with different levels of QoS requests compete for resources. While a task with low QoS request occupy high QoS resource, a task that requests a high QoS service can only be executed on a resource providing high quality of service. In this circs, low QoS tasks occupy high QoS resources while high QoS tasks wait as low QoS resources remain idle, and the resource serious wasteful. To overcome the shortcoming, it is necessary to add QoS restricts into algorithm.

As we know the task user submits have a deadline, because the task needs some time to run. And that user's expectation to task's deadline is different. So as the users, the tasks' deadline represents the demanded quality of service. On the other hand the network bandwidth increase rapidly, but the transmission speed is limited, and different resource's bandwidth has difference, there is a notable difference between huge data transmitssion and data submission in the computer system. In the grid computing applications always need to transport lots of data. Because of these concerns, we choose the deadline and bandwidth as two kinds of QoS request. Here we define the quality of service as QoS. {bandwidth, deadline}.

2.3 Benefit Function

As users, they can get some benefits when tasks are completed after some time, so users could define their benefit functions. Here, we present images of four common benefit functions (Fig.1). We adopt (a) in the text. If the tasks are completed before the deadline users could gain the highest benefit, otherwise they gain none.

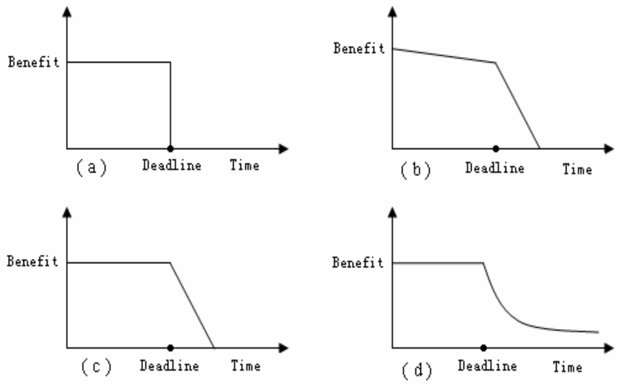


Fig.1. Images of benefit functions

2.4 Terminology

The expected execution time ET_{ij} of task t_i on machine m_j is defined as the amount of time taken by m_j to execute t_i given that m_j has no load when t_i is assigned. The expected completion time CT_{ij} of task t_i on machine m_j is defined as the wall-clock time at which m_j completes t_i (after having finished any previously assigned tasks). Let m be the total number of the machines in the HC suite. Let the arrival time of the task t_i be a_i, and the beginning time of t_i be b_i. From the above definitions, $CT_{ij} = b_i + ET_{ij}$. Let CT_i be CT_{ij}, where machine m_j is assigned to execute task t_i. The makespan for the complete schedule is then defined as max CT_i. Makespan is a measure of the throughput of the heterogeneous computing system. And if a task's CT is less than the deadline, we consider it can be completed.

3. ALGORITHM DESCRIBING

We know the grid environment is extremely complex, in order to reflect the dynamic and the autonomy of the grid environment we adopt a repeated mapping strategy. When the scheduling period comes scheduler collect the information of the current grid resource status. In the Meta-task there are not only the new tasks but also the unexecuted tasks that have mapped. In this way a task may map many times so that the system loads increase, but because of collecting more new information, it can gain better scheduling effect.

Conventional scheduling algorithms use the completion time as the main scheduling goal, but in the grid environment, user's action and system resource's status are very complex, it isn't reasonable to pursue the shortest completion time simply. For the grid user, the more tasks complete in the given time, the more satisfied they feel. So we define the user's satisfaction degree S. (the amount of tasks completed on time / the amount of submitted tasks) * 100% as one of the scheduling goals. In the text the algorithm we introduce guided by two dimensions QoS constraint, under the leading of the network bandwidth, increase the amount of the tasks which are completed in the deadline. Generally speaking, grid tasks can be divided into two kinds. One is network task, it needs to transport huge data between users and servers, although transmission times are different because of resources' different useable bandwidth, the transmission time of this kind of task takes up a good proportion in the total execution time; the other one is computing task, it has a huge amount of computing, but a small amount of data, and the transmission time of this kind of task just takes up a small proportion in the total execution time. We allocate the emergent tasks to the high capability resources in precedence, we call it Capability Resources First Scheduling Algorithm, and CRF for short. The pseudo code of the algorithm as follow:

```
T=$t_0$//start to schedule
(1) while (true)
(2)   t=t+⊿t
(3)   while (current time<t)
(4)     collect the arriving tasks into Mete-task
(5)   end while
(6)   collect the hosts' state informations and get CT and b
(7)   collect tasks which have been mapped but haven't been
      executed into Meta-task
(8)   for all the n resources
(9)       sort the sequence of the resource ordered by
          capability descended and represents it as
          capabilityOrder[1…n]
(10)      sort the sequence of the resource ordered by
          bandwidth descended and represents it as
          bandwidthOrder[1…n ]
(11)  end for
(12)  for all the m tasks in the Meta-task
(13)      sort the sequence of the tasks ordered by deadline
          ascended and represents it as
          deadlineOrder[1…m]
(14)  end for
(15)  for each t ∈ the Meta-task
(16)      if  ( t is a computing task )
(17)          if （CT <= deadline）
(18)              Submit[i] = t //the list of the tasks submitted
(19)              Resource[i] = capabilityOrder[i]//the list of
                                          resources allocated
(20)          else the task can't be completed on time
(21)          end if
(22)      end if
(23)      if  ( t is a network task )
(24)          if （CT <= deadline ）
(25)              Submit[i]=t
(26)              Resource[i]= bandwidthOrder [i]
(27)          else the task can't be completed on time
(28)          end if
(29)      end if
(30)      delete t from Meta-task
(31)  end for
(32)  assign tasks in the Submit[i] to resources in the
      Resource[i]
(33)  updates hosts' state informations
(34) end while
```

First collect tasks of new arriving or having been mapped but unexecuted into Meta-task, and collect hosts' state informations, get beginning time and completion time(1)-(7).Then sort resource ordered by capability and bandwidth respectively(8)-(11), and sort tasks according the deadlines(12)-(14). After complete these, we fetch a task from Meat-task, justify its task type (computation task or network task)(16)(23) and whether it coulde complete on time(17)-(24). If it can be completed before its deadline, add the task to the set Submit[i](18)(25) and add the related resource to the set Resource[i](19)(26). At last delete the task which is assigned to its resource from the Meat-task(30), assign tasks in the Submit[i] to the resources in the Resource[i](32) and update hosts' state informations(33).

4. SIMULATION

In order to validate the validity of the algorithm, we simulate the algorithm with simulating software called Gridsim. Gridsim provides lots of API, it could generate different the parameters of the capability of hosts, network bandwidth and so on randomly. We adopt a emulate to compare the CRF with the Min-min in two aspects of percentage of completed tasks and makespan. We introduce four sets of tasks differ in amount, simulate times, and take the average. The amount of computing tasks and network tasks generates randomly and the execution time of these tasks are known. Fig.2 shows the difference in the percentage of completed tasks of the two algorithms, and Fig.3 shows the difference in makespan.

From Fig.2 and Fig.3 we can see the makespan and the percentage of completed tasks get different degree of improving with submitting more tasks. Although the effect of improving makespan isn't as good as improving percentage of completed tasks, it leads to complete more tasks and fulfill users' requests much more.

5. CONCLUTIONS

Since the current scheduling algorithms have no QoS constraints or only one dimension QoS constraints, we introduce a algorithm that is guided by two dimensions of QoS based on improving the algorithm of Min-min, considering the factors of bandwidth and deadline. It fulfills the demands of users and system preferably. We adopt the repeated mapping strategy and make the algorithm be more suitable to the dynamic and the autonomy, and get satisfying simulation result. In this text we just consider two dimensions of QoS, and our work is adding more QoS constraints into the algorithm in the future, and makes the scheduling algorithm more suitable to the complex environment of the grid, and gets a better scheduling effect.

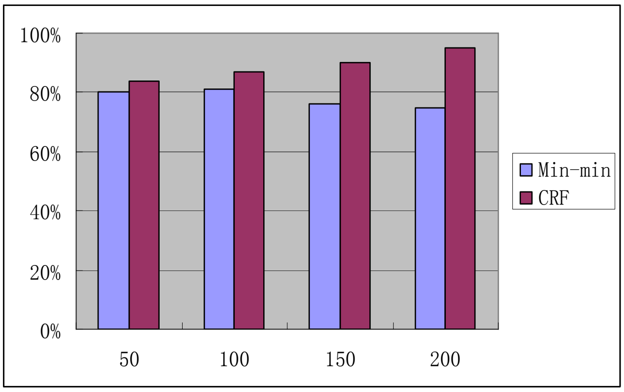


Fig.2. The difference of two algorithms in the percentage of completed tasks

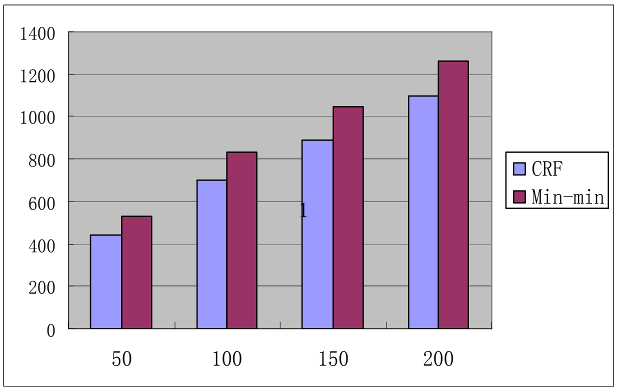


Fig.3. The difference of two algorithms in makespan

REFERENCES

[1] Jia Yu, Buyya, R., Chen Khong Tham, “Cost-Based Scheduling of Scientific Workflow Application on Utility Grids,” *e-Science and Grid Computing*, 2005, First International Conference on 05-08 Dec. 2005 Page(s):140 – 147.

[2] HE Xiaoshan, Xian-He Sun, Gregor von Laszewski, “QoS guided min-min heuristic for grid task scheduling,” *Journal of Computer Science and Technology*, 18, 2003, pp.442-451.

[3] Zhang Jinquan Ni Lina Jiang Changjun, “A Heuristic Scheduling Strategy for Independent Tasks on Grid,” in High-Performance Computing in Asia-Pacific Region, 2005 Proceedings, *Eighth International Conference* on Publication Date: 30 Nov.-3 Dec. 2005 On page(s): 6 pp.

[4] Buyya R, Murshed M, Abramson D, “A Deadline and Budget Constrained Cost-Time Optimization Algorithm for Scheduling Task Farming Applications on Global Grids [A],” in *Proceedings of the 2002 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA702)* [C] 2002.

[5] R. Buyya and M. Murshed, “GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing,” *Concurrency and Computation: Practice and Experience (CCPE)*, pp. 1-32, May 2002.

[6] Gong L., Sun X.H., and Waston E, “Performance modeling and prediction of non-dedicated network computing,” *IEEE Trans. on Computer*, September, 2002, 51(9): 1041-1055.

[7] Min-You Wu, Wei Shu, Zhang,H., “Segmented min-min:a scheduling algorithms for independent coarse-grained tasks,” *Applications and the Internet Workshops*, 2004. SAINT 2004 Workshops. 2004 International Symposium on 26-30 Jan. 2004 Page(s):674 – 680.

[8] Yu, J. and R. Buyya, “A Taxonomy of Scientific Workflow Systems for Grid Computing,” *ACM SIGMOD Record*, 2005. 34(3): p. 44-49.

[9] Noriyuki Fujimoto, Kenichi Hagihara, “A Comparison among Grid Scheduling Algorithms for Independent Coarse-Grained Tasks,” *SAINT Workshops* 2004: 674-680.

[10] Fujimoto, N., Hagihara, K., “A comparison among grid scheduling algorithms for independent coarse-grained tasks,” *Applications and the Internet Workshops*, 2004. SAINT 2004 Workshops. 2004 International Symposium on 26-30 Jan. 2004 Page(s):674 – 680.

[11] gan A,Ozguner F., “Schdeuling Independent Tasks with QoS Requirements in Grid computing with Time-Varying Resource Prices [A],” *Grid Computing-GRID* 2002[C] 2002.58-69.

Xiangchun Han is a associate Professor. His main research interests include computer architecture, parallel computation and CG.

Tao Zhang is a post graduated; his main research interests include gird computing and parallel computing.

A Study on Grid-based Information Organization and Management *

Jieli Sun[1,2], Yanfeng Bai [1], Qingyun Sun[3]
[1] College of Information Technology, Hebei University of Economics and Business
Shijiazhuang, 050061, China
[2] Graduate University of Chinese Academy of Sciences
Beijing, 100049, China
[3] College of Mechanical and Electronic Engineering, Nanjing Forestry University
Nanjing, 210037, China
Email: sunjieli@126.com , sunqingyun@126.com

ABSTRACT

An application of grid technologies in information organization and management is proposed. Based on traditional information organization and management mechanism and the technologies of grid, this paper introduces OGSA technologies and analyzes its theory and methods of information organization and management. It focuses on discussing the model framework of OGSA-DAI. Then the function and implementation methods of OGSA-DQP and its effect are proposed. At last it provides a new way to develop grid-based information organization and management. The relationship between OGSA-DAI and OGSA-DQP are also discussed.

Keywords: Grid, OGSA, Information Organization and Management, OGSA-DAI, OGSA-DQP

1. INTRODUCTION

In 2002, expert on Grid，Forster，described Grid computation as a process in which the dynamic resources sharing and collaborative problem solving are realized within several virtual community organizations[1]. Grid provides all sorts of controlling strategies and methods for this cooperation, and at different request, it can dynamically establish various levels of relation with different organizations or individuals. With the wide spread of the Gird projects and development of Gird technology, Grid can be defined as a system which can coordinate the distributed resources by standard, open general protocol and interface to provide the best service.

In this paper, a application of grid technologies in information organization and management is proposed. Based on traditional information organization and management mechanism and the technologies of grid, this paper introduces grid technologies and analyzes its theory and methods of information organization and management.

The rest of the paper is organized as follows. The theory of open grid service architecture is given in section 2 as a preliminary work. Information organization and management in Grid environment is given in Section 3. The model framework of OGSA-DAI, the function and implementation methods of OGSA-DQP and the relationship between OGSA-DAI and OGSA-DQP are also given in Section 3. Section 4 concludes the paper and points out the directions of future work.

2. GRID AND OPEN GRID SERVICE ARCHITECTURE

The purpose for applying the Gird technology is to visualize the resources and services of virtual construction including site distribution, heterogeneous (which is supported by different software and hardware and different web corresponding technology) and dynamic changes, and then to integrate and manage it. Once necessary, those computers, services, data and other resources can be visited. The key for this purpose is the standardization so that those resources and services distributed in site or belonging to different constructions can be managed as an individual virtual system to find, visit, adjust and supervise.

Grid resources includes such management activities as resources detection, resources category, fault isolation, resources supply, resource supervision, and varied self ability and service level. The most important one is to select appropriate resources from the Grid Resources Pool at the request service level, and then provide client effectively to meet his demand [1, 2].

By different data module of grid object, Grid-based information services module can be divided into four types, which are different in its relation with grid entity as followings.

(1) Grid Information Service based on layer model, the representative finding is the MDS (Metacomputing Directory Services/Monitoring and Discovering Services) designed and explored by Globus project group.
(2) Grid information service based on object model, the description on Grid object by GGF[3] (Global Grid Forum) is the defined language for Grid information service data, which remains further study and exploration.
(3) Grid Information Service based on the relation model, the disadvantage of relation model applying to Grid environment is its lack of capability in the distributed management, and no major software package on Grid Information Service based on relation model exists at present
(4) Grid Information Service based on Open Grid Services Architecture (OGSA), it inherits the advantages of Grid information Service based on layer model and relation model and is the ideal solution to Grid Information Service.

OGSA is a set of norms and standardization based on the integration of the Globus Grid Computing toolkit and Web service technology, and OGSA is composed of two essential technologies---Grid and Web Service.

On the previous Web Service concept, OGSA proposed the concept of Grid Service to solve problems like service detection, service establishment and life long management.

Grid Service is a kind of web service, providing a set of interfaces which are clearly defined and following specific routine to solve problems like service detection, dynamic

* This work is supported by the project sustentation fund of Hebei Province (No. 042135121).

service establishment, lifetime management and notification. In OGSA, everything is considered as Grid service, so actually Grid is an expandable Grid service integration. Grid Service may be collected in different ways to meet the request of virtual organization which also can be partly defined by its operation and sharing service.

Because all parts in Grid environment are virtualized, provided with a set of relatively united essential interface, all the Grid service can be implemented by these interfaces, therefore, the layered structure and more advanced services will be formed. These services can also transfer different abstract layers and be treated as a uniform approach.

Virtualization makes it possible to reflect a variety of logical resources instance to the same physical resources, regardless of the specific realization while integrating the services. It may also perform resource management in virtual organization on the basis of the bottom layer resource. By virtualization of Grid service, general service semantic and behavior can be seamlessly reflected to the basic utilities of local platform[4, 5].

The significance of OGSA lies in that it expands Grid from academic field majoring in science and engine computing to a more universal social economic area featuring as Distributed Systematic Service Integration.

3. INFORMATION ORGANIZATION AND MANAGEMENT IN GRID ENVIRONMENT

3.1 Outline of Information Organization

Information organization is the process in which the disorderly specific information is made into the orderly state by certain principles and approaches, aiming to turn the disorderly information into orderly ones for the convenience in using and transferring information. Information organization is a ordering process , whose purpose is to help people make use of information conveniently[6, 7].

Grid technology focuses on enormous resource sharing, emphasizes the resource sharing and collaboration among constructions, and reveals its unique advantages in solving compatible of inhomogeneous platforms and system integration. The purpose for applying the Gird technology is to visualize the resources and services of virtual construction, and then to integrate and manage it. Once necessary, those computers, services, data and other resources can be visited. With the development at full speed of grid technology, it is necessary that the technology and theory of grid can be applied in information organization.

3.2 Information Organization and Management in Grid Environment

Because of the main purpose of Grid to support the collaborative working in sharing resources, the database integration in Grid environment becomes the popular study of the present data management in Grid environment. In Grid environment, data resource is heterogeneous, and data can be stored in different data carriers, such as all kinds of relation database, various file system. For the realization of resources sharing and collaborative usage on those heterogeneous data resources, Grid middleware must first make these heterogeneous data resources integrated, that is to say, the united visit technology of the heterogeneous data resources is the first step to realize data resources sharing and dispatch. By the united visit technology, storage retrieval of the request of higher layer client will be reflected as lower operation of data resources, realizing the effective united visit and management of data in extended field[8].

The existing three Grid database integration strategies are:

1) Virtual Database. Virtual Database contains only one database model. And all clients can not sense the existence of various independent database. Virtual database is welcomed in concept but difficult to realize. Several kinds of transparence should be considered while constructing virtual database: heterogeneous transparence, name transparence, owner and cost transparence, parallel transparence and distribution transparence.
2) Customize Integration. It refers that the application programs themselves will finish the integration of database. For instance, in some scientific application programs, the programmer personally finds the related data source, then divides the integration task into query, future executive program, establishment of middle data source, displayed data delivery and shift and findings storage. Grid database management system should provide support for this kind of integration to cut down cost, consumption on time and the occurrence of errors.
3) Increment Integration. Virtual Database is an ideal and tailored integration paying too much attention to details, but Increment Integration is between them. In Increment Integration, the explorer needs not to finish every detail, for the advanced data access and integrated set (like OGSA-DAI) will automatically finish the rest of the integration.

The main purpose of GGF is to establish a technology standards of Grid. It is composed of seven fields, and each field contains several working and studying groups. DAIS Working Group[9]（Database Access and Integration Services Working Group）belongs to GGF database field, purposing to study how to apply database to Grid. OGSA only supplies lower layer integration for data visit and integration, but Grid application requires more advanced level data visit and integration approach. DAIS is working on the standards of the Grid data base service. The reference implementation of the draft made by DDAIS on the standards of Grid database service is OGSA-DAI[10] (Open Grid Services Architecture-Data Access and Integration) Project.

The OGSA-DAI Project explored by British e-science center is the representation in database Grid study field. OGSA-DAI is the middleware platform of database visit and integration, purposing to construct a middleware which will help data access and integration in Grid environment, and it realized the service visit to various database which make them share data with outside in form of Grid service.

OGSA-DAI is middleware explored on Globus Toolkit3(GT3) to visit data in database. GT3 is a referred realization of OGSA, and it has mainly solved the problems like the security in Grid, information base utilities, resources management, corresponding and error detection, providing a good processing environment for Grid application program. OGSA-DAI project attempts to realize uniform access to different database system by Grid Services. OGSA-DAI packs complicated data operation and provides a uniform Grid service interface, making Grid client or service conveniently visit and integrate all kinds distributed heterogeneous data sources in Grid by service interface in Grid environment.

OGSA-DAI obtains data via services. Client interacts indirectly with date source by those services, and it may provide three services compatible with Web Service and Grid Service. Three services include WS-I(Web Service Inter-operability)， WS-RF(Web Service Resource Framework) and OGSI(Open Grid Services Infrastructure).

The framework of OGSA-DAI includes four parts as following:

1) Data Layer. OGSA-DAI may access to all sorts data sources, including relational database(as MySQL，SQL Server， Oracle) XMI database(as Xindice) and document(as Xindice).
2) Business Logic Layer. This layer packs essential functions of OGSA-DAI, receiving and responding to request of client server (like inquiry, revise, and etc.), managing data delivery, announcements, the connection, management and interaction with data sources.
3) Presentation Layer. This layer packs all services shown by OGSA-DAI to Grid, including DAISGR(Data Access Interface Service Group Register)，GDSF(Grid Data Service Factory) and GDS(Grid Data Service), all of which are performed by Web Service or Grid Service access, adapting WSDL(Web Service Description Language)and XML Schema to depict the corresponding access.
4) Client End. OGSA-DAI support the client access compatible to OGSI，WS-RF，WS-I, which mainly depends on the service provided by the Presentation layer of server end and may realize the operation on the retrieval renew to low data sources.

OGSA-DAI framework also offers access to Data and Business Logic Layer, adopting JDBC and XMLDB driver; access to Presentation and Business Logic Layer is responsible for the corresponding between the two layers, supporting the adaptation of core function of OGSA-DAI.

The present OGSA-DAI projects in use are AstroGrid, Biogrid, BioSimGrid, Bridges, FirstDIG, GeneGrid, ODD-Genes, OGSA-WebDB, etc. With huge fund support and technology accumulation, together with the constant increasing version, OGSA-DAI will acquire much greater functionality, including better quality, supporting more DBMS and more SQL access, DBMS management operation, document visit and class libraries of client end, etc.

3.3 Grid-Based Information Query

If Grid-based information system has established database on several Grid nodes, the application should perform distributive query. The distributive query management in Grid environment differs from that in other environment mainly in the following:

1) Grid not only provides a set of systematic approaches to make secure use of the distant data resources, it also makes the distant computing resources available in security.
2) Grid provides mechanism for detection, distribution and supervision of dynamic resources
3) Grid provides mechanism monitoring network state, which is very important for the query in Wide Area Network environment.
4) Grid follows the present standards and all sorts of database on Grid can access in uniform approach.

The problems OGSA-DAI mainly solved are the dynamic visit to various heterogeneous data sources in Grid environment and the delivery of data among each Grid service. In the application of Digital Library, a more advanced date integration approach is required. The present distributive query projects on Grid technology are polar*[11] and OGSA-DQP[12, 13] (Open Grid Services Architecture － Distributed Query Processing), etc.

Polar* is an distributed query system applied to Grid environment explored by researchers in British Manchester and Newcastle universities, and it studies the problems concerning distributed query management in Grid Environment and performs advanced distributed query according to Grid features.

The researchers of Polar* believe that the biggest feature of the distributed query in Grid environments is its self-adaptability and it must be able to make full use of the powerful parallel computational capability of Grid and adapt to the dynamic evolutive features of Grid resources.

Polar* designed a descriptive query language, realized the parallel query functionality to heterogeneous database, it also studied on inquiring how to dynamically respond to the changes of resources.

Polar project has done and is going to fulfill the following works:

1) Design a declarative query language for clients to express data access and data analysis;
2) Observe workable Grid resources, client request and cost restriction, and research on the best optimization method for queries of this query language and cost model;
3) Study parallel executive approach of query plan produced by Query Optimizer；
4) Study how to make the long-time running task use those extra usable resources emerging in running to improve its quality;
5) Research on how the query responds dynamically to the resources changes including those newly emerged useable resources, return of distributed resources, resource invalidation and the changes on load on sharing resources, and so on.

Another project researching on the distributed query management is OGSA-DQP building upon OGSA-DAI project. OGSA-DQP is a Grid service-based distributed query system for clients explored on the basis of OGSA-DAI. With parallel database technology, it realized the distributed query with parallel functionality at the complicated data intensive request.

OGSA-DQP conforming to the Grid standards based on OGSA is established on OGSA-DAI ,data access and integration on GGF, with the lower set adopting optimization parallel machine of Polar*. OGSA-DQP made use of the reference implementation of OGSA/OGSI -- Globus Toolkit 3, which formed a serviced based framework by virtualizing data into Grid Services.

OGSA-DQP is virtually a huge throughput distributed data flow engine, relying on the access to service-oriented in Grid environment and assuming data resource can be visited over service based interface. In addition to the query to the Grid data,

OGSA-DQP also can transfer the Web Service in Grid, therefore, it makes combination of data visit and analysis and produces more useful value.

OGSA-DQP has designed a framework: supporting the pubic query through Grid Database Services(GDSs) and any other Web Services applicable on Grid , then analyzing integrated data visit; supplying the potential parallel mechanism for complicated intensive data request referring to parallel database technology; realizing the automation to complicated, heavy and exclusive installment and resources utilization by way of query optimizer; with the set standards of GDSs, providing uniform visit to database metadata and interactions with Grid database; dynamically acquiring the necessary resources by OGSA tool to meet the purpose of effective evaluation on Distributed Query.

OGSA-DQP offered two kinds of services to fulfill its function: Grid Distributed Query Service(GDQS) and Grid Query Evaluation Service(GQES). With the increasing demand for such middleware in practical use, OGSA-DQP will develop to be more powerful and more convenient.

4. CONCLUSIONS

The information environment represented by e-science, e-learning, e-business and e-government is bringing in new users demands, users behavior and application mechanism of user information. Orienting to the users, the present the information environment accumulates information resources, information services, information applying activities with the ultimate purpose of supporting users to use information, refine knowledge, solve problem , and making it become the combination of work and study environment for the users.

At present, Grid service technology has not been well-developed but it keeps developing, and more new technologies and standards are adopted continuously, such as Data Flow Language. However, Grid Service technology can solve problems like heterogeneity, interoperability and distributed query of data reproduction among different systems, which is the direction of the study. Though the related projects have been explored abroad, the detailed standard is still on establishment and the practicality of those projects remains to be strengthened, so if exploring our related research as soon as possible, we might harvest valuable findings.

REFERENCES

[1] Ian Foster, Carl Kesselman. The Grid: Blueprint for a New Computing Infrastructure, second edition. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2003.

[2] Francine Berman, Geoffrey C. Fox, Tony Hey. Grid Computing: Making the Global Infrastructure a Reality. Wiley, New York, 2003.

[3] GGF. http://www.ggf.org

[4] Globus. http://www.globus.org.

[5] Joshy Joseph, Craig Fellenstein. Grid Computing

[6] Zhou Ning. Information Organization. Wuhan University Publishing Company, Wuhan, China, 2006.

[7] Weimin Dai. Information Organization. Higher Education Publisher: Beijing, China, 2004.

[8] Ingo Frommholz,Predrag, et a1. Supporting Information Access in Next Generation Digital Library Achitectrures. Pre-proceedings of the Sixth Thematic Workshop of the EU Network of Excellence DELOS. 2004.

[9] DAIS. http://www.gridforum.org/6_DATA/dais.htm

[10] OGSA-DAI.http://www.ogsadai.org.uk/

[11] Polar*.http://www.ncl.ac.uk/polar/

[12] M Nedim Alpdemir, Arijit Mukherjee, Norman W Paton et a1. OGSA-DQP: A Service for Distributed Querying on the Grid. In: EDBT2004. 2004.

[13] M. Nedim Alpdemir, Arijit Mukherjee, Anastasios Gounaris,Alvaro A. A. Fernandes, Norman W. Paton, Paul Watson. An Experience Report on Designing and Building OGSA-DQP: A Service Based Distributed Query Processor for the Grid. http://www.ogsadai.org.uk/documentation/publications/alpdemir.pdf.

Jieli Sun was born in Heilongjiang, China, in 1969. She is a Associate Professor of College of Information Technology, Hebei University of Economics and Business, Shijiazhuang, China. She graduated from Northeast Normal University in 1992 with specialty of Computer Science and Technology. She received the M.Eng. degree from Beihang University of School of Computer Science and Engineering in 2004. She is a doctoral student in Graduate School of Chinese Academy of Sciences. She has published two books, over 20 Journal papers. Her research interests are in grid computing, semantic web，ontology and text mining.

Yanfeng Bai is an associate Professor of Information Technology College, Hebei University of Economics And Business, Shijiazhuang, China. He graduated from the University of Electronic Science and Technology of China, graduated from Ordnance Engineering College of P.L.A. He has published two books, over 30 Journal papers. His research interests include text mining, grid technology, operation system and digital library.

A New Scheduling Strategy in Grid Computing *

Hao Tian[1], Lijun Duan[2]
[1]School of Computer Science and Technology, Hubei University of Economics
Wuhan, Hubei 430205, China
[2]Department of Computer and Information Engineering
Hubei Institute of Economics Management
Wuhan, Hubei 430079, China
[1]Email: haotian@mail.whut.edu.cn
[2]Email: duan_lj@sina.com

ABSTRACT

Grid system consists of a wide variety of geographically distributed resources. Efficient scheduling of complex applications in a grid environment reveals several challenges due to its high heterogeneity, dynamic behavior, and space shared utilization. In this paper, first we compared some typical scheduling strategies and pointed out their shortcomings, and then we proposed a new scheduling strategy based on an agent-oriented algorithm, finally we simulated the strategy with the aid of SimGrid toolkit and it was proved reasonable and efficient. It is an effective approach for tasks scheduling in gird computing.

Keywords: Grid Computing, Task, Scheduling, Strategy, Agent-Oriented.

1. INTRODUCTION

Grid computing is an important network technology that has growing up globally during the recent years. Its object is to build a universal and mass computing process virtual system consisted by several distributed resources. Scheduling strategy is one of the key technologies of grid computing, a powerful scheduling strategy can improve highly the whole grid system's performance.

Grid applications can be divided into several parallel independent tasks by some famous methods, therefore, a scheduling strategy's goal is to schedule these tasks efficiently and improve the load balance in distributed heterogeneous environment. There are many typical scheduling strategies in grid computing, but most of them focus on the adaptation of the workload during the execution, and their solutions are all based on some evaluation by the scheduler of the agents' capacities and of the tasks workload[1]. This means the scheduler of grid has to maintain the information during its scheduling process, and could not refresh the information in time.

In this paper, we compare some typical scheduling strategies and propose a new scheduling strategy of which main originally is to be agent-oriented. It can fit well the applications with more agents and more tasks.

2. MAIN SCHEDULING STRATEGIES

There are many scheduling strategies in grid computing and their difference lies in their scheduling algorithms.

2.1 Scheduling Strategies Based on Genetic Algorithm

The scheduling strategies based on GA are widely used in grid computing, we also adopted a scheduling strategy based on an advanced algorithm in our previous projects[2]. Its scheduling algorithm is usually a genetic algorithm or a hybrid genetic algorithm. They express solution of scheduling problem as chromosome. They produce a group of tentative solutions before executing tasks, and put the tentative solutions into the real grid environment, then choose the chromosomes that can fit the environment well from the group on the principle of survival of the fittest, produce a new generation group of chromosomes that fit the environment better through crossover and mutation. Anagenesis generation to generation, there will be a new group of chromosomes fittest for the gird environment, it means that the final group is the best solution of the problem. In fact, these scheduling strategies' originally is a method of stochastic optimization and search, which has the self-adapted search ability and the potential ability of learning.

This class of scheduling strategies based on GA is a useful way for grid computing, but it would spend more time if there are more agents and more tasks in gird.

2.2 Scheduling Strategy Based on Trust Mechanism

The scheduling strategies based on trust mechanism collect and feed back resource information through some grid service tools such as MDS and NWS[3]. They import trust mechanism by using interpersonal trust relationship in human society for reference, and adopt the dynamic self-adaptive distributed replica location method for the management of data repository and the method of dynamic transfer for the tasks in the trouble agents. Their scheduling algorithms are usually the Min-min algorithm or the Max-min algorithm[4].

These strategies could improve some performances such as costs and tasks finishing time, and they also have a successful rate of tasks execution. They work better in local scheduling than in large-scaled scheduling as building trust mechanism in a large scale is difficult.

2.3 Scheduling Strategies Based on Self-scheduling Algorithm

These scheduling strategies' scheduling algorithm are self-scheduled, they divide equally the tasks in several chunks on a specific distribution, then they usually build a decreasing-size chunking scheme and distribute the chunks to agents with a fixed ratio[5]. An agent obtains a new chunk whenever it becomes idle.

This class of scheduling strategies is well suited for dynamic and heterogeneous environments such as grids, but due to the scheduling overhead and communication latency incurred in each scheduling operation, the overall finishing time may be greater than optimal.

* This work is supported by the Educational Ministry of Hubei Province, China (Grand No.D200619006)

2.4 Other Scheduling Strategies

There are so many other classical scheduling strategies in grid such as the scheduling strategy Using Legacy Codes[6], the scheduling strategy Using Predicted Variance [7], Optimal Job Scheduling [8], PUNCH[9] and so on. Most of them adopt a conventional strategy where a scheduling component decides which tasks are to be executed at which agent based on functions driven by system-centric parameters.

In all scheduling strategies shown above, the amount of tasks sent to each agent is decided by the scheduling component. Once the scheduling process begins, all changes about agents could not feedback to the scheduling component in time. The strategies could not realize real-time scheduling.

3. SCHEDULING STRATEGY BASED ON AN AGENT-ORINTED ALGORITHM

In this section we propose another scheduling strategy, where the amount of the tasks to be assigned to each agent is decided by the agent itself. The strategy is divided the following phases: initialization, evaluation, adaptive and finalization phase.

3.1 Initialization Phase

The initialization phase's goal is to divide every grid application into several independent tasks and get the information about all available agents.

Grid applications can be divided into independent tasks by many classic methods. Under the condition of taking complete time as the target of system optimization, in the process of grid scheduling, we divide a gird application into several independent tasks, and their relationship as shown in Fig. 1. Scheduling component can find out how many available agents there are in current grid and the performances about these agents by MDS and NWS.

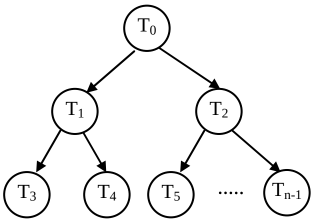


Fig.1. Tasks relationship

3.2 Evaluation Phase

The initialization phase's goal is to divide every grid application into several independent tasks and get the information about all available agents.

The evaluation phase's goal is to setup and refine the performance model of each agent. The scheduling component sends tasks to agents using a fixed quadratic increment, at the same time the scheduling component send a execution time limit *StdExecTime* to each agent. Each agent computes the task received and includes this information in its performance model, calculates the best workload size to be computed at the next iteration based on the analysis of its performance model. The *StdExecTime* was calculated using Eq. (1).

$$StdExecTime = \frac{\sum_{i=0}^{m-1} task_i}{\sum_{j=0}^{n-1} processor(agent_j) \times speed(agent_j) + 1} \qquad (1)$$

where:

speed(*agent$_j$*) is the average processing speed of *agent$_j$*, *Processor (agent$_j$)* is the amount of processors of *agent$_j$*, *m* is the number of tasks to be scheduled, *n* is the number of available agents. They can be gotten by NWS and MDS.

This process continues until the scheduling component receives a signal from the agent in order to start the adaptive phase. This signal is generated when the agent performance model starts presenting estimates with minimum error.

3.3 Adaptive Phase

The phase's goal is to adapt each agent's performance model of any variation is observed and to generate appropriate estimates of workload size to be computed in the next iteration.

The scheduling decision model sends to every available agent a task using the same fixed increment, but it includes information about the time slice the agent has to compute at the next iteration. After the processing of the task by the agent, the scheduling component receives the execution time of the task computed and an estimation of the next workload size in order to accomplish to the time slice define at its side. At the next workload to assignment for the agent, the scheduling component just change the workload size to send to the agent accordingly to the estimate previously received. It keeps using this procedure until it reaches a specified limit *availableTasks*, the *availableTasks* was calculated using Eq. (2).

$$availableTasks = \sum_{i=0}^{r} taskSize_i \qquad (2)$$

$$(0 < r \le n-1)$$

where:

r is the number of remain tasks.

In this phase each agent starts a loop receiving tasks to be computed. Together with the task, it receives the workload size and execution time inserted in a prediction table, which is used to compute the performance model. Using this prediction table and the execution time limit value received from the scheduling component, it computes the next workload size *nextTaskSize*. After that, the agent sends to the scheduling component the result, execution time *execTime* of the task received and an estimation of the next workload size *nextTaskSize*. The agent gets out from the loop when it receives a signal message from the scheduling component to initiate the next phase. The *execTime* was calculated using Eq. (3), and the *nextTaskSize* was calculated using Eq. (4).

$$execTime_i = \frac{taskSize_i}{processor(agent_j) \times speed(agent_j) + 1} \qquad (3)$$

$$(0 \le i \le m-1, 0 \le j \le n-1)$$

where:

execTime$_i$ is the execution time of task *i* that executed by agent *j*.

$$nextTaskSize_j = taskSize_i \times \frac{StdExecTime}{execTime_i} \qquad (4)$$

$$(0 \le i \le m-1, 0 \le j \le n-1)$$

where:

nextTaskSize$_j$ is the estimation of the next workload size on agent *j* after it finished executing task *i*.

3.4 Finalization Phase

The finalization phase adjusts the workload size computed in each agent in order to achieve load balancing, resulting in a better performance. When the scheduling component switched from the adaptive phase to this phase, it stops using agents' predictions and starts using current scheduling method till the end of tasks processing. After assigning the remaining tasks, the scheduling component starts a loop receiving the remaining results from the agents.

4. SIMULATION

We used the SimGrid[10] toolkit to evaluate our scheduling strategy. The platform used for simulation is an example of grid model included in the SimGrid package.

We adopted this platform to simulate applications with different number of tasks (1000 and 10000 tasks) and quantity of computation per task (100, 500, 1000 and 2000 MFlop/s) using deployments with 50 and 100 agents, and we assumed that communication costs to send one task to an agent is fixed (0.001 Mbyte/s) and to receive the result is irrelevant. The simulation results as shown in the following four figures.

Fig. 2 illustrates the measurements obtained for an application containing 1000 tasks begin executed in 50 agents of the platform, with computation amount per task varying from 200 to 2000 MFlop/s. Using 50 agents and computation quantity ranging from 600 to approximately 1000 MFlop/s, the strategy presented the best results.

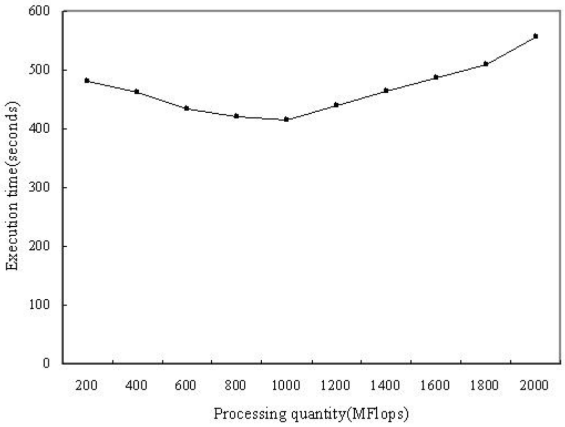


Fig.2. Measurements scheduling 1000 tasks using 50 agents

Fig. 3 illustrates the measurements obtained for an application containing 1000 tasks begin executed in 100 agents of the platform, with computation amount per task varying from 200 to 2000 MFlop/s. Using 100 agents and computation quantity ranging from 1400 to 2000 MFlop/s, the strategy behaved stable.

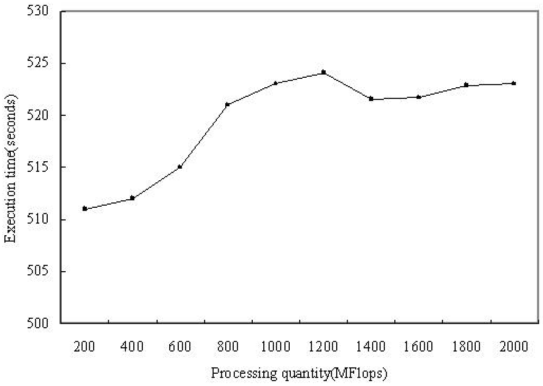


Fig.3. Measurements scheduling 1000 tasks using 100 agents

In Fig. 4 the same measurements are presented for an application with 10000 tasks. Using tasks with no more than 1200 MFlop/s the strategy presented the best results. It behaved better from 200 to approximately 600 MFlop/s than after 600 MFlop/s.

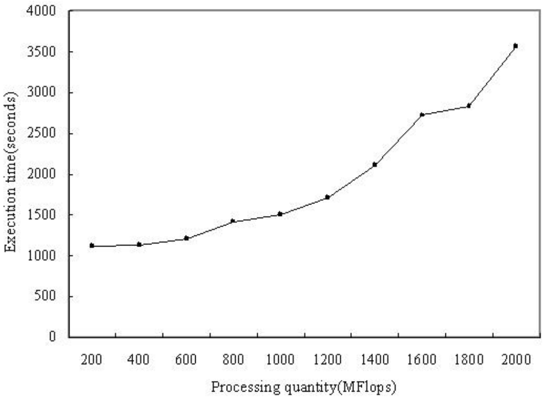


Fig.4. Measurements scheduling 10000 tasks using 50 agents

In Fig. 5 the measurements are presented for an application with 10000 tasks executed in 100 agents. Using tasks with no more than 1200 MFlop/s the strategy presented the best results.

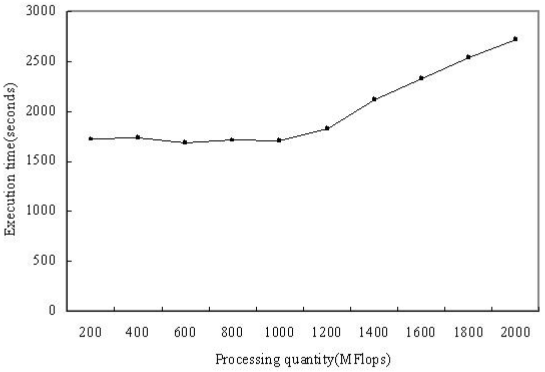


Fig.5. Measurements scheduling 10000 tasks using 100 agents

5. CONCLUSIONS

In this paper, we compared several typical scheduling strategies, and on the base of analyzing result we proposed a scheduling strategy based on an agent-oriented algorithm, then we simulated it by using SimGrid toolkit. The result shows that the scheduling strategy has good expandability, can realize global optimum scheduling, and can also achieve load balancing well. It is an efficient approach for grid scheduling. It is promising and a good alternative to be considered when a scheduling algorithm is needed for scheduling applications in grid computing.

REFERENCES

[1] T. Ferreto, C. De Rose, and C. Northfleet. “Scheduling BoT Application in Grids Using A Slave Oriented Adaptive Algorithm”, *Proceeding of the Parallel and Distributed Processing and Applications: Second International Symposium (ISPA2004),* Hong Kong, China, December 13-15, 2004, Lecture Notes in Computer Science, vol. 314, September 2004, pp. 392-398.

[2] H. Tian and L. Duan, “Resource Management and Scheduling Model in Grid Computing Based on An Advanced Genetic Algorithm”, *Proceeding of the 5th International Conference on Distributed Computing and Applications for business, engineering and sciences (DCABES'2006),* Hangzhou, China, October 2006, Shanghai University Press, Vol.1, pp. 238-242.

[3] Y. Du, Y. Chen and P. Liu. *Grid Computing,* 1st ed., Tsinghua University Press, 2002, pp.86-90.

[4] S.Vismanathan and B.Veeravalli, “Design and Analysis of A Dynamic Scheduling Strategy with Resource Estimation for Large-scale Grid Systems”,*Proceeding of the 5th IEEE/ACM International Workshop on Grid Computing* (GRID’04),2004.

[5] A.Chronopoulos et al, “A Class of Loop Self-Scheduling for Heterogeneous clusters”, *Proceeding of CLUSTER’2001*, 2001.

[6] P. Kacsuk et al, “High-Level Grid Application Environment to Use Legacy Codes as OGSA Grid Services”, *Proceeding of the 5th IEEE/ACM International Workshop on Grid Computing* (GRID’04), 2004.

[7] L. Yang, J.M. Schopf and I. Foster, “Conservative Scheduling: Using Predicted Variance to Improve Scheduling Decisions in Dynamic Environments”, *Proceedings of the ACM/IEEE Supercomputing* 2003 Conference, November 2003, p.31-36.

[8] J. Soldatos, E. Vayias and L. Polymenakos, “Grid Donors Resources Utilization Analysis towards Optimal Job Scheduling”, *Proceedings of the DPSN '04, Workshop, held in the Scope of IFIP Networking,* 2004, Athens, May 14th 2004.

[9] R. Buyya, D. Abramson and J. Giddy, “Nimrod/G: An Architecture for A Resource Management and Scheduling System in a Global Computational Grid”, *Proceedings of the International Conference on High Performance Computing in Asia-pacific Region* (HPC Asia 2000), Beijing, China. IEEE Computer Society Press, USA, 2000.

[10] H. Casanova, “Simgrid: A Toolkit for the Simulation of Application Scheduling”, *Proceedings of the IEEE Symposium on Cluster Computing and the Grid* (CCGrid’01). 2001.

Hao Tian is an instructor of the School of Computer Science and Technology, Hubei University of Economics. He graduated from Wuhan University of Technology and received his Master degree in 2005. His research interests are in distributed parallel processing, grid computing and communication technology.

Lijun Duan is an instructor of the Department of Computer and Information Engineering, Hubei Institute of Economics Management. She graduated from Wuhan University of Technology in 2003. Her research interests are in computer applications, soft engineering and grid computing.

Software Sharing Technology Based on Grid

Hong Zhao [1,2], Woxur[1], Mixia Liu [2], Yong[1] Hou, Shengwei Tian [1]
[1] College of Information Science and Engineering, Xinjiang University, Urumchi, Xinjiang 830046, P.R. China,
[2] School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, P.R. China
Email: zhaoh@lut.cn

ABSTRACT

A grid platform constructed by GT4 based on OGSA is put forward to solve the problem of disk space wastage, process speed slowness and overmuch cost on buying software because of installing the same software in different computers. And with the remote desktop process technology, the remote software sharing and dynamic load balance of software servers is realized. The software server is integrated in the grid by deploying Web Service on this software server. The CA server and UDDI server are configured on the grid platform. The grid portal is also constructed on the grid platform. User certificate and operate privilege are tested by CA server; the service of uniform description, register, notification and invocation are realized by UDDI server; the uniform interface of software processing is provided by grid portal. The realization and processing effect are illustrated at last.

Keywords: Grid; Web Service; Remote Desktop Process; Uddi

1. INTRODUCTION

At present, the same software is installed in different computers repeatedly. This phenomenon results in disk space waste, process speed slowness and overmuch cost on buying software, which is popular, especially in each kind of computer laboratories where the identical software is equipped to some computers through “cloning” by the “Ghost” software.

With the development of computer hardware, computer process ability is improved greatly, but its using rate is becoming lower. According to the statistics of 1990s, average using rate of computer process ability is 15% during daytime, only 5% during night [1]. The using rate of computer process ability will become much lower as the computer’s performance is improved and the quantity is added.

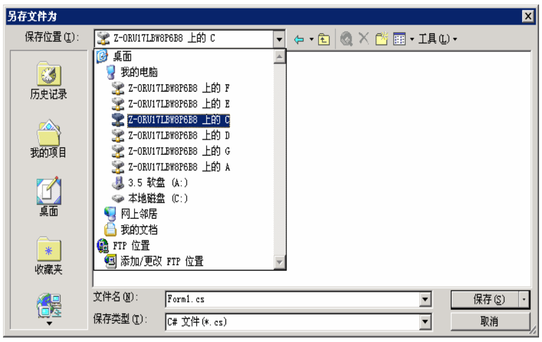
The problems stated above can be well solved by using grid and remote desktop process technology developed in recent years.

2. REMOTE DESKTOP PROCESS

The RDP (Remote Desktop Process) technology is upgraded from TELNET [2], which can support graphical interface. The RDP technology is integrated in Windows 2000 Server firstly by Microsoft, and is improved in Windows 2003 for more convenient and credible accessing. Similar software is VNC (Virtual Network Computing) in Linux. The similar function of RDP is one of functions provided by most modern operating systems.

If a computer is deployed RDP as a server, other computers can log on it through network as remote clients. After username and password are verified, user on clients can operate the server remotely as if on the local computer, for example, installing software, running program and so on.

While logging on the server remotely, the user can map client disks to server to save data produced by server as shown in fig.1.



保存位置(I): Z-ORU17LBW8P6B8 上的 C
文件名(N): Form1.cs
保存类型(T): C# 文件(*.cs)

Fig.1. Saving data produced by server in the local computer disk

3. GRID

The grid technology is one kind of new technology developed recent years. It attempts to share all resources of the Internet, such as resources of computation, memory, communication, software, data, information, knowledge and so on. The grid technology enables these resources possible to cooperate with others on the Internet, then makes the Internet become a giant super computer, and provides plug and play services for the users [3].

The research of the grid structure has already been transformed from the layered hourglass model that focuses on protocol [3,4] to OGSA (Open Grid Service Architecture) that focuses on service [3,6]. In the OGSA, all kinds of resources including that of computation, memory, procedures and databases are encapsulated in form of services. From resources to services, the transformation can unify resources, information and data so to realize flexible, consistent and dynamic share mechanism easily. All resources are considered as grid service in OGSA. So grid is the extensible set of Grid Services, namely Grid={Grid Services}.

In a Grid Service there are several interfaces that resolve operations of discovering service, creating services dynamically, managing life period of service and notifying changes of service, i.e., grid service = interfaces + service data. The simple description of grid service is shown in fig.2.

From fig.2, we can see that grid service can send data of service to request and receive request of service through grid service interface and other interfaces.

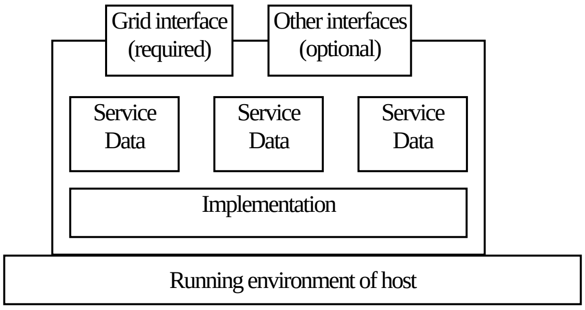

Grid interface (required)
Other interfaces (optional)
Service Data
Service Data
Service Data
Implementation
Running environment of host

Fig.2. Simple description of grid service

Grid service is based on Web Service and consistent with WSRF (Web Service Resource Framework) criterion. It results from Web Service appended OGSI (Open Grid Service Infrastructure)[1].

4. REALIZATION OF SOFTWARE SHARE

4.1. Design of Software Share Scheme

GT4[7,8] (Globus Toolkit 4) is a new grid tool according with WSRF criterion developed by the Globus Alliance. Grid platform is established by GT4, and then software servers are integrated in the grid for software sharing by deploying Web Service on these servers. The UDDI (Universal Description, Discovery, and Integration) server is deployed on the grid platform that realizes service register, issue, query and matching. All services provided by the grid need to be registered, issued and then accessed. For ensuring legal users to access software service provided by the grid, the CA (Certificate Authority) server is deployed on the grid platform. The CA is the center of issuing, signing and verifying digital certificate. Any action needs to be corresponding with the digital certificate in the grid which indicates legal identity and relevant right. The grid portal provides users the uniform interface of accessing software. Architecture scheme is shown in fig.3.

In fig.3, the function of CA server, UDDI server and grid portal may be deployed on one computer, also may be distributed to several computers.

Each kind of software is installed in the software servers for sharing. When a software server is initiated, the Web Service deployed on the software server will register the services in the UDDI server automatically that the software server can provide. UDDI server always registers the service information provided by each software server. When one software server is shut down or stopped providing services, the Web Service deployed on the software server will report to the UDDI server and request for stopping services. Then the UDDI server will delete the service information of the software sever from its service registration automatically. When one software server becomes invalid due to any faults, it cannot communicate with the UDDI server. In this case, the UDDI server cannot receive the messages subscribed from the software server in scheduled time, so may think the software server has withdrawn from the grid, and then will renew its service registration information to avoid user access invalid service. All information of software services provided to the user by the grid has been recorded in the UDDI server.

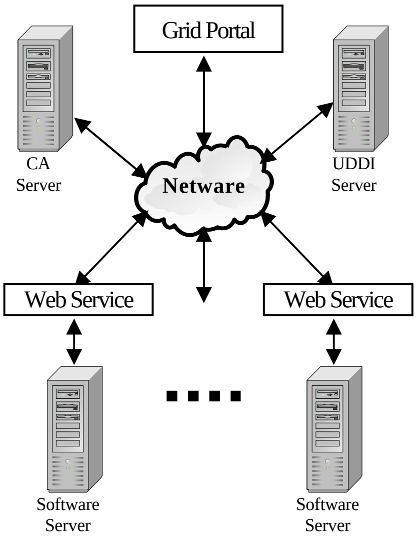


Fig.3. System architecture scheme

4.2 Realization of Dynamic Load Balance

The identical software is simultaneously installed in different software servers possibly, and different users possibly need to access different software or the identical software at the same time. Therefore, the load of various software servers needs to be balanced dynamically in the grid.

Besides collecting information of services provided by the software server, the Web Service deployed on the software server is also responsible for regularly collecting dynamic information of current free size of memory, CPU using rate and processor quantity of the software server. The UDDI server has subscribed the information from various software servers, and the Web Services will transmit the information of the software servers to the UDDI server regularly. Length of the “Regular time” is assigned by the UDDI server and may be modified according to the condition of being busy of the UDDI server.

When one user accesses some software, the UDDI server will query which software servers may provide the software service according to user’s request, then calculate usable computation capacity of the candidate software servers according to the empirical formula (1). The software server that owns the biggest usable computation capacity will be selected to provide the service for the user.

$$C_{spare}=CPU_{num}*CPU_{speed}*(1-CPU_{used})*40\%+M_{free}*40\%+(P_{max}-P_{cur})*20\% \quad (1)$$

Note: The influence of network transmission performance to usable computation capacity is not considered in formula (1).

In formula (1), C_{spare} is usable computation capacity of the software server, CPU_{num} is CPU quantity, CPU_{speed} is CPU basic frequency, CPU_{used} is CPU current using rate, M_{free} is free size of memory, $P_{\max}$ is maximal quantity of processes running in the software server, and P_{cur} is quantity of processes running in the software server currently.

From formula (1) we can know that CPU ability accounts for 40%, free memory accounts for 40%, and the quantity of permitted processes accounts for 20% in the usable computation capacity of software server, C_{spare} .

4.3 The RDP file for Connecting

The RDP file is a kind of text file with RDP expanded name, which can be built by the UDDI server automatically and can be edited directly by the text editor. When one carries out remote desktop connection, the RDP file may be used to connect the software server. The RDP file contains the information of remote desktop connection, for example, server's IP address, username and password for connecting, the information of whether mapping the local disks to the server and so on.

When user accesses some software, the UDDI server will select an appropriate software server that can provide the software service according to the situation of various software servers and their current usable computation capacity. Then the UDDI server organizes the RDP file according to the selected software server, and then returns this RDP file to the user's computer. The RDP file will be opened on the user's computer by the grid portal and connects remote software server automatically. The user does not know previously which software server will be selected to provide the software service, namely the selection is transparent to the user.

4.4 The Entire Procedure of Accessing Software Service

When one user accesses some software on the grid platform, the unified interface provided by the grid portal will be visited. After confirming of the identity and corresponding right, the grid portal transmits user's request to the UDDI server. The UDDI server matches the request, and selects the best candidate of software server, then organizes the RDP file according to the matched result, and returns the RDP file to user. The RDP file will be opened automatically in user's computer by the grid portal and the user will log on the software server in remote mode automatically.

5. CONCLUSIONS

By using the technology described as above, we have organized 15 PC servers as software servers which can be shared by all members of the whole college, these servers have been placed in network laboratory, software training room, microcomputer laboratory and communication laboratory respectively, installed MATLAB, 3DMAX, PRO/E and so on. We integrated CA server, UDDI server and grid portal function in an ordinary PC. The running effect is quite satisfied with three months testing.

We plan to carry on the experiment under the Linux environment next step, and consider the influence of network transmission to the usable computation capacity of software server in formula (1), as noted, this influence has not been considered previously.

REFERENCES

[1] I. Foster, C. Kesselman. *The GRID 2[M]*. Beijing: China Machine Press, 2005, 4.

[2] Lu Aiqing, Zhang Huiyong, Zhao Zheng. "Implementation Principle and Application of Telnet Protocol [J]." *Compute Engineering*, 2002, 28(11), pp.268~270.

[3] Dou Zhihui,Li Sanli,Chen Yu,Liu Peng,*Grid Computing* [M],Beijing: Tsinghua University Press,Oct 2002.

[4] I.Foster,C.Kesselman,J.Nick et al,*The physiology of the grid:*"An open grid services architecture for distributed systems integration [EB]." http://www.globus.org/research/papers/ogsa.pdf,Oct 2004.

[5] I. Foster. *A globus Primer [EB]*. http://www.globus.org/toolkit/docs/4.0/, 2005.6.

[6] Borja Sotomayor. *The Globus Toolkit 4 Programmer's Tutorial* [EB]. http://gdp.globus.org/gt4-tutorial/,Jun 2005.

[7] GGF. GT4 Admin Guide [EB]. http://www-unix.globus.org/toolkit/docs/4.0/admin/docbook/,Jun 2005.

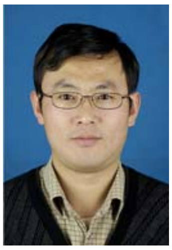
[8] I. Foster, C. Kesselman, S Tuecke. "The anatomy of the grid: Enabling scalable virtual organizations[J]," *International Journal of Supercomputer Applications*, 2001,15(3),pp.200~222.

[9] Joshy Joseph, Craig Fellenstein. GRID COMPUTING[M]. Beijing: Tsinghua University Press, 2005,1.

[10] Zhao Hong, Yu Dongmei. "A study on improving data transmission capability in grid using compression technology [J]". Journal of Lanzhou University Technology, 2005, 31(4), pp.101~103.

[11] Xun Zhiwei, Feng Baiming,Li Wei. *Grid Computing Technology [M]*.Beijing:Publishing House of Electronics Industry,May 2004.

[12] Zhao Hong, Liu Peng,"Data grid construction and system integration [J],"*Computer Engineering and Design*, 2006, 23(12),pp.4424~4426.



Hong Zhao is an associate professor of Lanzhou University of Technology and a doctor student of Xinjiang University. He graduated from Northwest Normal University in 1993, since then he has been working in Lanzhou University of Technology as a faculty. Since 2006, he has studied in the college of information science and engineering of Xinjiang University for doctor degree. He has taken charge nine projects and over 20 journal papers. His interests of research are grid computing, intelligent information processing and distributed parallel processing.

Research Summary on Synergy between P2P Computing and Grid Computing

Hongwei Chen
School of Computer Science and Technology, Hubei University of Technology
Wuhan, Hubei Province, 430068, China
Email: chw2001@sina.com

ABSTRACT

Peer-to-peer (P2P) Computing and Grid Computing are promising distributed computing. With their development, the synergy between Grid Computing and P2P Computing is inevitable, and it will benefit them to make up their deficiencies. Their synergy can utilize more desktop PC resources, effectively avoid single failure of server, extend application domain of Grid Computing, adopt excellent resource discovery means such as DHT, and make use of standard architecture OGSA to realize genuine Grid Computing. In this paper, the difference between Grid Computing and P2P Computing is analyzed, and synergic phase, synergic significance, synergic pattern, synergic paradigm, synergic problem between Grid Computing and P2P Computing are summarized.

Keywords: Synergy, Peer-to-peer Computing, Grid Computing, Open Grid Services Architecture, Autonomic Domain

1. INTRODUCTION

P2P Computing and Grid Computing are two new approaches to distributed computing. Both claims to address the problem of organizing large scale computational societies. P2P Computing is the sharing of computer resources and services by direct exchange between systems, and it does not have fixed clients and servers, but a number of peer nodes that function as both clients and servers to the other nodes on the network [1]. Grid Computing is concerned with “coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations.” The key concept is the ability to negotiate resource-sharing arrangements among a set of participating parties (providers and consumers) and then to use the resulting resource pool for some purpose [2]. Many researchers are looking into the practical uses of these two technologies.

Synergy between P2P Computing and Grid Computing has experienced three phases: (1) Year 2002 or so, several papers had brought forward concepts about synergy between them. (2) From year 2002 to now, several projects have researched about synergy between them. (3) Since year 2005, the standard document about synergy between them has been published.

2. DIFFERENCE BETWEEN P2P COMPUTING AND GRID COMPUTING

Foster [3] and Talia [4] have compared and contrasted with P2P Computing and Grid Computing. Both the goal of P2P Computing and that of Grid Computing are resource sharing and collaboration. However, their initial application is different: P2P computing initiates from File Sharing, and Grid Computing initiates from Scientific Computing. This is the core difference between P2P Computing and Grid Computing. Other differences such as resource type, system scale, resource discovery, connectivity, service and infrastructure, position in WAN, and security policy between them are all based on the difference of their initial application. The following table is the difference between P2P Computing and Grid Computing.

Table 1. Difference between P2P Computing and Grid Computing

Difference	P2P Computing	Grid Computing
Initial Application	File Sharing	Scientific Computing
Resource Type	Mainly home PC	More powerful and diverse
System Scale	Millions	Thousands
Resource Discovery	More mature, active report policy	Centralized or layer model
Connectivity	Flexible, but not reliable	Reliable, but not Flexible
Service and Infrastructure	Simple service integration	Service aggregation
Position in WAN	Edge of WAN	Center of WAN
Security Policy	Low	High

3. SYNERGIC SIGNIFICANCE BETWEEN P2P COMPUTING AND GRID COMPUTING

P2P Computing and Grid Computing are new generation and promising distributed computing technology. Although they have significant differences, they can complement each other. Synergy between Grid Computing and P2P Computing has the following huge significance.

(1) Synergy can utilize low-end desktop PC resources, so it sufficiently utilizes all resources in WAN.
(2) Synergy can weaken function of centralized server, and pay more attention to function of all members in the network. So it can effectively avoid influence from single failure.
(3) Synergy can make use of superiority of P2P technology such as file searching, and extend application area of Grid Computing.
(4) Synergy can make use of superiority of P2P technology such as resource discovery (Distributed Hash Table), and make up deficiency of grid resources.
(5) P2P can sufficiently utilize existing architecture of Grid Computing (Open Grid Services Architecture, OGSA). So different P2P systems can interact in the standard protocol, and one can obtain resources and services from other P2P systems.

4. SYNERGIC PATTERN BETWEEN P2P COMPUTING AND GRID COMPUTING

Kazuyuki Shudo[6] has summarized three kinds of synergic pattern between P2P Computing and Grid Computing: Desktop Grid pattern, P2P over Grid pattern, and Grid over P2P pattern. In the above three patterns, “Grid” means traditional centralized resource management or cluster mode, “P2P” means full distributed resource management mode. However, his division method is not perfect.

Foster [5] has suggested the essence of Grid Computing which can be captured in the following three points: (1) coordinates resources that are not subject to centralized control, which means that a Grid integrates and coordinates resources and users that live within different control domains. (2) using standard, open, general-purpose protocols and interfaces. (3) to deliver nontrivial qualities of service. From the above definition, we consider that a Grid is consisted of many self-managed domain (autonomic domain). Grid resources can be divided into intra-domain resource and inter-domain resource. Furthermore, grid resources can be managed in full distributed mode or centralized mode. So we can divide different kinds of Grid Computing according autonomic domain and resource management mode. The following table is Synergic Pattern between P2P Computing and Grid Computing.

Table 2. Synergic Pattern between P2P Computing and Grid Computing

Synergic Pattern	Description
Desktop Grid	It aggregates, discovers and uses resources based on P2P mode in a single Autonomic Domain.
P2P over Grid	It aggregates resources based on cluster mode in an Autonomic Domain, and discovers inter-domain resources based on P2P mode among different Autonomic Domains.
Grid over P2P	It aggregates resources based on P2P mode in an Autonomic Domain, and schedules inter-domain resources based on cluster mode among different Autonomic Domains.
P2P over Hybrid	It aggregates resources based on P2P mode or cluster mode in Autonomic Domain, and discovers inter-domain resources based on P2P mode among different Autonomic Domains.
P2P over P2P	It aggregates resources based on P2P mode in an Autonomic Domain, and discovers inter-domain resources based on P2P mode among different Autonomic Domains.

5. SYNERGIC PARADIGM BETWEEN P2P COMPUTING AND GRID COMPUTING

Table 3. Synergic Paradigm between P2P Computing and Grid Computing

Synergic Paradigm	Synergic Pattern	Application
GRIDNUT [7]	P2P over Grid	Resource Discovery
ATLAS [8]	Grid over P2P	Resource Discovery
DDGrid (Drug Discovery Grid) [9]	P2P over Grid	Resource Discovery
PGS(P2P Grid Scheduler) [10]	Grid over P2P	Job Scheduling
P2P Grid [11]	Grid over P2P	Desktop Computing
IRTL (Information Resource Transaction Layer) [12]	Grid over P2P	Message Interaction
GSM(Grid Service Monitor) [13]	Grid over P2P	Performance Monitoring
Self-Optimizing Grids [14]	Grid over P2P	Data Management
PCMA (P2P-based Content Management Architecture with Grid Properties) [15]	P2P over Hybrid	Content Management

6. SYNERGIC PROBLEM BETWEEN P2P COMPUTING AND GRID COMPUTING

The P2P Working Group (P2PWG) and Global Grid Forum (GGF) are the main academic organizations focusing on research of P2P. The P2PWG and GGF have merged, and GGF is responsible for standardization of P2P and Grid Computing. A new Global Grid Forum paper examines ways to make grid computing and peer-to-peer (P2P) applications work together. The paper, “Peer-To-Peer Requirements on the Open Grid Services Architecture Framework” [16], is the work of the grid forum's OGSAP2P research group that looked at ways to use the open-source grid framework for P2P applications. Synergy between Grid Computing and P2P computing needs taking the following pivotal technologies into account [16].

Table 4. Synergic Problem between P2P Computing and Grid Computing

Synergic Problem	Description
Scalability	It must resolve scalability of resources, users and organizations.
Connectivity	It must make sure peer-to-peer, end-to-end and transparent communication; and allow distributed sharing of resources, information and services.
Discovery	It must resolve discovery problem of other resource nodes, metadata from other nodes, or virtual organizations.
Security	It specially needs resolving trust problems such as identity trust, resource trust and data trust.
Resource Availability	It mainly considers offline, running failure and QoS of resources.
Location Awareness	It provides absolute, relative and context support capability via applications.
Group Support	It aggregates peers, resources and users which have common features.

7. CONCLUSIONS

P2P Computing and Grid Computing are promising distributed computing. The synergy between P2P Computing and Grid Computing is inevitable, and it benefits them to make up their deficiencies. P2P Computing and Grid Computing will unite as a whole, and Grid Computing is the final synergic objective. Their Combination will generate many new changes such as trust management and resource discovery and so on. In the future, we will aim at these changes to study some key technologies about synergy between P2P Computing and Grid Computing.

REFERENCES

[1] Dejan S. Milojicic, Vana Kalogeraki, Rajan Lukose, Kiran Nagaraja, Jim Pruyne, Bruno Richard,Sami Rollins, Zhichen Xu.Peer-to-Peer Computing. http://www.pl.hp.com/techreports/2002/HPL-2002-57R1.pdf

[2] Forster I, Kesselman C, Tuecke S. "The Anatomy of the Grid: Enabling Scalable Virtual Organizations". *International Journal of Supercomputer Applications*, 2001, 15(3): 200-222.

[3] Foster I., Iamnichi A. "On Death, Taxes, and Convergence of P2P and Grid Computing", *IEEE Internet Computing*, Jan. 2003.

[4] Talia D. , Trunflo P. "Toward a synergy between P2P and grids". *Internet Computing, IEEE*. 7[4]. 2003:94-96.

[5] Foster I. "What is the Grid? A three-point checklist". *Grid Today-Daily News and Information for the Global Grid Community*, 2002(6).

[6] Kazuyuki Shudo, Integration Patterns of P2P and Grid Technologies. http://www.shudo.net/publications/ KR-JP-Grid-Symposium-200412-P3/shudo-KR-JP-Grid- Symposium-200412.pdf.

[7] Talia D., Trunflo P. Peer-to-Peer Services for Distributed Resource Discovery on Grids. http://dcs.ics.forth.gr/coregrid-workshop/bstracts/ Domenico-Talia.pdf.

[8] Manolis Koubarakis, Iris Miliaraki, Zoi Kaoudi, "Matoula Magiridou and Antonis Papadakis-Pesaresi, Semantic Grid Resource Discovery using DHTs in Atlas". *3rd GGF Semantic Grid Workshop. Athens Greece*. February 13-16, 2006.

[9] Shudong Chen, Xuefeng Du, Fanyuan Ma, Jianhua Shen. "A Grid Resource Management Approach Based on P2P Technology.High-Performance Computing in Asia-Pacific Region, 2005". *Proceedings. Eighth International Conference on 2005 Page*(s):362 - 369.

[10] Cao J., Kwong O., Wang, X., Cai W. "A Peer-to-Peer Approach to Task Scheduling in Computation Grid". *In Proceedings of the 2nd International Workshop on Grid and Cooperative Computing* (GCC 2003), Shanghai, China, December 2003.

[11] Prem Uppuluri, Narendranadh Jabisetti, Uday Joshi, Yugyung Lee. "P2P Grid: Service Oriented Framework for Distributed Resource Management". *2005 IEEE International Conference on Services Computing* (SCC'05) Vol-1, pp. 347-350.

[12] Junseok Hwang, Aravamudham P., Liddy E., Stanton J., Maclnnes I. "IRTL (information resource transaction layer) middleware design for P2P and open GRID services". *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference* on 6-9 Jan 2003 Page(s):10 pp.

[13] Yin Chen. Grid Service Monitor: Performance Monitoring and Measurement of Grid Services Using Peer-to-Peer Techniques. http://www.inf.ed.ac.uk/publications/thesis/online/IT050246.pdf.

[14] Schintke F., Schutt T., Reinefeld A. "A framework for self-optimizing grids using P2P components Database and Expert Systems Applications", 2003. *Proceedings. 14th International Workshop* on 1-5 Sept. 2003 Page(s):689 – 693.

[15] Qian Zhang, Yu Sun, Zheng Liu, Xia Zhang, Xuezhi Wen. "Design of a distributed P2P-based grid content management architecture". *Communication Networks and Services Research Conference, 2005. Proceedings of the 3rd Annual* 16-18 May 2005 Page(s):339 – 344.

[16] Karan Bhatia , Per Brand , Sergio Mendiola, Alex Mallet , Karlo Berket. "Peer-To-Peer Requirements On The Open Grid Services Architecture Framework," http://www.ggf.org/documents/ GFD. 49. pdf.

Hongwei Chen (1975-), male, from Hubei Province, PHD, Lecturer of Hubei University of Technology, interested in Grid Computing, Peer-to-Peer, Information Security, Mobile Agent.

Multi-agent Application

A Multi-agent Cooperative System Based on 3APL*

Shengfu Zheng[1], Shanli Hu[1,2], Xinjian Lin[1], Chaofeng Lin[1], Shexiong Su[1]
[1]Department of Computer Science and Technology, Fuzhou University. Fuzhou 350002, China
[2]Key Laboratory for Computer Science Chinese Academy of Sciences, Beijing 100080, China
Email: sfz_roger@yahoo.com.cn

ABSTRACT

Most traditional AOP (agent-orient programming) approaches lack special consideration for CPS (cooperative problem solving).A kind of multi-agent cooperative system based on 3APL (An Abstract agent Programming Language) is presented. This paper presents an EBNF (Extended Backus-Naur Form) based 3APL syntax for individual agent with introducing cooperative plans expression. Refer to a new kind of multi-agent cooperative operation semantics, which combines reasoning rules, stack, configuration, and transition rules. Additionally, we study the multi-agent cooperative deliberation cycle processes deeply. Further, a cooperative example is analyzing for addressing how to generate and revise plans between agents.

Keywords: 3APL, Cooperation, Transition rules, Cooperative Deliberation Cycle

1. INTRODUCTION

The agent-oriented theory and technology is an important field of DAI, software engineering, robot technology, etc. AOP (agent-orient programming) is a new programming paradigm that supports a societal view of computation and extends conventional OOP (Object-Oriented Programming). Several AOP approaches have been proposed (e.g. Agent-0[10], CONGOLOG[2], 3APL[3], etc.), each one with its own characteristics and applicability.

Agents are autonomous entities able to perform their task without requiring a continue user interaction. However, new domains are emerging that impose new requirements for teamwork. It is a pity that most of these languages however lacked an important element of multi-agent system (i.e. cooperative problem solving).

3APL (An Abstract Agent Programming Language) is a programming language for implementing cognitive agents that have beliefs, goals, and plans as mental attitudes, can generate and revise their plans to achieve their goals, and are able to interact with each other and with the environment they share with other agents. The first version of 3APL was designed by Hindriks[3] et al.

In order to consider specified cooperative problem solving between agents, we present a kind of multi-agent cooperative system based on 3APL. We also propose a new multi-agent cooperation operational semantic based on transition system. Additionally, we study the multi-agent cooperative deliberation cycle processes deeply, and extends Hindriks' work.

2. SYNTAX

Let L be the first order language with negation ($\neg$) and conjunction ($\wedge$).The symbol $\models$ will be used to denote the standard entailment relation for L . $\top\in L$ will be used to denoted a tautogy, $\bot\in L$ to denoted falsum. Further, we assume belief formulas L_B , goal formulas L_G plan formulas L_P , they satisfy $L_B, L_G, L_P \subseteq L$.

Definition 1 beliefs and goals formulas

The belief formulas L_B with typical element β and the goal formulas L_G with typical element κ are defined as follows[6].

$\top\in L_B$ and $\top\in L_G$

If $\phi\in L$ then $B\phi\in L_B$, and $G\phi\in L_G$

If $\beta,\beta'\in L_B$ and $\kappa,\kappa'\in L_G$ then $\neg\beta,\beta\wedge\beta'\in L_B$ and $\neg\kappa,\kappa\wedge\kappa'\in L_G$.

Definition 2 Plans

If well-formed formula $\beta,\beta'\in L_B$ and $BasicActions$ be the set of basic actions, then the set of plans, denoted by $Plans$ is defined as follows:

$BasicActions \subseteq Plans$

if $\pi,\pi'\in Plans$, then $if\ \ \beta\ \ then\ \ \pi\ \ else\ \ \pi'\ \ \in Plans$

if $\pi\in Plans$, then $while\ \ \beta'\ \ do\ \ \pi\ \ \in Plans$

if $\pi,\pi'\in Plans$, then $\pi;\pi'\ \ \in Plans$

if $\pi,\pi'\in Plans$, then $\pi\,||\,\pi'\ \ \in Plans$

There are four 3APL program operators: the sequential operator (denoted by ;), the iteration operator (denoted by a while-do construct), the conditional choice operator (denoted by an if-then-else construct), and the cooperative concurrence operator (denoted by ||). In particular, we use $\pi\,||\,\pi'$ to denote agent executes plan π itself and others which join the CPS execute plan π' . And it satisfies $\pi\,||\,\pi'\not\models\bot$ .We use Ω to denote the empty plans, and identify $\Omega;\pi$ and $\pi;\Omega$ with π .

* This paper is supported by the National Natural Science Foundation of China under Grant No.60573076.

Definition 3 Reasoning Rules
If $\beta \in L_B$, $\kappa, \kappa_h, \kappa_b \in L_G$,and $\pi, \pi_h, \pi_b \in L_P$,then reasoning rules for goals GR ,their interactions IR ,and plans PR are defined as follows[5].

$$\kappa_h \leftarrow \beta \mid \kappa_b \in GR$$

$$\kappa \leftarrow \beta \mid \pi \in IR$$

$$\pi_h \leftarrow \beta \mid \pi_b \in PR$$

The goals reasoning rules are used to generate , revise or drop goals.

$\top \leftarrow \beta \mid \kappa_b$ goal κ_b generate

$\kappa_h \leftarrow \beta \mid \kappa_h \wedge \kappa_b$ goal κ_h extends to $\kappa_h \wedge \kappa_b$

$\kappa_h \leftarrow \beta \mid \top$ goal κ_h drops

The interactions reasoning rules are used to generate plans to achieve goals. Finally, the plan reasoning rules are used to revise and drop plans.

Definition 4 Agent
An agent is a tuple $<\sigma, \gamma, R, \tau, \xi>$, where $\sigma \in L_B$ is the belief base, $\gamma \in L_G$ is the goal base , $R = 2^{GR} \times 2^{IR} \times 2^{PR}$ is the rules base, $\tau : BasicActions \times L_B \rightarrow L_B$ is the belief update function resulting from the execution of basic actions, ξ is the environment it interacts with. It satisfy that $<\sigma,\gamma,R,\tau,\xi> \models B\beta \Leftrightarrow \sigma \models \beta$, $<\sigma,\gamma,R,\tau,\xi> \models G\kappa \Leftrightarrow \gamma \models \kappa$.

Definition 5 Stack
The set of agent stack $S = s \mid s.S$, where the symbol "." denotes "push" a stack onto a stack. Single element of a stack $s =< \phi, \pi, \delta, \rho >$, where current goal $\phi \in L_G$,the plan $\pi \in L_P$, the interactions rules $\delta \subseteq IR$ and the plans rules $\rho \subseteq PR$, which correspond to ϕ

Definition 6 Configuration
A configuration of individual agent is a tuple $< \sigma, \gamma, S >$, where σ, γ, ξ are as in Definition 4, S in Definition 5. Multi-agent configurations is a tuple $< A_1, \cdots\cdots, A_n, \xi >$, where $A_i (1 \le i \le n)$ is individual agent configuration, ξ is the specifications of the share environment.

In order to describe the cooperation for multi-agent, we introduce cooperation plans expression <co_plans> (see Fig.1). The expression <co_plans> depicts agent sends messages and synchronizes its plan with others'. Extended Backus-Naur Form(EBNF) is frequently used to specify programming languages, operating system commands, and other types of computer input. The symbol on the left-hand side of the "::=" must be replaced by one of the alternatives on the right hand side(we use"::="instead of ":=", but the meaning is the same). The symbol "?" means that the symbol (or group of symbols in parenthesis) to the left of the operator is optional (it can appear zero or one times). The symbol "*" means that something can be repeated any number of times (and possibly be skipped altogether). The alternatives are separated by symbol "|". In that[7] specifications, we use <atom> to denote an atomic formula the terms of which can include Prolog-like list representations of the form [a,b,[3,f]], [X|T], and [a,[4,d]|T], etc. Moreover, we use <ground_atom> to denote a ground atomic formula, which is an atomic formula that contains no variables. The terms of ground atomic formulae can include Prolog-like list representations such as [a,b,c], [e,[9,d,g],3]. Finally, we use <Atom> to denote atomic formulae where the predicate letter starts with a capital letter, <ident> to denote a string, <var> to denote a variable，and <query> to denote a belief and goal query expression which are either the special atomic formulae true or a well-formed formula (i.e. <wff>) constructed from atoms and logical connectors[7].

```
<Program> ::=          "Program" <ident>
                       ("Load" <ident> )?
                       "Capabilities :" ( <capabilities> )?
                       "BeliefBase :" ( <beliefs> )?
                       "GoalBase :" ( <goals> )?
                       "PlanBase :" ( <plans> )?
                       "PG-rules :" ( <p_rules> )?
                       "PR-rules :" ( <r_rules> )?
<capabilities>::=      <capability> ( ";" <capability> )*
<capability> ::=       "{"<query>"}" <Atom> "{"<literals>"}"
<beliefs> ::=          ( <belief> )*
<belief> ::=           <ground_atom> "." | <atom> ": -"
                       <literals>"."
<goals> ::=            <goal> ( ";" <goal>)*
<goal> ::=             <ground_atom> ( "and" <ground_atom> )*
<plans> ::=            <plan> ( ";" <plan> )*
<plan> ::=             <basicaction> | <composedplan>
<basicaction> ::=      "nil" | <Atom> |
                       "Send("<iv>,<iv>,<atom>")"
                       |"Java("<ident>,<atom>,<var>")" |
                       <wff>"?" | <atom>
<composedplan> ::= "if" <wff> "then" <plan> ( "else" <plan> )?
                       |"while" <query> "do" <plan> |<plan> ";"
                       <plan>
<p_rules> ::=          <p_rule> ( ";" <p_rule> )*
<p_rule> ::=           <atom> "<-" <query> "|" <plan>
<p_rule> ::=           "<-" <query>  "|" <plan>
<r_rules> ::=          <r_rule> ( ";" <r_rule> )*
<r_rule> ::=           <plan> "<-" <query> "|" <plan>
<co_plans> ::=         <co_plan> ( ";" <co_plan> )*
<co_plan> ::=          <p_rule> ( ";" <r_rule> )* |
                       "Send("<iv>,<iv>,<atom>");JavaSync("<i
                       dent>,<atom>,<var>")"
<literals> ::=         <literal> ( ";" <literal> )*
<literal> ::=          <atom> | "not("<atom>")"
<wff> ::=              literal> | <wff> "and" <wff> | <wff> "or"
                       <wff>
<query> ::=            wff> | "true"
<iv> ::=               ident> | <var>
```

Fig.1. The EBNF based 3APL syntax for individual agent

3. SEMANTIC

Multi-agent cooperation semantic for 3APL is an operational semantic. The operational semantics of 3APL is defined using transition system [1]. Transition system is a structure $<\Gamma,\rightarrow>$ where Γ is a set (of elements, called configurations) and $\rightarrow\subseteq\Gamma\times\Gamma$ is a binary relation (called the transition relation) and it corresponds to a single computation step. Here, Γ is the configuration of an agent. In the transition rules below, several cooperative operational semantics are given.

$$\frac{<\sigma,\gamma,s>\rightarrow<\sigma',\gamma',S'>}{<\sigma,\gamma,s.S>\rightarrow<\sigma',\gamma',S'.S>}$$

where $s\neq E$ (E is empty stack) TR1

$$\frac{\gamma\vDash\phi \quad \sigma\vDash\beta \quad \top\leftarrow\beta\,|\,\phi\in\theta \quad <\theta,\delta,\rho>\in R}{<\sigma,\gamma,E>\rightarrow<\sigma,\gamma,<\phi,\Omega,\delta,\rho>>}$$

where R is agent rules TR2

$$\frac{A_1\rightarrow A'_1\cdots\ A_i\rightarrow A'_i}{<A_1,\cdots,A_i,A_{i+1},\cdots,A_n,\xi>\rightarrow<A'_1,\cdots,A'_i,A_{i+1},\cdots,A_n,\xi'>}$$

$(1\leq i\leq n)$ TR3

$$\frac{\kappa\leftarrow\beta\,|\,\pi\in IR \quad \kappa\leftarrow\beta\,|\,\pi'\in IR' \quad \sigma\vDash\beta \quad \phi\vDash\kappa \quad \sigma\nvDash\kappa}{<\sigma,\gamma,<\phi,\Omega,\delta,\rho>>\rightarrow<\sigma,\gamma,<\phi,\pi\,\|\,\pi',\delta',\rho>>}$$

where $\delta'=\delta\setminus\{\kappa\leftarrow\beta\,|\,\pi\in IR,\kappa\leftarrow\beta\,|\,\pi'\in IR'\}$ TR4

$$\frac{\tau(a,\sigma)=\sigma'}{<\sigma,\gamma,<\phi,a;\pi,\delta,\rho>>\rightarrow<\sigma',\gamma',<\phi,\pi,\delta,\rho>>}$$

where $\gamma'=\gamma\setminus\{\kappa\,|\ \ \sigma'\vDash\kappa\}$ TR5

$$\frac{\pi_h\leftarrow\beta\,|\,\pi_b\in PR \quad \pi'_h\leftarrow\beta\,|\,\pi'_b\in PR' \quad \sigma\vDash\beta}{<\sigma,\gamma,<\phi,\pi_h\,\|\,\pi'_h;\pi,\delta,\rho>>\rightarrow<\sigma,\gamma,<\phi,\pi_b\,\|\,\pi'_b;\pi,\delta,\rho>>}$$

TR6

$$\frac{\sigma\vDash\phi}{<\sigma,\gamma,<\phi,\Omega,\delta,\rho>.S>\rightarrow<\sigma,\gamma',S>}$$

where $\gamma'=\gamma\setminus\phi$ TR7

TR1 specifies how a transition for a composed stack can be derived from a single stack. Here, we specify that s cannot be the empty stack[4]. In particular, when S is an empty stack, it identifies the transition above the line with the below.

TR2 specifies the transition form initial stack (empty stack) to non-empty stack. If agent believe β , ϕ is sub-goal of goal γ , $\top\leftarrow\beta\,|\,\phi$ is a element of the agent goal reasoning rules and $<\theta,\delta,\rho>$ is a element of the agent rules R , then the agent' initial stack is transformed to be $<\phi,\Omega,\delta,\rho>$.

TR3 is the rules for a transition between the configurations of multi-agent. The changes of single agent configuration bring about the multi-agent configurations transition via cooperation.

TR4 specifies how multi-agent generates cooperation plans. If the two plans generation rules are applied for the same goal κ , the cooperation plans $\pi\,\|\,\pi'$ becomes the plans of the resulting stack element. Further, agent applies plan π , others agent(s) apply plan π' . The reason for removing the two plans generation rules is that we do not want to the agent try the same plan generation rule twice, to achieve a certain goal.

TR5 describes the execution of basic action which is at the head of a plan if the function τ is defined for action a and the belief base σ in the configuration. Further, goals from the goal base which is reached, is removed.

TR6 points cooperation plans revision for multi-agent if the two plans revision rules are applied for the cooperation plan $\pi_h\,\|\,\pi'_h$ and β holds. Further, the plans after the revised plan are in reserve.

TR7 depicts the stack element is removed after agent' achievement or multi-agent successful cooperation. In particular, goal γ is empty and S is empty stack, after agent reaches all goals.

4. COOPERATIVE DELIBERATION CYCLE

The above we have discuss the syntax based on EBNF and cooperative operational semantic based on transition system for multi-agent cooperation. However, in order to run 3APL we also need an interpreter that determines the order in which rules are applied, when the actions should be performed, when belief updates should be made, etc[5]. The multi-agent cooperative deliberation cycle plays the interpreter role.

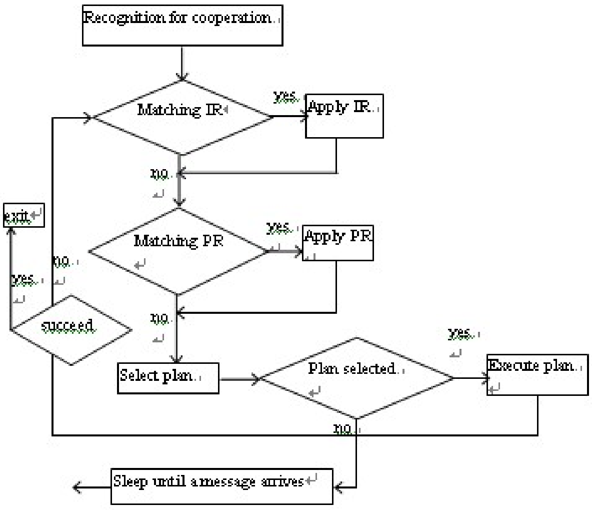


Fig.2. Deliberation Cycle for Multi-agent Cooperation

The cooperative deliberation cycle combines with the four stages for cooperative problem solving process[9] (see Fig.2.). Agent matches and applies IR after generating recognition for cooperation. Then agent revises cooperative plans via matching and applying PR. Finally, agents select and execute the best plan until agents achieve the goal. Otherwise, agent sleeps until a new message arrives to wake it up. If failing in the executing plan, agent would match IR once again according to the change of environment.

5. EXAMPLE

For illustration, we present a simple example of cooperation between two agents. Agent a and agent b have to cooperate to get the apple which is in a high position. The initial share environment beliefs are that agent a and agent b are near a chair, agent a is taller than b, the apple is in a high position, and that something is top if there is no other thing on top of it. Below, we define the two agents' initial belief base σ_a, σ_b and goal base γ_a, γ_b .

$$\sigma_a = \sigma_b = \{near(chair,a), near(chair,b), taller(a,b), high(apple), top(Y) :- not(on(X,Y))\}$$

$$\gamma_a = \gamma_b = \{get(apple)\}$$

$$\phi = near(chair, apple)$$

//according to TR2, the current goal ϕ is found

$$\pi \,||\, \pi' = movenear(chair, apple) \,||\, \Omega$$

// according to TR4, the cooperation plan is generated

$$on(block, chair)$$

// agent a and b discover there is a $block$ on top of the chair

$$\pi_h \,||\, \pi'_h \leftarrow \neg top(chair) \,|\, \{movenear(chair, apple) \,||\, move(block, floor)\}$$

// according to TR6, cooperation plan revision rule that agent a move the chair near the apple and agent b move block on the floor is applied.

$$\top \leftarrow top(chair) \,|\, \{movenear(chair, apple) \,||\, goto(apple)\}$$

//according to TR4, agent a move chair near the apple, and agent b go to the position where the apple is at.

$$\pi \,||\, \pi' = buttress(chair) \,||\, climb(chair)$$

// according to TR4

$$\pi_h \,||\, \pi'_h \leftarrow taller(a,b) \,|\, \{climb(chair) \,||\, buttress(chair)\}$$

// because agent a is taller than b, according to TR6, the cooperation plan revision rule is applied again to exchange their roles.

$$get(apple)$$

//finally, agent a get the apple and goal is achieved.

The above example researches on that how the cooperation performs between the two agents. Agents find the plans through cooperation plans generation rules and plans revision rules until that is a feasible plan to achieve the goal.

6. CONCLUSIONS AND FUTURE WORK

We have presented a new kind of multi-agent cooperative system based on 3APL. And we described the syntax and semantics of multi-agent that includes all the classical elements of the theory of agents (i.e. beliefs, goals, plans). In order to depict cooperation between agents, we bring about cooperation expression. Then we refer to a new kind of multi-agent cooperative operation semantics, which combines reasoning rules, stack, configuration and transition rules. Additionally, we study the multi-agent cooperative deliberation cycle processes deeply. Further, a cooperative example is analyzing for addressing how to generate and revise plans between agents.

With regard to future work, we will further research and design dynamic plans generation and revision rules in order to adapt the dynamic, unpredictable environment. On the other hand, multi-agent coordination logic and the finding of the best plans etc. will be our key problems in future work.

REFERENCES

[1] G. D. Plotkin,"A Structural Approach to Operational Semantics,"in *Technical Report DAIMI FN-19*, University of Aarhus, 1981.

[2] G. De Giacomo, Y. Lesperance, and H. J. Levesque, "ConGolog, a concurrent programming language based on the situation calculus,"in *Artificial Intelligence*, 121:109--169, 2000.

[3] K. V. Hindriks, F. S. de Boer, W. van der Hoek, and J.-J. Ch. Meyer,"agent programming in 3APL,"in I*nt. J. of Autonomous agents and Multi-agent Systems*,2(4),pp357–401,1999.

[4] M. B. van Riemsdijk, M. Dastani, J.-J. Ch. Meyer, Frank S. de Boer,"Goal-Oriented Modularity in agent Programming,"in *Proceedings of the Fifth International Joint Conference on Autonomous agents and Multiagent Systems (AAMAS'06)*,2006.ACM Press.

[5] M. Dastani, M. B. van Riemsdijk, F. Dignum, and J.-J. Ch. Meyer,"A programming language for cognitive agents: goal directed 3APL,"in *Programming multiagent systems, first int. workshop (ProMAS'03)*,LNAI,pp111–130. Springer, Berlin, 2004.

[6] M. B. van Riemsdijk, M. Dastani, and J.-J. Ch. Meyer,"Semantics of declarative goals in agent programming,"in *Proc. of AAMAS'05*,2005.

[7] M. Dastani, M. B. van Riemsdijk, and J.-J. Ch .Meyer,"Programming multi-agent systems in 3APL,"In R. H. Bordini, M. Dastani, J. Dix, and A. El Fallah Seghrouchni, editors,*Multi-agent Programming: Languages, Platforms and Applications*,Springer,Berlin, 2005.

[8] M. Dastani. 3APL platform: User Guide. http://www.cs.uu.nl/3apl/download.html.

[9] Wooldridge, M. and Jennings, N.R. "The Cooperative Problem Solving Process". Journal of Logic &Computation. 9(4):563-592.1999.

[10] Y. Shoham. agent-oriented programming. Artificial Intelligence, 60:51–92, 1993Bowman, B., Debray, S. K., and Peterson, L. L,"Reasoning about naming systems,"in *ACM Trans. Program. Lang. Syst.*, 15, 5 (Nov. 1993), pp795-825.

Shengfu Zheng, born in 1983, master candidate in Fuzhou University. His current interest includes AI, multi-agent systems.

Xinjian Lin, born in 1982, master candidate in Fuzhou University. His current interest includes AI, multi-agent systems.

Shanli Hu, born in 1944. He is a professor in Fuzhou University. His current interest includes AI, multi-agent systems.

Chaofeng Lin, born in 1983, master candidate in Fuzhou University. His current interest includes AI, multi-age systems.

Shexiong Su, born in 1982, master candidate in Fuzhou University. His current interest includes AI, multi-agent systems.

A QoS-aware Multicast Routing Algorithm Based on Ant Agents

Jiabao Hu
Department of Computer Science, Wuhan University of Technology
Wuhan, Hubei 430063, P.R. China
Email: hjbyp@sina.com

ABSTRACT

The existing schemes based on ant agents don’t take into account the impact of the imprecision of network state information on routing performance. In this paper, we design a novel ant agent-based routing (AAR) model with bandwidth and delay guarantees, in which we consider the influence of state information imprecision. Moreover, on the basis of the model, we propose a new QoS-aware multicast routing algorithm called QMRA. Extensive simulations show QMRA can achieve high routing success ratio, low average packet delay and fast convergence when the network state information is inaccurate.

Keywords: QoS, Multicast Routing, Ant Agent, Imprecision, State Information

1. INTRODUCTION

With the rapid development of communication networks (including circuit-switching networks and packet-switching networks), more intelligent techniques are needed to deal with the complexity problems of network routing such as QoS-aware routing, network dynamics and load balancing, and improve the reliability and scalability of network routing. The network routing problem is intrinsically a distributed, dynamic and multi-objective problem; as well as the routing decisions should only be made on the basis of the local state information by each network node. These features make it well adapted to solve the routing problems mentioned above by using ant agent approaches.

Ant agents, also called mobile agents based on swarm intelligence, can be used to make a distributed and adaptive network routing model, which is inspired by the ability of discovering the shortest path to a food source and the trail-laying/trail-following behaviors of a real ant colony. A reinforcement learning technique is applied to the routing model so that the ant agents can collaboratively find the optimal path between a source node and a destination node through the positive feedback mechanism [1].

At present, there exist several schemes to apply ant agents in network routing. In [2], an ant based control (ABC) system was designed to solve the load-balancing problem in circuit-switching networks by Schoonderwoed et al. It can efficiently cope with the dynamic aspects of network routing problem, but doesn’t work in asymmetric networks any more. Caro and Dorigo’s AntNet was originally designed for the network routing in packet-switching networks [3]. Unlike the ABC system, AntNet isn’t only restricted to the routing applications in symmetric networks. Nonetheless, the routing in AntNet is determined by means of a very complex procedure and involved in too many control parameters. In addition, Bonabeau et al. improved the performance of the ABC system in a reminiscent way of dynamic programming [4]. In [5], an agent-based routing system (ARS) with bandwidth and hop count constrains was proposed by K. Oida and M. Sekido as an extended version of AntNet. On the basis of AntNet, G. Lu and Z. Liu presented a multicast routing algorithm with delay and delay variation constraints as well [6].

Being similar to traditional networks, the state information in the networks based on ant agents is inherently imprecise due to the non-negligible link propagation delay and the dynamic variation of traffic load [7]. The imprecision has a noticeable impact on not only the speed of discovering the optimal path that satisfies certain QoS constrains but also the routing success ratio, because ant agents need to probe the links that can’t meet QoS requirements really. However, the ant agent-based schemes mentioned above do NOT take into account the imprecision of state information while founding their network routing models.

In this paper, we design a novel ant agent-based routing (AAR) model with bandwidth and delay guarantees, in which we consider the influence of state information imprecision on routing performance. And that, on the basis of the model, we propose a new QoS-aware multicast routing algorithm called QMRA, which works for packet-switching networks where the state information is imprecise. In our AAR scheme, an ant agent use the cost of a path instead of the trip time or age of an ant agent to determine the amount of pheromone to deposit, so that it has a simpler migration process and less control parameters. Furthermore, because of the usage of backward ants in the AAR model, it can also be applied in asymmetric networks efficiently.

2. ANT AGENT-BASED ROUTING MODEL

A communication network is usually represented as a weighted, connected graph G=(V,E), where V denotes the set of nodes and E denotes the set of full-duplex, directed communication links connecting the nodes. $|V|$ and $|E|$ denote the number of nodes and links in the network, respectively. Without loss of generality, only simple graphs are considered, in which there exists at most one link between a pair of ordered nodes.

A. Pheromone Table Structure

In order to apply the trail-laying and trail-following behaviors to packet-switching network routing, we replace the traditional routing tables in the network nodes by tables of probabilities, which are called “pheromone tables”, as the pheromone strengths are represented by these probabilities. Suppose that a node i ($i \in V$) with k neighbor nodes in a packet-switching network possesses a pheromone table $R_i = [r^i_{j,d}]_{|V|-1,k}$ with $|V|$-1 rows and k columns. Each row in the pheromone table corresponds to a destination node and each column corresponds to a neighbor node. The value $r^i_{j,d}$ expresses the probability of choosing the neighbor j as the next hop on the way to the destination d. In a pheromone table, all of the probability values in each row must conform to the constraint:

$$\sum_{j \in J(i)} r^i_{j,d} = 1 , \tag{1}$$

where $J(i)$ denotes the set of the neighbors of node i,and $d \in V$.

B. Imprecise State Information

As mentioned before, the local state information maintained at every node is dynamic and imprecise in a network based on ant agents. It includes: 1) the connectivity of network topology; 2) the link state, which consists of the residual bandwidth on a link, the propagation delay along the link, and the link cost; 3) the node state, which consists of the queuing delay at a node and the node cost, For the purpose of simplicity, we only take into account the imprecision of the residual bandwidth on a link. Such a simplification will not degrade the routing performance significantly, in that the residual bandwidth is most representative among the above state information [8].

In order to capture the imprecision of link residual bandwidth, $\forall(i,j)\in E$, we can use $\Delta b(i,j)$ to denote the estimated maximum change of the link residual bandwidth $b(i,j)$. That is, based on the recent state history, the actual residual bandwidth of link (*i,j*) is expected to be between $b(i,j)-\Delta b(i,j)$ and $b(i,j)+\Delta b(i,j)$ when a backward ant arrives (see the next subsection). We'll try to calculate $\Delta b(i,j)$ through the following way. Let $\Delta b_{old}(i,j)$ and $\Delta b_{new}(i,j)$ be the values of $\Delta b(i,j)$ before and after a backward ant arrives, respectively. Similarly, let $b_{old}(i,j)$ and $b_{new}(i,j)$ be the values of $b(i,j)$ before and after the backward ant arrives, respectively. The value of $\Delta b_{new}(i,j)$ is calculated as follows:

$$\begin{aligned}\Delta b_{new}(i,j) &= \alpha\times\Delta b_{old}(i,j)\\ &+(1-\alpha)\times\left|b_{new}(i,j)-b_{old}(i,j)\right|\end{aligned}. \qquad (2)$$

The factor $\alpha(\alpha<1)$ determines how fast the history information $\Delta b_{old}(i,j)$ is forgotten, and $(1-\alpha)$ determines how fast $\Delta b_{new}(i,j)$ converges to $\left|b_{new}(i,j)-b_{old}(i,j)\right|$.

We further assume that the residual bandwidth on link (*i,j*) are uniformly distributed in the range [$b_{new}(i,j)-\Delta b_{new}(i,j)$, $b_{new}(i,j)+\Delta b_{new}(i,j)$], and that, *B* denotes the minimum residual bandwidth required to the link. Therefore, we can calculate the probability which the link satisfies the bandwidth requirement. If $B>b_{new}(i,j)+\Delta b_{new}(i,j)$, then $P(b(i,j)\geq B)=0$, otherwise,

$$P(b(i,j)\geq B)=\frac{b_{new}(i,j)+\Delta b_{new}(i,j)-B}{2\Delta b_{new}(i,j)}. \qquad (3)$$

C. AAR Scheme

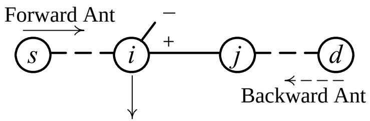
In our routing model, we devise two kinds of ant agents: forward ants and backward ants to discover adaptively the least-cost path that satisfies the bandwidth and delay requirements. The packet format of each ant agent is defined in Fig. 1, where each item between a pair of parentheses is the comment to the corresponding field name.

ant.ID (ant agent's identifer)
ant.node (visited node set)
ant.cost (visited link cost set)
ant.time (trip time record)
ant.bw (visited node's bandwidth set)
ant.br (bandwidth requirement)
ant.dr (delay requirement)
ant.s (source node)
ant.d (destination node)

Fig.1. The packet format of each ant agent

Forward ants are launched periodically from certain source to a destination selected randomly at an equal time interval. Suppose that a forward ant travels from the source *s* to the destination *d* along path (*s,…,i,j,…,d*), which is illustrated in Fig. 2. When the forward ant arrives at node *i*, it records the visited nodes and the cost of path (*s,…,i*), and calculates the accumulated trip time along the path. If $ant.time>ant.dr$, the forward ant dies, otherwise it starts the following operations.

Assume that node *j* is chosen as the next node to the destination by the forward ant, in that it has the maximum probability value in the pheromone table maintained at node *i* and $j\notin ant.node$. The forward ant detects the current residual bandwidth of link (*i,j*) as $b_{old}(i,j)$, collects its old residual bandwidth value that is detected by the previous backward ant and cached at node *i*, and calculates the value of $\Delta b_{old}(i,j)$. Subsequently, $b_{old}(i,j)$ and $\Delta b_{old}(i,j)$ are put into the stack of field *ant.bw*.



Destination \ Neighbor	*s*	*j*	…
s			
j			
d		+	−

Fig.2. The operation process of ant agents

When the destination *d* is reached, the forward ant generates a backward ant, dumps all of its memory to the backward ant and dies. Then, the backward ant moves from the destination *d* to the source *s* along the same path as that of its corresponding forward ant, but in the reverse direction. Arriving at node *i* from node *j*, the backward ant detects the current residual bandwidth of link (*i,j*) as $b_{new}(i,j)$, and evaluates the values of $\Delta b_{new}(i,j)$ and $P(b(i,j)\geq ant.br)$ according to (2) and (3) respectively. If $P(b(i,j)\geq ant.br)=0$, the backward ant dies, otherwise it updates the entry corresponding to the destination *d* in the pheromone table at node *i* in terms of the following equations:

$$r^{i}_{j,d}=\frac{r^{i}_{j,d}+\delta r}{1+\delta r},\ r^{i}_{k,d}=\frac{r^{i}_{k,d}}{1+\delta r}. \qquad (4)$$

In the equations, $k\in J(i)$ and $k\neq j$, δr is the reinforcement parameter of pheromone.

$$\delta r=\frac{a}{c(i,j)}+b\cdot P(b(i,j)\geq ant.br)\text{'} \qquad (5)$$

where *a* and *b* are the system control parameters, $c(i,j)$ denotes the cost of link (*i*,*j*). That is, the probability of node *i*'s neighbor *j* is increased while the probabilities of the other neighbors are decreased. Furthermore, the increase is in proportion to the reciprocal of $c(i,j)$ and the value of $P(b(i,j) \geq ant.br)$. Similarly, the backward ant carries out the above scheme at each intermediate node on its trip.

3. QOS-AWARE MULTICAST ROUTING ALGORITHM

Although the multicast routing problem with two or more additive metric constraints is a NP-complete problem [9], we can efficiently solve it by applying the preceding AAR model. Moreover, the algorithm based on ant agents has the better performances on network dynamics, load balancing and reliability than the traditional heuristic methods.

A. Preliminaries

Suppose that $s \in V$ is the source node of a multicast tree, and $M \subseteq \{V-\{s\}\}$ is a set of destination nodes of the multicast tree. Let R^+ be the set of positive real numbers. For any link $e \in E$, we define the link state information: the bandwidth function *bw*(*e*): $E \rightarrow R^+$, the delay function *delay*(*e*): $E \rightarrow R^+$, and the cost function *cost*(*e*): $E \rightarrow R^+$. Similarly, for any node $n \in V$, we define the node state information: the delay function *delay*(*n*): $V \rightarrow R^+$, and the cost function *cost*(*n*): $V \rightarrow R^+$. We use *T*(*s*,*M*) to denote a multicast tree that has the following relations:

$$
\begin{aligned}
& bw(p(s,d)) = \min\{bw(e), e \in p(s,d)\} \quad , \qquad (6)\\
& delay(p(s,d)) = \sum_{e \in p(s,d)} delay(e) + \sum_{e \in p(s,d)} delay(n)\\
& \cos t(T(s,M)) = \sum_{e \in T(s,M)} cost(e) + \sum_{e \in T(s,M)} cost(n)
\end{aligned}
$$

where *p*(*s*,*d*) denotes the path from the source *s* to the destination *d* of *T*(*s*,*M*). Let *B* and *D* be the minimum residual bandwidth and the maximum end-to-end delay required to *p*(*s*,*d*), respectively. The problem of the multicast routing with bandwidth and delay constraints can be represented as follows:

$$
\begin{aligned}
& bw(p(s,d)) \geq B \wedge delay(p(s,d)) \leq D\\
& \wedge \cos t(T(s,M)) = \min[...] \quad . \qquad (7)
\end{aligned}
$$

B. Algorithm Description

When a node receives a call request for opening a multicast session with the destination set *M*, the bandwidth requirement *B* and the delay requirement *D*, it becomes the source *s* of the multicast session. The pheromone table of each node in the network is initialized using an initial value, which is determined by the current status of bandwidth distribution.

1) We select a destination d_1 at random in *M*, and set the number of forward ants to *L*. Before a forward ant being launched from the source *s*, some of fields in it are initialized: *ant.node*={*s*}, *ant.cost*={0}, *ant.time*=0, *ant.br*=*B*, *ant.dr*=*D*, *ant.s*=*s*, *ant.d*=d_1. The forward ant starts from the source to the next node that has the maximum probability value in the pheromone table maintained at the source. If there's more than one node with the maximum probability value, we select randomly one from them as the next hop to the destination. After the next node being determined, field *ant.bw* in the forward ant is initialized in light of what we narrated in section II.
2) A forward ant moves from one node to another, one by one, as shown in Fig. 3 in which the network topology is generated by Waxman's algorithm [10]. The forward ant implements our AAR scheme at each intermediate node during the movement towards the destination d_1. As mentioned in section II, when the forward ant arrives at node d_1, the backward ant is generated and travels towards the source *s*, updating the entry corresponding to node d_1 in the pheromone table at each intermediate node along the path in terms of (2), (3), (4) and (5).
3) After a time interval *D* (i.e. the end-to-end delay constraint), the second forward ant is launched from the source *s*. And then, step 1 and 2 are repeated until all of *L* forward ants with the same destination d_1 are launched completely. Eventually, the optimal-cost path *p*(*s*,d_1) satisfying the bandwidth and delay requirements is established, and all the nodes along the path are designated as set M_1.
4) Similar to step 1, we select randomly a destination d_2 in set *M*-{d_1}, and set the number of forward ants to *L* too. The fields in the forward ants are initialized: *ant.node*={*s*,M_1} ($s \in M_1$, but it doesn't affect the operation process), *ant.cost*={0}, *ant.time*=0, *ant.br*=*B*, *ant.dr*=*D*, *ant.s*=*s*, *ant.d*=d_2. A forward ant chooses node *u* as its next hop to the destination d_2 by comparing the probability values in the pheromone table maintained at the source *s*. If $u \in ant.node$, the similar process continues until the first node *v* that is subject to $v \notin ant.node$ comes forth as shown in Fig. 3. When $v \notin ant.node$, the forward ant reinitializes its fields as follows: *ant.node*={*v*,M_1}, *ant.cost*={0}, *ant.time*=0, *ant.br*=*B*, *ant.dr*=*D*, *ant.s*=*v*, *ant.d*=d_2. From now on, the process of discovering the feasible and optimal path between node *v* and the destination d_2 corresponds with the above steps.
5) If *M*-{d_1,d_2,…,d_k}=*Φ*, the multicast session is established, otherwise step 4 is repeated.

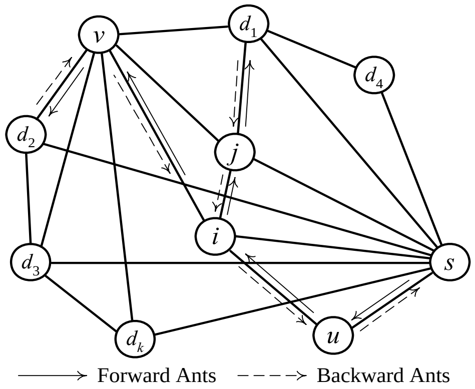


Fig.3. Network illustration

Seen from the above process, our proposed algorithm, QMRA is also an algorithm based on on-demand routing, which can decrease the quantity of ant agents needed in the case of ensuring the convergence of the algorithm.

4. SIMULATION

We implement our proposed algorithm by modifying and developing the Antnet model in OMNeT++, and evaluate its routing performance via two measures: routing success ratio and average packet delay.

$$Rs = \frac{Ns}{Nr}, Dp = \frac{Na}{Ti}, \qquad (8)$$

where *Rs* and *Dp* denote routing success ratio and average packet delay, respectively. In the equations, *Ns* is the number of the requests accepted, and *Nr* is the total number of call requests; *Na* is the sum of ant agent's number, and *Ti* is a given time interval. For the sake of comparison, another ant agent-based algorithm with QoS guarantees, ARS [5] is simulated as well. Each of the experimental results is the average value obtained by simulating it time after time.

A. Experimental Environment

The network topology with 50 nodes used in our experiments is generated by Waxman's algorithm [10], in which the bandwidth and delay of each link are uniformly distributed in the range of [10Mbps, 50Mbps] and [0.5ms, 2.5ms] respectively. For the purpose of simplicity, the cost of each link is configured to one unit.

In the experiments, multicast sessions are activated following a negative exponential distribution for the inter-arrival times, of which the mean value is fixed to 15sec. The source node and its destination nodes of each multicast session are generated at random. In order to simulate real situations, each group size (i.e. the number of destination nodes) is always less than 10% of the total nodes. The number of ant agents needed in each multicast session (i.e. parameter *L*) is uniformly distributed in the range of [50, 100]. The size of each ant agent is fixed to 256Bytes. System control parameters α, *a* and *b* are configured to 0.35, 0.08 and 0.005 respectively.

In addition, we import another performance metric: imprecision rate λ, which specifies the largest percentage difference of the actual residual bandwidth of certain link from its predicted residual bandwidth.

$$\lambda = \sup\ remum\ \{\frac{|b_{act}(i,j) - b_{new}(i,j)|}{b_{new}(i,j)}\}, \qquad (9)$$

where $b_{act}(i,j)$ is the actual residual bandwidth value of link (*i,j*) at the time of routing, and $b_{new}(i,j)$ is its residual bandwidth value used to compute routing. And that, $b_{act}(i,j)$ is uniformly distributed in $[(1-\lambda)b_{new}(i,j), (1+\lambda)b_{new}(i,j)]$.

The simulation parameters used in ARS are deployed as depicted in [5].

B. Experimental Results

Fig. 4 compares the routing success ratios of QMRA and ARS, which is a function of bandwidth requirement *B*, end-to-end delay requirement *D*, and imprecision rate λ. As expected, from this figure it can be seen that our algorithm performs much better than ARS as the imprecision rate grows higher, though there's an exception while λ <15%. QMRA still has a high success ratio even when the imprecision rate is as high as 50%.

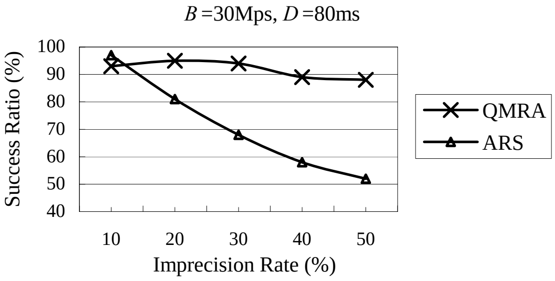


Fig.4. Routing success ratio

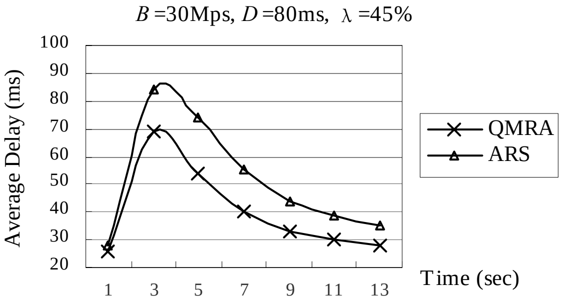


Fig.5. Average packet delay

The variation of average packet delay with respect to simulation time is able to reflect the convergence speed of the algorithms based on the trail-laying and trail-following theory. Seen from Fig. 5, our algorithm has a lower average delay, and converges faster than ARS when the imprecision rate is high. The reason is that QMRA based on a prediction mechanism can more accurately discover the path satisfying QoS constraints in the ant agent-based networks with imprecise state information, in which the optimization process is usually slow.

5. CONCLUSIONS

In this paper, we design a novel ant agent-based routing model with bandwidth and delay guarantees. Furthermore, on the basis of the model, we propose a new QoS-aware multicast routing algorithm working for packet-switching networks with inaccurate state information. Our simulations show the proposal can achieve high routing success ratio, low average packet delay and rapid convergence when the network state information is imprecise.

Our ongoing work includes: 1) the theoretical analysis and verification about the complexity and convergence of the algorithm based on ant agents; 2) a further investigation on how to accelerate the convergence of the algorithm and decrease the amount of ant agents needed under the precondition of convergence.

REFERENCES

[1] E. Bonabeau, M. Dorigo, and G. Theraulaz, "Inspiration for optimization from social insect behavior," *Nature,* vol. 406, pp. 39-42, July 2000.

[2] G.D. Caro and M. Dorigo, "AntNet: Distributed stigmergetic control for communications networks," *Journal of Artificial Intelligence Research,* vol. 9, 1998, pp. 317-365.

[3] E. Bonabeau, F. Henaux, S. Guerin, D. Snyers, P. Kuntz, and G. Theraulaz, "Routing in telecommunications networks with "smart" ant-like agents," in *Proc. of 2nd International Workshop on Intelligence Agents for Telecommunications Applications,* Paris, July 1998.

[4] K. Oida and M. Sekido, "An agent-based routing system for QoS guarantees," in *Proc. of IEEE International Conference on Systems, Man, and Cybernetics,* vol. 3, pp. 833-838, Oct. 1999.

[5] G. Lu and Z. Liu, "Multicast routing based on ant-algorithm with delay and delay variation constraints," in *Proc. of IEEE Asia-Pacific Conference on Circuits and Systems,* pp. 243-246, Dec. 2000.

[6] R.A. Guerin and A. Orda, “QoS routing in networks with inaccurate information: Theory and algorithms,” *IEEE/ACM Transactions on Networking,* vol. 7, no. 3, pp. 350-364, June 1999.

Integrated Supply Chain Management Based on Multi-agent

Guangchao Wu, Shu Yu
School of Mathematical Sciences, South China University of Technology
Guangzhou 510640, China
Email: wuguangchao@tom.com, yushu_scut@126.com

ABSTRACT

According to some existing problems in traditional supply chain management and incorporating the characteristics of multi-agent such as pro-activeness, reactivity and autonomous, this paper proposes an integrated supply chain management system based on multi-agent, MA-ISCM, and presents its architecture. MA-ISCM also introduces QoS management and defines the corresponding QoS model to optimize the integrated supply chain. Finally, the paper points out the disadvantages of current agent communication languages and proposes the XML-based agent communication language. The structure of communication layer and framework are also demonstrated.

Keywords: Integrated Supply Chain Management, Multi Agent, Knowledge Query and Manipulation Language, Agent Communication Language, Quality of Service, Extended Markup Language

1. INTRODUCTION

With the globalization of economy and rapid development of high technology, the requirements of agility and intelligence by supply chain become more urgent and customers' demands on products are more complicated, diversified and individualized. The competition among enterprises turns to the competition among supply chains. Traditional supply chain management (SCM) can't meet the requirements of enterprises' development any more and the integrated supply chain management (ISCM) offers a very good resolution.

Integrated supply chain (ISC) means that all members in supply chain form into a virtual organization based on common goals and each member inside the organization cooperates with others in the fields of information, funds and materials in order to optimize the performance. While ISCM refers to the management of the whole ISC, which means to plan, coordinate and control the business flow, material flow, information flow and fund flow among suppliers, manufacturers, distributors, customers and consumers. As a result, the ISC can become a seamless process and fulfill the business goals effectively. However, the highly dynamic ability, retractility and flexibility of ISCM pose a lot of challenges to computer and network technologies. As an important and successful breakthrough in the field of distributed artificial intelligence, the autonomous, reactivity, social ability and pro-activeness characteristics of agent make it an effective way to tackle these challenges.

In this paper, we first analyzes the current researches on supply chain management and points out some existing problems. Then, the paper proposes the architecture of ISCM based on multi-agent, MA-ISCM, and discusses how to introduce quality of service (QoS) management in supply chain. In addition, the paper also compares two kinds of agent communication languages, KQML (Knowledge Query Markup Language) and FIPA ACL (Agent Communication Language), and proposes the XML-based ACL. The conclu- sions and future research directions are presented at last.

2. RELATIVE RESEARCH

Traditional SCM researches mainly focus on the solution of individual problems inside enterprises and seldom consider the management problems of supply chain as a whole, which is just the key to set up competition advantages against others. Reference [1] proposed a SCM with hierarchical structure to improve the adaptability and automaticity of supply chain, but it couldn't solve the highly dynamic problem of ISC. Reference [2] researched on the purchase management in ISCM and presented its theory model and running model, but the authors didn't analyze how to implement the models effectively and the multi-agent technology proposed by this paper would be a very good supplement. The authors in [3] used multi-agent to integrate supply chain intelligently and discussed agents' applications from three layers (information layer, knowledge layer and decision layer) in SCM, but they didn't demonstrated how to build the architecture of ISCM. Reference [4] considered the supply chain as a whole and showed the application architecture and software implementation model of ISCM, but the authors didn't explain the specific technologies.

In addition, most of the current researches only analyze the mutual cooperation between nodes in neighbor supply chain or use multi-agent to integrate supply chain, but they don't mention how to manage these entities and optimize the resources based on their QoS, which we think is a very important point for ISCM in business environment.

3. ISCM BASED ON MULIT-AGENT

3.1 Definition of Agent

Presently, there are many different definitions about agent and here we use the following one, "Agent is the entity that has a life cycle and is designed to finish some tasks and can act independently in certain environment. Four of the most fundamental properties of agent are autonomous, reactivity, social ability and pro-activeness。

3.2 Agent Group

Agent is the smallest unit for ISCM, while agent group is the basic component part. An agent group is composed of several agents who have very close relationship or similar functions. The group members usually have some common characteristics and these characteristics can be used by the group manger to manage the group. Additionally, agent group also provides a convenient way to the cooperation among the agents inside the group and using the group as a unit can allow them to study and optimize the tasks more effectively.

In ISCM system, manufacturer, supplier and distributor can be regarded as an agent group respectively, and each agent group includes some agent entities with different functions. For example, supplier agent group can involve order agent, inventory agent and material plan agent. So, different entities in ISCM can be divided into some agent groups according to their relationships. These groups have different problem solving ability and they communicate and coordinate through some established protocols, which can form the whole ISCM system into an integrated part. As a result, the performance can be improved and the problems that a single agent can't solve are able to be tackled easily. An example of agent groups in ISCM is shown as follows:

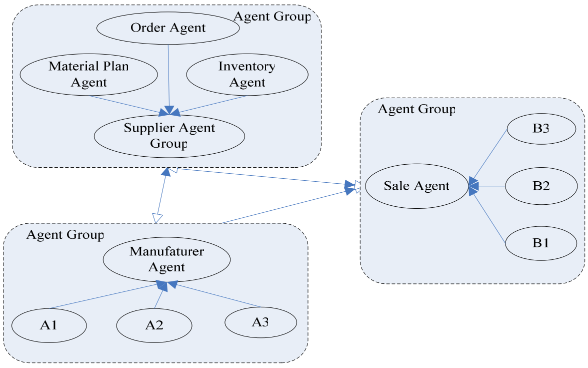


Fig.1. Agent Group Examples in ISCM

3.3 ISCM Architecture Based on Multi-agent

In the following, we propose a multi-agent-based ISCM (MA-ISCM). MA-ISCM regards order management, purchase management, inventory management, cost control and sale management as the basic functional units and organize them in the form of agent group, which can finally become an ISCM system with high flexibility, autonomy and retractility. At the same time, in order to configure and manage the resources in supply chain more effectively, we introduce QoS management in MA-ISCM. The architecture of MA-ISCM is presented in Fig.2.

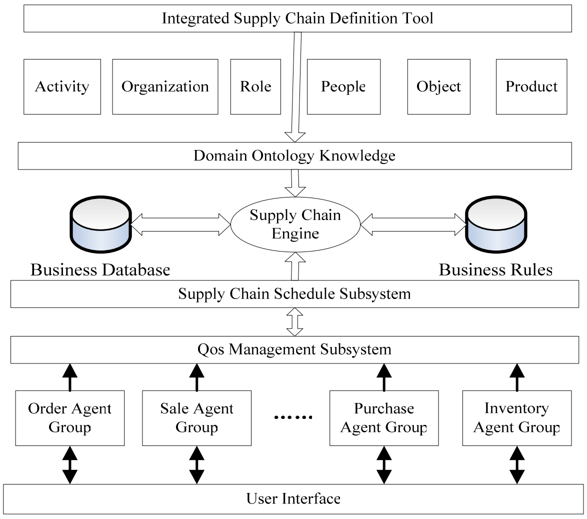


Fig.2. ISCM Architecture Based on Multi-agent

MA-ISCM mainly includes the following parts:

(1) ISCM Definition Tool

It defines each basic entity in the business activities of ISCM, such as product, people, role and activity.

(2) Domain Ontology Knowledge

It describes each entity inside enterprises and forms them as domain ontology knowledge. As the resource data and objects in ISCM are very complicated and have different representation means and semantic meanings, using ontology to represent agents' knowledge and semantic meanings can make their communication more convenient and accurate and the problems of data exchange and cooperation can also be tackled.

(3) Data Engine

It maps the business data, ontology knowledge and business rules to the multi-agent system.

(4) Schedule Subsystem

It takes charge of the coordination and dispatching of each agent group to improve the agility of ISCM.

(5) QoS Management Subsystem

It manages the QoS of each agent group in supply chain and chooses the one that satisfies the QoS restriction conditions according to the requirement of schedule subsystem.

3.4 QoS Management in MA-ISCM

With the rapid increasing number of entities in supply chain, it is inevitable that many agent groups who provide the same or similar functions appear and these agent groups might have different QoS. So, during the optimization process of MA-ISCM, it is necessary to select some best agent groups according to the QoS requirements. For example, we might require the supplier to provide products with lower cost besides meeting the requirements of dependability and reputation. As a result, the MA-ISCM needs to manage the QoS effectively and selects the agent group that meets the restriction conditions dynamically to optimize the resource and improve the agility.

QoS usually includes some non-functional properties, such as cost, execution time, reputation and dependability. In this paper, we define a basic QoS model, which is shown in Fig.3. It can be extended at any time when needed.

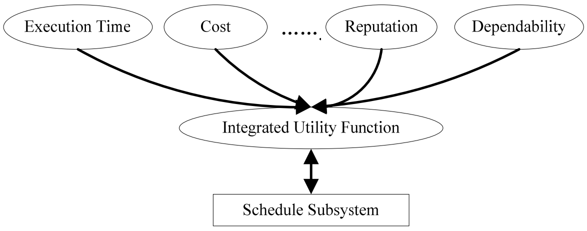


Fig.3. QoS Model

The QoS model describe agent groups' ability to meet the requirements of supply chain and it is a set of a few service quality standards, each of which demonstrates some non-functional properties and has its own computation and evaluation means. The QoS model is composed of four performance parameters: cost, execution time, reputation and dependability. Each parameter has its own weight when being composed and they exchange information with schedule subsystem to direct the task schedule of supply chain. All initial or modified performance values related with a task are input through a visual interface. After the task is finished, all performance values are shown, which can provide references for further processes. The definition of each QoS standard is shown as follows (A denotes a certain agent group or entity):

(1) Execution Time (denoted as T): refers to the time from sending out task request to finishing it. It is made up of two parts: task processing time (TP_A) and service delay time (TD_A), which means $T_A=TP_A+TD_{A.}$

(2) Cost (denoted as C): means the A's cost of finishing the task. Supposed N denotes the number of resource units used by the task, UT denotes the time of using the

resources and P denotes the price of the resources in a unit time. Then, under the condition of unified "currency" unit, the cost for this specific task is $C=N\times UT\times P$.

(3) Dependability (denoted as R): A's dependability, R_A, refers to the possibility of correctly providing service by A and it is the metric of A's reliability. Normally, R_A is computed as follows: $R_A=N_A/K$, in which N_A means the number of finishing the task successfully by A until now and K means the total number of executing the tasks by A until now.

(4) Satisfaction Degree (denoted as SD): A's satisfaction degree SD_A refers to users' subjective satisfaction degree about A and is usually expressed as a decimal number between 0 and 1. SD_A is computed through the ratings from users and calculated as follows: $SD_A=(SD_{A1}+SD_{A2}+\ldots+SD_{An})/n$, in which SD_{Ai} refers to the users' satisfaction degree about A at the time of i^{th} service.

From the above definitions, we can lean that A's QoS model is a quaternion $Q_A=(T_A, C_A, R_A, SD_A)$. Each quality standard has its own computation and evaluation means and the whole QoS model also has a global evaluation method like the one discussed in reference [5].

3.5 Agent Communication Language

To communicate and coordinate in MA-ISCM, agents need to rely on the agent communication protocol. For the different agent groups in ISCM, the communication languages used by them are not necessary the same because the agents locate in different enterprises.

Currently, the relative successful agent communication languages include KQML [6] (Knowledge Query and Manipulation Language) from DARPA and FIPA ACL [7] (FIPA Agent Communication Language). The basic concepts and principles of these two languages are similar, but their definitions of syntactical format and semantic explanation are different. The major differences between them are their semantic framework. But both of them have some disadvantages, such as difficult ACL message parsing, poor expansibility and bad cross-platform ability.

XML (Extended Markup Language) offers a feasible way to solve these problems and its advantages, including simplicity, flexibility and good expansibility, can meet the requirements of ACL very well.

When agents communicate with each other through ACL, the messages among them can be divided into two relative independent layers: (1) Communication Meta Layer; (2) Communication Content Layer. As a result, no matter KQML or FIPA-ACL is chosen as the communication language, XML can always be used to wrap the ACL meta message or describe the contents and convert the ACL document into XML document before sending it.

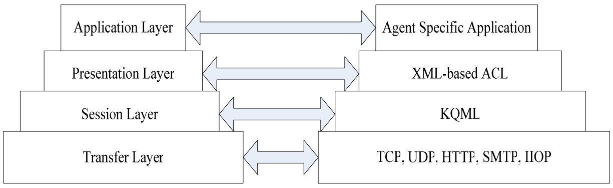


Fig.4. Structure of Agent Communication Layer

Using XML to code the message contents in ACL document, we can get the unified description and avoid the communication problems caused by using different communication languages. The communication framework based on XML is depicted in Fig.4.

As the agents in the same group usually rely on the same ACL, so they can use a common XML wrap and parsing module to allow conversion between ACL document and XML document. A concrete communication process can be depicted as follows: (1) A inserts the requests into the content layer of ACL and creates the corresponding ACL document; (2) XML wrap and parsing module convert ACL document into XML document; (3) the XML document is sent to B by the schedule of control and execution module.

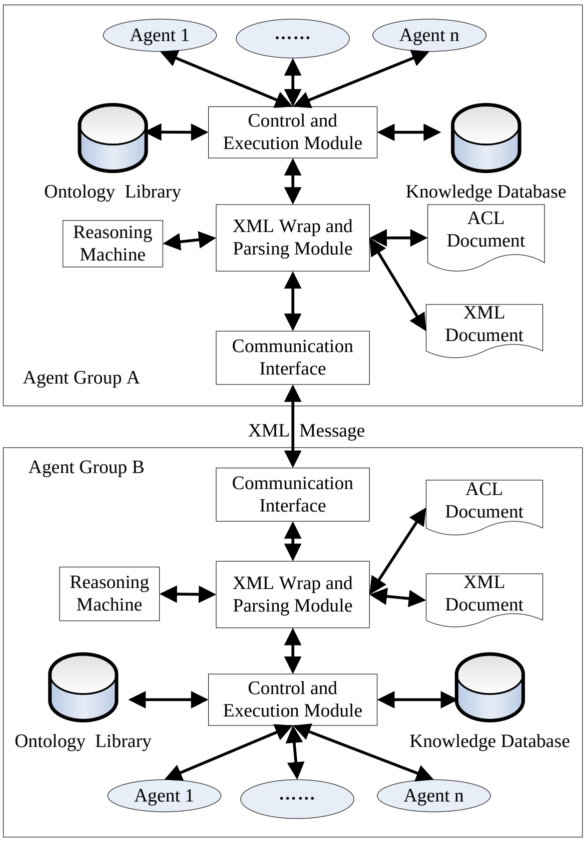


Fig.5. XML-based Communication Framework of Agent

After B receives the document, it takes the following actions: (1) parse the XML document and create the correspondent ACL document through XML wrap and parsing module; (3) use reasoning machine to produce the ACL results based on ontology library and knowledge database; (3) convert the results into XML document again and send it back to A.

4. CONCLUSIONS

Based on the characteristics of multi-agent, this paper first uses agent group to achieve the modular design and management of ISCM and proposes an ISCM system based on multi-agent, MA-ISCM, and present its architecture. Secondly, the paper introduces QoS management into MA-ISCM system and defines the QoS model to optimize the MA-ISCM. Finally, the paper also tries to propose XML-based agent communication language and gives the relevant communication layer structure and communication framework.

Our future researches will focus on the following three directions:

(1) Try to improve the QoS model in ISCM and figure out the global QoS computation method;
(2) Deeply analyze the interaction and communication protocols among multi-agent group;
(3) Apply multi-agent technology into the actual ISCM system and confirm its validity and feasibility.

REFERENCES

[1] Zheng wen-en, Sun Yao, Lu Ming-hua, “Real-time Multi-agent Structure and Implement of Group Intelligent Decision System,” *Journal of System Simulation*, 2003, 12(15): 1718-1720.

[2] Bao Jun-je, Tang Min, “A Design for the Information System by Supply Chain Management,” *Journal of Chongqing Normal University (Natural Science Edition)*, 2004, 21(1):38-40.

[3] Sun Xiu-jie, Wu Ze-shu, Chen Ting-bin, “Intelligent Integration fro Supply Chain Enterprise Based on Multi-Agent,” *Logistics Sci-Tech*, 2007(2): 70-72.

[4] Wang Chun-xi, Zha Jian-zhong, “An Integrated System Structure for Supply Chain Application,” *Industrial Engineering Journal*, 2006, 9(3):59-64.

[5] Zhang Cheng-wen, Su Sen, Chen Jun-liang, “Genetic Algorithm on Web Services Selection Supporting QoS,” *Chinese Journal of Computers*, 2006, 29(7):1029-1037.

[6] Finin T et a, “Specification of the KQML Agent Communication Language,” *DARPA knowledge sharing initiative external interfaces working group*, 1993.

[7] FIPA, “Communicative Act Library Specification,” http://www.fipa.org/specs.

A New Car-Following Model Based on Multi-Agent System*

Chaozhong Wu, Xinping Yan, Xiaofeng Ma
Intelligent Transport System Research Center, Wuhan University of Technology
Wuhan, Hubei, 430063, China
Email: chaozhongwu@gmail.com

ABSTRACT

In this study, we developed a car-following model based on an intelligent agent. The intelligent agent was first introduced into a micro-traffic flow system. Each vehicle was expressed as an agent and many vehicles in traffic flow constitute a multi-agent system. Individual differences in drivers' reaction time and character were considered at the same time. The theory of probability was applied to reflect the distribution of drivers' stochastic characteristic dispositions, and function of probability density was used to express the reaction time of drivers. Each vehicle expressed its intelligence through following the leading vehicle by its own reaction time and character. A case was developed to validate the proposed model; the results showed that the model can reflect the drivers' reaction time and character.

Keywords: Car-Following Model, Multi-Agent System, Traffic Flow

1. IMPORTANT INFORMATION

Traffic flow models presently attract a rapidly growing community of traffic engineers. Recent research focuses on micro-simulation models[1],[2]. Car-following model is one of important micro traffic model, which based on the idea that each driver controls a car under the stimuli from the preceding car, which can be expressed by the function of headway distance or the relative velocity of two successive cars[3]. Car-following models examine the following status of vehicles traveling in the same lane. They usually express a convoy in terms of stable, accelerating, interference, and spread. Some of the many car-following models include the General Motor, Action Point, and Collision Avoidance models.

For decades, researchers have paid considerable attention to traffic flow problems[4][5]. Many studies from differing perspectives have examined various traffic phenomena[6]. These include methods based on many-particle and statistical physics, nonlinear dynamics, as well as complex systems and network and control theory[7].The car-following model is a typical microscopic traffic continuous model that simulates the movement of vehicles following each other in a single lane without any passing. It is based on the assumption that each driver will react in some specific way to a stimulus from the preceding vehicle[8]. These models can be used to simulate second-by-second velocity and acceleration. One typical improvement is the optimal velocity model, which assumes that a vehicle's speed is dependent on the distance from the previous vehicle. It can simulate many features of actual traffic such as chaos, evolution of an obstruction, and stop-and-go traffic[6]. Helbing and Tilch[4]proposed the generalized force model to solve this problem; in this model, the relative speed of the front and following vehicle is considered. A full velocity difference (FVD) model was addressed both positive and negative differences in velocity; it can correctly predict the delay times of car motion and kinematic wave speeds at jam density, and can also produce results such as the formation of congestion based on an initially homogenous condition[8]. The statistical method was used to analyze traffic data in complex networks [9][10]. Some dynamic traffic models were proposed [11]12]. Martin Schonhof, et al. investigate how the equipment level influences the efficiency and velocity of information propagation by analytical calculation and by microscopic simulation of freeway traffic with a given percentage of vehicles equipped for inter-vehicle communication[13].

However, most of the previous studies of car-following model had serious drawbacks. In the optimal velocity model, velocity and acceleration can be calculated in real time, but it does not consider the relative velocities of the front vehicle and the following vehicle. The generalized force model solved the problem of overly rapid changes in acceleration. However, it did not consider the positive difference in velocity, so results of its simulations differed greatly from real traffic. In fact, when a front vehicle has a much greater velocity than the following vehicle, the following vehicle will not decelerate even if the distance between the two is less than required for safety. In other words, difference in velocity has an effect not only when a following vehicle is faster, but also when it is slower than the vehicle in front. Neither the optimal velocity model nor the generalized force model can address these situations. In general, traditional car-following models are based on the following assumptions: (1) All vehicles travel in the same lane, no on-ramps or off-ramps exist, and passing is not allowed; (2) Every driver operates his or her vehicle based only on the performance of the preceding vehicle;(3) Drivers' individuality is not considered; all vehicles have the same performance. Under these assumptions, traditional car-following models have the same character: all following vehicles have analogical acceleration or deceleration based on the lead vehicle, and time delay is the primary difference.

In fact, a vehicle's functional status is decided by many factors, such as velocity of the front vehicle, velocity of the following vehicle, distance between the two vehicles, road conditions, vehicles' performance, and drivers' individuality. Due to this complexity, many models ignore some or all of these factors. For example, a model might not consider drivers' individuality, and assume all vehicles have the same performance. To avoid this shortcoming, we developed a model that considers a driver's individuality based on multi-agent system.

Therefore, the objective of this study was to develop an car-following model based on multi-agent system for simulating the vehicles moving on the same lane. The key elements in the proposed model include (a) multi-agent system for vehicle platoon; (b) systematic analysis of the complex relationships between the leading vehicle and the following vehicle; (c) drivers' reaction time and character; and (d) application of the developed model in a numerical study.

This paper is organized as follows. The next section will

*This study is sponsored by National Basic Research Program of China（2005CB724205, 2006CB705505）and the State Scholarship Fund (N0.2004842267).

present a car-following model considering reaction time and character of drivers at the same time. A case is developed to validate the proposed model in section 3. Conclusions and suggestions for future studies are summarized in Section 4.

2. METHODOLOGY

In traffic flow, vehicles are controlled by drivers. In other word, characteristics of drivers can express the characteristics of vehicles. Drivers are intelligent and they can make decisions according to the information of leading vehicle and their own property. Agent is capable of representing the characteristics of driver. In this study, the character and reaction time are used to express the property of drivers. Fig. 1 shows the architecture of vehicle agent.

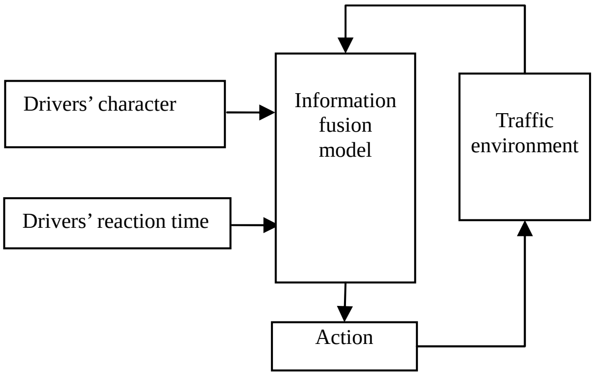


Fig.1. The architecture of vehicle agent

There are many vehicles in traffic flow, which can be expressed by Multi-Agent Systems (MAS). The development of the MAS has progressed since the 1990s as a subfield of distributed artificial intelligence[14].

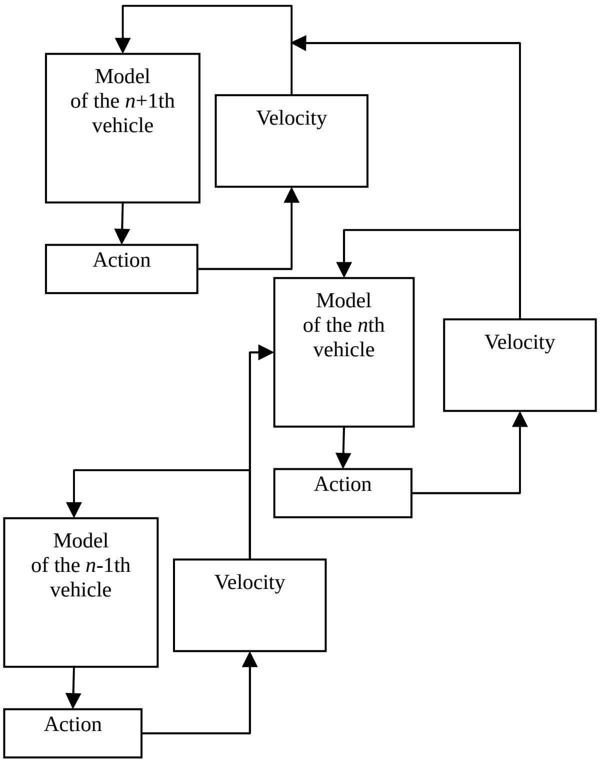


Fig.2. The architecture of multi- agent car-following vehicles

A platoon of vehicles can be designed into the MAS. Every agent perceives the velocity and acceleration information of the vehicle to the front, deals with the information based on the model, and implements actions. We assumed that the *n*+1th vehicle following the *n*th vehicle, and the *n*th vehicle following the *n*-1th vehicle. The architecture of multi- agent car-following vehicles is showed in Fig.2.

In this study, drivers' property were developed into the model. We assumed that the *n*th vehicle follows the n-1th vehicle and that the *n*+1th vehicle follows the *n*th vehicle.

The reaction time and character of drivers were introduced into the agent mod el. The statistical analysis showed that the distribution of drivers' reaction time is off-center Gaussian distribution[15]. The function of probability density is as follows:

$$f(T)=\frac{1}{\sqrt{2}\xi T}\exp[-(\tfrac{\ln(T)-\lambda}{\xi})^2] \qquad (1)$$

where $\xi^2=\ln(1+\frac{\sigma^2}{\mu^2})$, $\lambda=\ln(\frac{\mu}{\sqrt{1+\sigma^2/\mu^2}})$. The reaction time of drivers can be calculated by $f^{-1}(T)$.

We incorporated different kinds of drivers with different characteristics, as some drivers are reckless while others are careful. As they drive, they accelerate or decelerate differently; reckless drivers make sudden changes, whereas careful drivers make smoother and more controlled changes. We let $q(i)$, (*i*=1,2, …, *m*) denote the degree of acceleration, obeying the probability distribution. $P_c(i)$ (*i*=1,2, …, *m*) denotes the probability of all kinds of drivers. $P_c(i)$ (*i*=1,2, …, *m*) is subject to $\sum_{i=1}^{m}P_c(i)=1$.

Drivers' reaction time and characteristics were introduced into full velocity difference (FVD) model. The FVD model was proposed[8] as follow:

$$\frac{dv_n(t)}{dt}=k[V(\Delta x_n)-v_n(t)]+\lambda\Delta v_n \qquad (2)$$

where *x* is the vehicle position; *v* is the vehicle velocity; *t* is the time; *n* is the vehicle sequence; $\Delta x_n = x_{n-1}-x_n$, $\Delta v_n = v_{n-1}-v_n$; *i* is the type of driver.

Thus, the proposed model includes the drivers' reaction time and characteristics as follows:

$$\frac{dv_n(t)}{dt}=[k(V(\Delta x_n)-v_n(t-f^{-1}(T)))+\lambda\Delta v_n]\times q(i) \qquad (3a)$$

$$V(\Delta x)=v_1+v_2\tanh[C_1(\Delta x-l_c)-C_2] \qquad (3b)$$

where $q(i)$ is the character of the i -type driver. v_1, v_2, C_1, C_2 and l_c, are coefficients and $P_c(i)$, (*i*=1,2, …, *m*) is the probability of a driver's characteristic; $f^{-1}(T)$ is the inverse function of drivers' reaction time. Every vehicle expresses its intelligence through its own reaction time and character and by perceiving the leading vehicle. The simulated vehicles in this model function as a multi-agent system.

3. NUMERICAL SIMULATION AND DISCUSSION

As a case study, the starting process of vehicles was examined at an intersection in the city of Wuhan. When the traffic light changed to green, all ten cars started and moved forward. The average distance between the cars was 7.4*m*. The resulting optimal parameter values are k=0.85s^{-1}, v_1=6.75m/s, v_2=7.91 m/s, C_1=0.13m^{-1}, C_2=1.57m^{-1}, l_c=5m, m=3, $P_c(1)$ =0.3, $P_c(2)$ =0.4, and $P_c(3)$ =0.3. The simulation covered 30 seconds.

The FVD model is tested firstly. Fig.3 presents the results of the FVD simulation of the vehicle convoy’s starting process at the intersection.

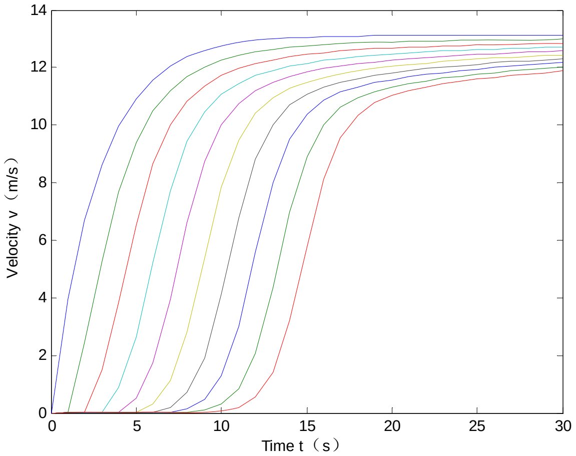


Fig.3. The simulation results of FVD model

Fig.3 presents that the velocity curve is very uniform; movement of following vehicles is identical to the lead vehicle. Acceleration and deceleration are expressed, while the drivers’ individuality is not expressed. This differs from reality, when every driver has individual characteristics such as personality, driving ability, physical status, psychological state, etc. Each factor has a different effect on vehicle functioning. The results shows that the drivers’ reaction time and character are not reflected in the FVD model.

In order to compare with the FVD model, we simulate the same scene using the model developed in this study with the same parameters and simulation time. The simulation results are showed in Fig.4

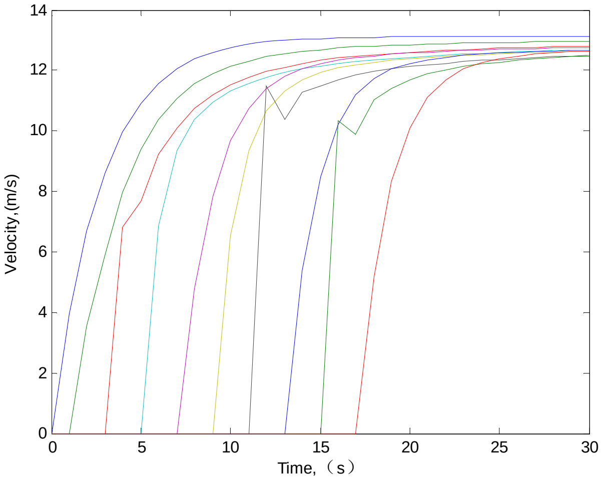


Fig.4. The simulation results of the proposed model

Fig.4 illustrates a non-uniform velocity curve; the functioning of following vehicles was not identical to the lead vehicle. All the following vehicles are delay in different degree, which determined by the reaction time. The characters of drivers are also reflected in Fig.4. At times, the velocity of following cars was greater than that of the lead car, which is closer to reality. The individuality of drivers was in accordance with the probability distribution. Every driver had a typical individuality of reckless, normal, or careful, with a probability of 0.3, 0.4, and 0.3, respectively. In the case study, the individuality types of ten drivers were generated according to the probability distribution. The first, second, fourth, and eighth drivers were normal. The third, seventh and ninth drivers were reckless. The fifth, sixth, eighth, and tenth drivers were careful. Fig.4 shows that the third, seventh, and ninth cars showed velocity undulations.

4. CONCLUSIONS

In this study, an intelligent agent system was designed for vehicles, which introducing drivers’ reaction time and characteristics into the FVD model. The proposed model improves upon the previous FVD model by providing the advantage of expressing drivers’ individualities (reaction time and character). Every vehicle expressed its intelligence through its own reaction time and character by perceiving the leading vehicle. The simulated vehicles in this model can be used as a simple intelligent agent system. A case study simulated the starting process of vehicles at an intersection, and the results indicated that the proposed model correctly exhibited drivers’ individuality and was more realistic.

This study is a new attempt and the proposed model can be used for analyzing various distributions of drivers with different individuality, which includes reaction time and character. Although the results suggest that the multi-agent approach is applicable to practical problems that involve a large number of vehicles in traffic system, the proposed model could be further enhanced through introduce other individuality (such as stochastic behaviors, traffic rules) into its framework.

REFERENCES

[1] Patrizi, Bagnerini et al., “On the role of source terms in continuum traffic flow models”, *Mathematical and Computer Modelling*, 44, 2006, pp. 917–930

[2] M. Krbalek, D. Helbing. “Determination of interaction potentials in freeway traffic from steady-state statistics”, *Physica A*, 333, 2004, pp.370－378.

[3] Zhao, X. and Gao, Z. "A new car-following model: full velocity and acceleration difference model", *European Physical Journal B* ,47(1), 2005, pp. 145-150.

[4] Helbing, D. and B. Tilch. "Generalized force model of traffic dynamics", *Physical Review E*, 58(1), 1998, pp. 133-138.

[5] Ou, Z. H., S. Q. Dai, *et al*. "Density waves in the full velocity difference model", *Journal of Physics a-Mathematical and General*, 39(6), 2006, pp. 1251-1263.

[6] Bando, M., K. Hasebe, *et al*. "Dynamical Model of Traffic Congestion and Numerical-Simulation." Physical Review E 51(2), 1995, pp. 1035-1042.

[7] Helbing, D., D. Armbruster, et al. "Information and material flows in complex networks", *Physica a-Statistical Mechanics and Its Applications*, 363(1),

2006, pp. XI-XVIII.

[8] Jiang, R., Q. S. Wu, *et al*. (2001). "Full velocity difference model for a car-following theory", *Physical Review E*, 6401(1), 2001, pp.1-4

[9] R. Albert, A.-L. Barabasi. “Statistical mechanics of complex networks”, Rev. Mod. Phys. 74, 2002, pp. 47 – 97.

[10] R.M. Colombo, “Hyperbolic phase transitions in traffic flow”, SIAM J. Appl. Math. 63 (2), 2002, pp. 708–721.

[11] C.F. Daganzo, “Requiem for high-order fluid approximations of traffic flow”, Trans. Res. 29B (4), 1995, pp. 277–287.

[12] R.M. Colombo, A. Corli, “Dynamic parameters identification in traffic flow modeling”, Discrete Contin. Dyn. Syst (Supplement Volume). 2005, pp.190–199.

[13] Martin Schonhof, Arne Kesting, et al., “Coupled vehicle and information flows: Message transport on a dynamic vehicle network”, Physica A , 363, 2006, pp.73 – 81.

[14] Liu, J.K. and L.J. Er, *Multi-agent Technique Applying Summarization Control and Decision*, 16(2), 2001, pp. 133-140.

[15] Wang, D.H. *Traffic flow theory*. China Communications Press, 2002, pp.29-30.

Chaozhong Wu is an Associate Professor of Intelligent Transport System Research Center, Wuhan University of Technology. He graduated from Wuhan University of Technology in 2002 and got Ph.D degree. He is Membership of National ITS Standardization committee, China and Associate Secretary-General of the Committee of Youth Scientific & Technological Workers for Transportation, China. He was a visiting scholar of University of Regina, Canada (2005~2006), His research interests are in traffic flow theory, intelligent transport system and transportation safety.

Xinping Yan is a full Professor and director of Intelligent Transport System Research Center, Wuhan University of Technology. He graduated from Xi’an Jiaotong University in 1997 and got Ph.D degree. He is editor of “Journal of Condition Monitoring and Diagnostic Engineering Management”, “Journal of Plant Engineering and Management”, and “Journal of Maritime Environment”. He is Membership of National ITS Standardization committee, China and chairman of the Committee of Youth Scientific & Technological Workers for Transportation, China. His research interests are in transportation safety, information fusion, and intelligent transport system.

Application of Petri Net in the Analysis of Agent Conversation Models

Weihong Yu
School of Economics and Management, Dalian Maritime University
Dalian, LiaoNing 116026, China
Email: yuwhlx@163.com

ABSTRACT

Petri Net models have emerged as a very promising performance modeling tool for systems that exhibit concurrency, synchronization, and randomness. A petri-net based approach for analyzing conversation models among agents is proposed in this paper. As an example, the application of petri net into search and rescue at sea decision supporting system is introduced. By using this method, coherency and coordination of agent conversations are verified.

Keywords: Petri Net, Agent, Agent Conversation Model

1. GENERAL DEFINITIONS ABOUT PETRI NET

Definition 1 Petri Net: A petri net is a four-tuple (P, T, IN, OUT) where P = {p1, p2, p3, ..., pn} is a set of places;T = {t1, t2, t3, ..., tm} is a set of transitions.
$P \cup T \neq F$, $P \cap T = \varnothing$.
IN: $(T \times P) \rightarrow N$ is an *input function* that defines directed arcs from places to transitions, and
OUT: $(T \times P) \rightarrow N$ is an *output function* that defines directed arcs from transitions to places.

Definition 2 Marking: A *marking* of a petri net with *n* places is an *(1 x* n) row vector. It associates with each place a certain number of tokens or marks represented by dots.

Definition 3 Transition Firing and Reachability: A transition ti of a petri net can only be fired if each of the input places of this transition contains at least IN (ti, pj) tokens. Firing of a transition ti involves withdrawing the number of tokens spent by the input function, from each of the input places of transition ti and adding them to each of the output places of transition ti. An enabled transition ti can fire at any time. When a transition ti enabled in a marking M fires, a new marking M′ is reached according to equation

$$M¢ (pj) = M (pj) + OUT (ti, pj) - IN (ti, pj)\ \forall\ pi \in P$$

We say marking M′ is reachable from M through firing of transition ti.

Definition 4 Incidence Matrix: IN: $T \times P \rightarrow \{0, k\}$ is the input incidence matrix;
OUT: $T \times P \rightarrow \{0, k\}$ is the output incidence matrix.
Here, $k \geq 1$ is an integer and denotes the weight of the associated arc. IN (ti, pj) is the weight of the arc pj → ti. OUT (ti, pj) is the weight of the arc ti → pj. This weight is 0 if the arc does not exist.
The incidence matrix is

$$W = OUT - IN = wij$$

The incidence matrix is used to find out the change in the marking of a petri net upon firing a given transition and the characteristic equation of the change in the marking is given by M′ = M0 + T•W, where T is the firing vector stating which transition has fired, M0 is the initial marking¸ is the modified marking, and W is the incidence matrix.

Definition 5 Safe Petri Net: A petri net is said to be *safe* for an initial marking M0 if for all reachable markings, each place will contain at most one token.

Definition 6 Liveness and Deadlock: A petri net with initial marking M0 is *live* if, no matter what marking has been reached from M0, it is possible to ultimately fire *any* transition by progressing through some further firing sequence. A live petri net guarantees *deadlock-free* operation, no matter what firing sequence is chosen.

2. HOW TO USE PETRI NET TO ANALYZE CONVERSATIONS

When we do some researches on the decision supporting system of search and rescue at sea based on multi agent, finite state machines is used to identifying the conversations of the agents. To analyze the conversation model and verify system coherency, first we combine these conversations to form a unified petri net, and then we build petri net’s incidence matrix to check conditions of liveness, safeness and reachability. The conversion of the concurrent finite state machines into a petri net representation is straight forward and proceeds as follows:

- First, each state in the automata is represented by a place in the petri net.
- For each conversation rule present in the state machine, a transition is added to the petri net.
- For each conversation rule adds an arc from the place corresponding to the beginning of the conversation rule to the transition.
- Then add another arc from that transition to the place corresponding to the end of the conversation rule.
- The transition must now be enabled by any and all utterances which trigger that conversation rule. For each conversation rule do the following:
- Add a place for each utterance that will enable that conversation rule.
- Add an arc from the transition, corresponding to each utterance that will enable it to the place(s) in the previous step.
- If there are multiple utterances that will enable a conversation rule, then the arc, transition and arc corresponding to that conversation rule will need to be duplicated.
- Add an arc from the place(s) in the previous step to this conversation rules transition (or duplicated transitions).
- For each final state in each conversation a transition is added. For each final state, an arc is added to its transition.
- For each conversation one place is added, and arcs are constructed from each of the transitions in the previous step to the place added in this step. This represents the corresponding conversation reaching a final state. When the place added in this step contains a token, that conversation has reached a final state.
- One more transition is added for the entire conversation

set which corresponds to all of the conversations reaching a final state and then resetting each conversation to their initial states. An arc is added from each place added in the previous step to the transition added in this step. One arc is added from the transition in this step to each place which is marked as an initial state, thus resetting the petri net.

- Finally one token is added to each place that corresponds to an initial state.

3. THE ANALYSIS OF AGENT CONVERSATION MODELS IN SEARCH AND RESCUE AT SEA DECISION SUPPORTING SYSTEM

3.1 Identifying the Agents and Using Finite State Machines to Represent Agent Conversation Model

In the search and rescue at sea decision supporting system based on multi agent, we identify four agent classes: user interface agent, asset allocation agent, search calculate agent, search and rescue coordinating agent. The conversation models of these agents are shown as Fig. 1 to Fig. 4.

In Fig. 1, Fig.2, Fig.3 and Fig.4, F1 to F16 are all conversation rules and were described in Table 1.

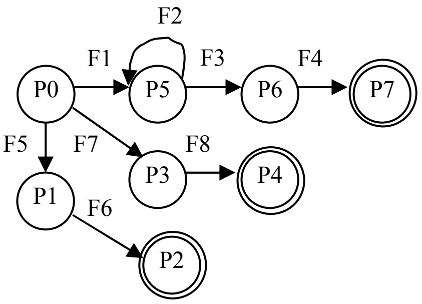


Fig.1. The conversation model of user interface agent

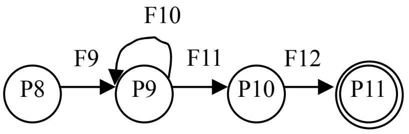


Fig.2. Asset allocation agent conversation model

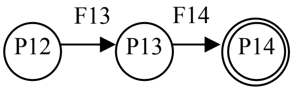


Fig.3. Search calculate agent conversation model

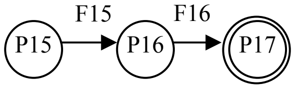


Fig.4. Search and rescue coordinate agent conversation model

3.2 Converting Conversation Models into a United Petri Net

According to this method we convert conversation models into a united petri net as shown in Fig.5.In order to analyze better, we simplify this petri net by referring to In-Out relationship of conversation rules. The simplified petri net is shown as Fig.6.

Table 1. the description of conversation rule

rule	the description of the rule
F1	transmit:propose a message containing the parameters for processing to the asset allocation agent.
F2	received:receive a “ask for answer” message from the asset allocation agent. transmit:reject to the asset allocation agent. suchthat:no answer now.
F3	transmit:propose answer to the asset allocation agent.
F4	received:receive “done” message from the asset allocation agent.
F5	transmit:propose a message containing the parameters for processing to the search calculate agent.
F6	received:receive ”done” message from the search calculate agent.
F7	transmit:propose a message containing the parameters for processing to the coordinate agent.
F8	received:receive “done” message from the search calculate agent.
F9	received:receive a message from the user interface agent.
F10	transmit:propose a “ask for an answer” message to the user interface agent. received:receive reject from the user interface agent. suchthat: no answer now.
F11	received:receive answer from the user interface agent.
F12	transmit:propose a “done” message to the user interface agent.
F13	received:receive a message containing the parameters for processing from user interface agent.
F14	transmit:propose a “done” message to the user interface agent.
F15	received:receive a message containing the parameters for processing from the user interface agent.
F16	transmit:propose a “done” message to the user interface agent.

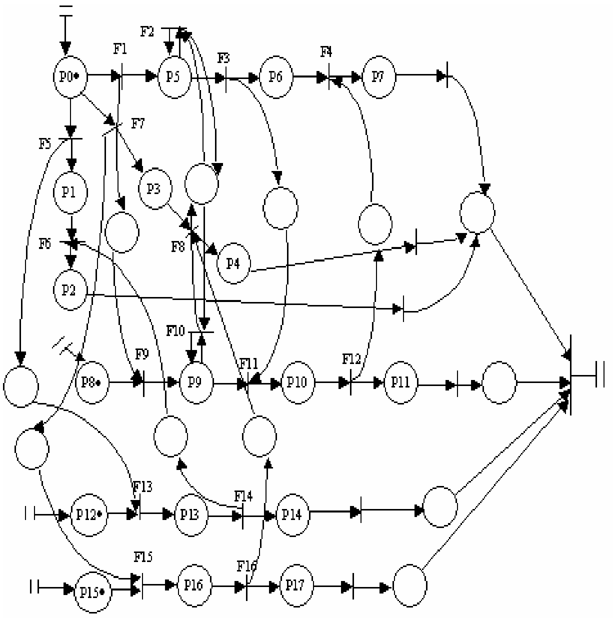


Fig.5. The petri net of conversation models

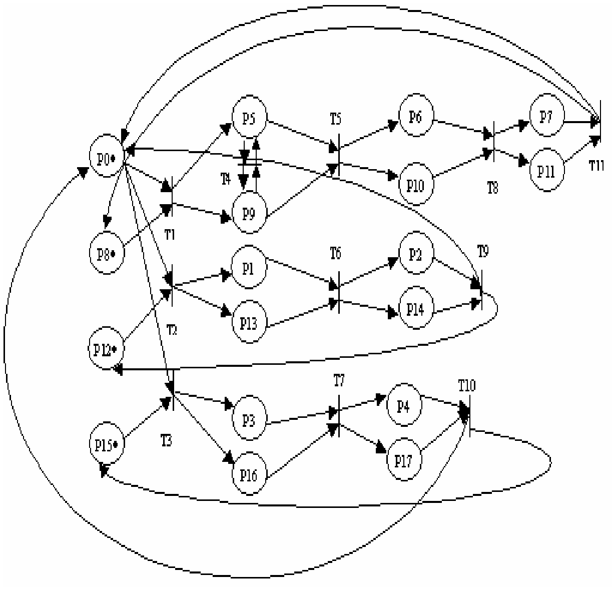


Fig.6. Simplified petri net model

3.3 Building Incidence Matrix to Analyze Safeness and Reachability of Petri Net

According to definition 4 of part one, we build the incidence matrix W of petri net as shown in Fig.6.

$$W=\begin{array}{c|cccccccccccccccccc}
 & p0 & p1 & p2 & p3 & p4 & p5 & p6 & p7 & p8 & p9 & p10 & p11 & p12 & p13 & p14 & p15 & p16 & p17 \\
t1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
t2 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
t3 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
t4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
t5 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
t6 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
t7 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\
t8 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
t9 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\
t10 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
t11 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

The initial marking of this petri net is M0

$$M0=\begin{array}{cccccccccccccccccc}
p0 & p1 & p2 & p3 & p4 & p5 & p6 & p7 & p8 & p9 & p10 & p11 & p12 & p13 & p14 & p15 & p16 & p17 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0
\end{array}$$

From the initial marking M0, we can see in the beginning place p0, p8, p12 and p15 are associated with a token. According to the characteristic equation of the change, we can calculate all reachable markings of M0.If in the initial marking, we fire transition t1, then,

$$T=\begin{array}{ccccccccccc}
t1 & t2 & t3 & t4 & t5 & t6 & t7 & t8 & t9 & t10 & t11 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

So a new marking M1 can be reached,

$$M1=\begin{array}{cccccccccccccccccc}
p0 & p1 & p2 & p3 & p4 & p5 & p6 & p7 & p8 & p9 & p10 & p11 & p12 & p13 & p14 & p15 & p16 & p17 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0
\end{array}$$

After that, t4 can be fired and a new marking M2 can be reached, the vector T can become,

$$T=[\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

According to this method, we can get all reachable markings from M0, so the marking matrix for the petri net is shown as following,

$$M=\begin{array}{c|cccccccccccccccccc}
 & p0 & p1 & p2 & p3 & p4 & p5 & p6 & p7 & p8 & p9 & p10 & p11 & p12 & p13 & p14 & p15 & p16 & p17 \\
m0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
m1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
m2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
m3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
m4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
m5 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
m6 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
m7 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{array}$$

It can be seen that in all possible marking of the petri net, the number of tokens at any place never exceeds one. Thus the petri net is safe. Fig.7 shows the reachability graph of the system. Each node of this graph represents a reachable marking of the petri net. We can see from this graph that from all the reachable markings of the system, a firing sequence exists, which contains every transition of the petri net. That implies that every transition of the petri net is live and the system is deadlock free. So we can draw a conclusion that the original conversation models of the agents are right.

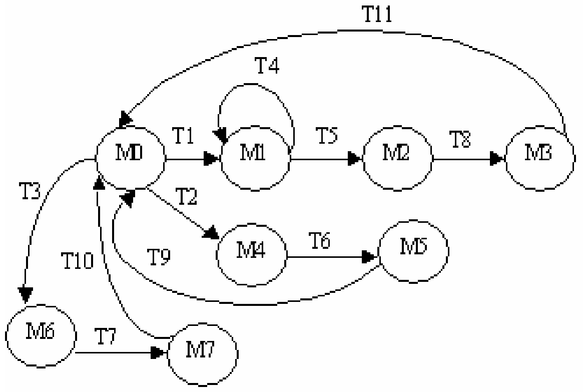


Fig.7. The reachability graph of the system.

4. CONCLUSIONS

Agents are becoming one of the most important topics in distributed and autonomous decentralized systems, and there are increasing attempts to use agent technologies to develop software systems. Such systems are complex and there is a pressing need for system modeling techniques to support reliable, maintainable and extensible design. This paper presents a petri-net based approach for analyzing conversation models among agents. This approach was applied into the search and rescue at sea decision supporting system based on multi-agent, thus coherency and coordination of agent conversations were verified.

REFERENCES

[1] Alan K, Galan and Albert D, Baker, *Multi-Agent Communication in JAFMAS*, http://www,boeing,com/special/agents99/galan_baker,pdf

[2] Alan K,Galan, *Petri Net Analysis of Application Examples*, http://www,ececs,uc,edu/~abaker/JAFMAS/JAFMASscianalysis,pdf.

[3] Søren Christensen and Niels Damgaard Hansen, “Coloured Petri nets extended with channels for synchronous communication, In Rober Valette”, editor, *Application and Theory of Petri Nets 1994*, Proc, of 15th Intern, Conf, Zaragoza, Spain, June 1994,LNCS, pp,159~178

[4] Michael K¨ohler, Daniel Moldt, and Heiko R¨olke, “Modeling the behaviour of Petri net agents”, In J, M, Colom and M, Koutny, editors, *Proceedings of the 22st Conference on Application and Theory of Petri Nets*, volume 2075 of LNCS, Springer-Verlag, June 2001,pp,224~241

[5] Daniel Moldt and Frank Wienberg, *Multi-agent-systems based on coloured Petri nets,In P, Az´ema and G, Balbo* [2], pp, 82~101.

Agent-enabled Tactics for Synchronal Cooperation by Interdisciplinary Professionals Avoiding Simultaneous Confliction *

Qingru Kong
Personal Power Department, Wuhan University of Technology,Wuhan,
Hubei Province 430070, China
Email: kong1973@mail.whut.edu.cn

ABSTRACT

Real-time cooperative design is a good approach to integrate the advantage interdisciplinary professional resources to improve the integrated competitive capacity of a development. However, the simultaneous confliction has always been the bottleneck for the application of interdisciplinary professional synchronal cooperation. Simultaneous confliction is induced mainly by two causes: simultaneous submission and transition delay. This paper employs agent technology to overcome the above two problems. An intelligent graph net topology structure, which nodes are classified as the client nodes, web server, and the application servers, is proposed firstly. Agents are employed as the bridges between these nodes to make the process of cooperative intelligent avoiding the simultaneous submission. The multi-layer buffers architecture is detailed to implement the in-phase sharing interface avoiding the transition delay. The methods have obviously improved the degree of intelligence and the in-phase capability of real-time cooperative design, consequently are helpful to make the interdisciplinary professionals synchronal cooperation technologies practical.

Keywords: Multi-Agent Systems, Personal Power Scheme, Synchronal Cooperative Design, Simultaneous Confliction.

1. INTRODUCTION

Product diversification and individuated requirements lead to complex product design with more functions consideration requiring the strong integrated design competencies in diverse domains. However, apparently a single professional cannot be high-class in any domains, while they are competitive in their core capabilities[1]. Two features of modern enterprises, distributed in geographic areas and professional in a certain domain, make product developments more and more complex, which forces companies to find new ways to improve their product development process. CSCD (Computer Supported Cooperative Design) is an effective approach to integrate diverse core capabilities of interdisciplinary professionals distributed in diverse enterprises or departments to develop a product with diverse consideration[2]. Considering the way of running, CSCD can be classified as asynchronous cooperative design and synchronous cooperative design.

Through an asynchronous cooperative design system, distributed participants involved in a product development process can work together through sharing documents and database of information that provide services for retrieving and managing objects in a repository. Version management is an important part for asynchronous cooperative design. Versioning comprises the specification storage, and maintenance of versioned design objects whereas activity management is responsible for cooperation control, design flow management and management of design transactions processing versioned design objects. Recent reports show that the construction of repositories to manage design artifacts by offering adequate storage and manipulation services has reached a good level of maturity[3]. However, a main disadvantage of asynchronous cooperation technology is that it does not provide a uniform environment for diverse professionals to real-time cooperation in some necessary cases.

Compared to asynchronous cooperation, synchronous cooperation provides a good opportunity for distributed designers to real-time discuss a product in respective viewpoints online. Synchronous cooperative design aims at intelligence real-time sharing. Considering a design, the estimations from designers of different engineering domains are different in respective viewpoints. Synchronous cooperative design is a good approach to integrate the interdisciplinary professional resources of designers of different domains to improve the integrated competitive capacity of a product[4]. Synchronous cooperative design is mainly introduced to enhance integrated advantages of product design by higher degrees of process parallelism and cooperation in diverse domains. The simultaneous confliction rises while professionals involved in the process of cooperative design, which limits the CSCD practical. This paper focuses on eliminating the bottleneck by employing multi-agent.

2. INTELLIGENT TOPOLOGY STRUCTURE

2.1 Substantial Framework of Net

Due to the increasing complexity of synchronous cooperative design systems, it gets harder and harder to guarantee desired system's properties. The usual approach to tackle this problem is frequently called "divide and conquer": systems will be composed of simpler aspects that are easier to handle. The topology structure of a synchronous cooperative design system is described in two aspects: one is the substantial framework of net, which is the frame and muscle of the system, while another is the architecture of information flow, which is the brain and nerves of the system.

To construct the substantial framework of net, firstly making the correlative functions involved in synchronous cooperation sure by analyzing the workflow of synchronous cooperation is necessary. The workflow of synchronous cooperation fell into three successive phases: unattached review, sub group discussion, and opinions compromise. These phases may be interferential during the whole process of cooperation. In the phase of unattached review, designers respectively visit the interested files through correlative software to draw the unattached original opinions on the design. In the phase of sub group discussion, part designers share a public interface to discuss the design in some correlative files and domains, meanwhile, any designer can unattached review the interested

* Hubei Province Natural Science Fund(No. 2005ABB023).

file in a sub window. Namely, the interface of cooperation comprises of three sub windows – public window including public interface of software used to review the design together by part group or all participants, private window including individual interface of software used to unattached review the interested file, and the discussion window used to exchange the discussion information. In the phase of opinions compromise, all participants share a uniform interface to draw an optimal design with the compromise of respective opinions. According to the functions involved in the workflow, a synchronous cooperative design system should comprise of the modules of software sharing, resource control, requirement control, and negotiation manage.

Software sharing is the foundation of cooperation, which offers a uniform interface to participants by mapping the interface of software in application server to the web server or respective browser. Software sharing is divided into two manners of sharing: remote calling and public interface sharing. Remote calling is used to call the correlative software installed in the application server in the process of unattached review. When the software control receives the requirement of remote calling, it will find and run the adaptive software in the application server, and then map the interface to the private window in the caller’s browser. Public interface sharing is used to share public interface to discuss the design. When software resource control receives the requirement from a group, it will find and run the adaptive software in the application server, and then distributes the interface to the public window of the member’s browser.

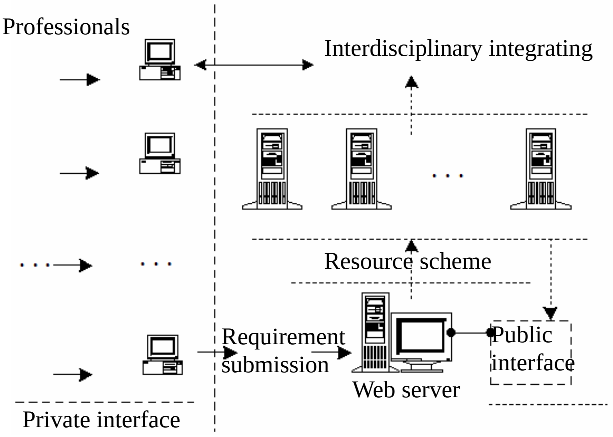


Fig.1. intelligent multi-lay net topology structure

The module of resource control, requirement control, and negotiation manage, which is the intelligent part of the net, are used to avoid the conflicts of requirements and realize parallelism control. These modules belong to the intelligence aspect and are built by multi-agent technology.

In order to effectively organize these modules above, a graph net topology structure illustrated as Fig1 is proposed based on the previous research of software sharing.

The nodes of graph net topology structure are classified as the client nodes, web server, and the application servers. Agents are special nodes regarded as the bridges between these normal nodes. Designers visit the system by web browse on the client ends. This system is a B/S structure so that no any plug-in is required to install on the client ends, which make the system more practical. A registered designer can take part in the cooperative design after his authority is confirmed. The interface of a client is divided into the public area including the public discussion module and the software environment, which is the interface mapping from the web server, and the private area, which is the relative interface mapped from the application server. Experts can discuss and modify the design of a uniform document in public area, and view the change of the relative file according to the public modification in private area.

Web server is the center of information flow. Web server offers public interface, which is mapped to the public areas on client ends, for software and the file to be discussed. The web server is divided into public discussion and public application areas. Public discussion area provides a uniform discussion environment to all experts so that the design can be optimized in multi viewpoints. Public application area is mapped from the selected application server to provide a uniform application environment to all experts for viewing or modifying.

2.2 Architecture of Intelligent Information Flow

The user of a synchronous cooperative design system is actually a geographically distributed team including various designers from various companies or departments, which manner of working is rapidly increasing in prevalence. However there is a big gap between research and practice in synchronous cooperation due to the bottleneck of complex problem of synchronization control and conflict resolution. On collocated teams, depending upon the nature of the conflict and how it is handled, conflict can spur innovation and improve performance or it can cause rifts and worsen performance. Currently much of the research on conflict has focused on understanding the antecedents and effects of conflict in teams in which members are physically collocated. In this research, because the conflicts come from not only various designers but also requirements and resource, the main task of intelligence aspect is coordinating various conflicts of the system. Issues of synchronization control and conflict resolution of separated resources, including the software resource, file resource and the expertise resource, are taken into mainly consideration. This research takes multi-agent technology to solve the problem of synchronization control and conflict resolution.

Agents and multi-agent systems, which can seamlessly integrate these independent components into a single system with interrelated information flows, are more and more successfully deployed and used in practice of distributed intelligence, even if many theoretical issues underlying them are not fully understood at this point and subject of current investigations. Agent can be roughly defined as independent intelligence unit, which can conduct the information and events from others. An agent can be a component or applet distributed in a net environment. Multi-agent systems comprises of interdependent agents, which mutually cooperate to complete a system decision-making through the intelligent information flows mass net. The agent-oriented software technology provides the right level of abstraction for distributed intelligence applications, representing a valuable help for handling their complexity.

This multi-agent system contains six kinds of agents: professional agent, resource agent, requirements agent, application agent, software agent, and file agent, which model is showed as Fig2.

Designer agent deputizes for actions of the designer. To review the design, a designer needs to call software, call file. To cooperate with others, a designer needs to join a present sub group, create an original sub group, invite anther one to join the sub group, and submit opinion. To realize the automation of cooperation, these actions are deal by designer agent. The requirements and feedback constitute the interface of the agent.

Requirement agent deals with the synchronization requirements to eliminate conflict. Requirements agent is constructed to control the requirements from the designers, such as the requirements of joining, discussion, calling of software, visiting and modifying files, to conduct constraints and wipe out locking conflicts. Requirements agent is the intelligence core of the system, which insures the mass information flow reasonable and efficient. Requirements agent dynamically allocates the software and files, which are called by designers, and arranges the authorities such as view, conversation, and modification to expert. Some authorities can be arranged to many designers, such as view and conversation, while some authorities can only be arranged to single such as modification. Requirement agent is the foundation of negotiation manage module, which is used to manage constraints and wipe out locking conflicts.

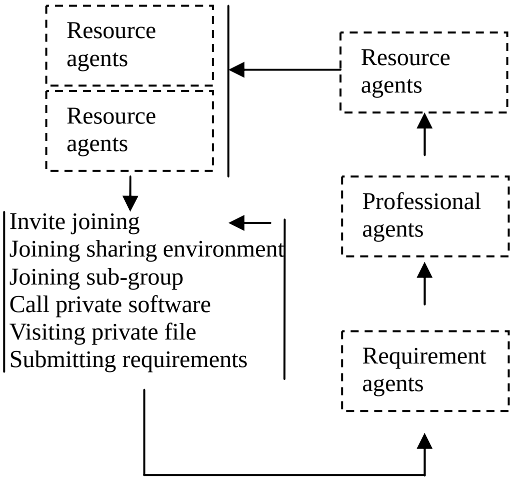


Fig.2. multi-agent models

Resource agent is employed to allocate useful resource such as the software or documents to corresponding designer, according to the mechanism of charge equilibrium. Software agent controls the schedule of the called software tool. File agent arranges files to corresponding designers.

Software agent and file agent assist the requirements agent to find and arrange resource by sending the software state and file state to the requirements agent. Application control supervises the charge of every application server, which is the criterion of charge equilibrium for requirements agent. Software agent and file agent navigate the software and files group in application servers to arrange software and files to designers reasonable according to the requirements and the state.

3. IN-PHASE COOPERATIVE INTERFACE

The image online rendering efficiency is another bottleneck of synchronization cooperative design. Since all data are transmitted by net in process of synchronization cooperation, the amount of data transmitted during the whole process is vital to the in-phase of interaction. The data transmitted in synchronization cooperation is classed as the action and the mapped interface. All manipulation and interface are transmitted as information bundle. To an interface, two kinds of authorities are allocated to diverse designers, view and manipulation. In the unattached review section, designers can fully manipulate the private sub windows. In the sub group discussion section, only a manipulator is arranged, the others are viewers. The manipulator owners the authority to manipulate the public interfaces, and the others all can only view it. Namely, a designer may owner two roles as viewer and manipulator. The actions from manipulators are transmitted on net to remote control the software. Generally the manipulation and requirements contains small data packages, which are not crucial to the transmission efficiency. However, to all designers, the interface in the client ends actually are mapped from the display buffers of servers, which contains a great deal of data thus is vital to the cooperation efficiency. It is difficult to lower the amount of data of a mapped interface, since which is a color image. Apparently it is unpractical that a designer receives the manipulation information of interface from others with required a long waiting. This research employs some novel tactics such as the discrete pixels transmitted on web only, interface reconstructed through interpolation on client end, to solve the problem.

The Fig 3 illustrates the way of transmitting interface through multi-layer buffer. The way is divided into five phases. First, a manipulator sends his actions to the application server. Second, the disposed data are transmitted to the first layer buffer – display buffer in application server. Third, these data in the first layer buffer is transmitted to the discrete processor, which abstracts the discrete pixels of interface in server. Fourth, these data of discrete pixels are transmitted to the client end via net. Five, discrete data are transmitted to the interpolator to reconstructed the interface in the client end.

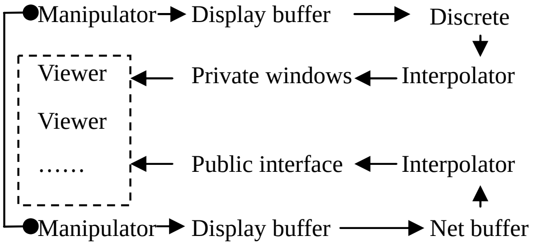


Fig.3. Manner of transmitting interface

With multi-layer tactics, only some key pixels data are necessary to be transmitted to client ends from servers, the image interpolation is employed to reprint the whole interface of client ends. As the operations of interpolation are distributed to respective client ends and only a few data are transmitted through web, the real time capability is improved substantially.

This research applies image interpolation technology to lower the transmission charge on web. Designers can remotely share and exchange multi-dimensional information, including the dynamical image information of application environment, interaction manipulation information, and the text information. The dynamical image information is not necessary to be transmitted in a uniform high level of detail. This paper proposes a rendering tactics in multi precisions to improve the efficiency of rendering, namely lower important areas, such as the areas to be viewed without having to be modified, are rendered in lower precisions. The precisions of rendering can be controlled by the precisions of image interpolations. Respective precisions of interpolations can be defined in respective client ends with designers, which make it possible that respective special requirements of higher quality of interface require higher efficiency of public net. This tactic improves the efficiency of rendering without evidently depressing the capability of cooperation.

An image can be represented as a 2-dimension array. Every element of the array representing a pixel is a matrix including three colors variables R, G, B. These elements can be taken as distinct points to reconstruct the interface on client ends with proper interpolation functions. On a closed interval, some approximation of arbitrary functions such as Lagrange interpolating polynomial, which is determined simply by specifying certain points on the plane through which they must pass, is available. However, two causes limit the use of Lagrange interpolating, one is that the oscillatory nature of high-degree polynomials and the property that a fluctuation over a small portion of the interval can induce large fluctuations over the entire image range restrict their use. The second is that a reconstructed interface has to employ interpolation function twice respectively in transverse and vertical directions since an image is a 2-dimension array, which will seriously enhance the oscillatory nature of high-degree polynomials. The piecewise polynomial interpolation technique is an alternative approach to control the errors of image interpolation. Some piecewise polynomial interpolation techniques are alternative, such as piecewise-linear interpolation, a piecewise polynomial and cubic interpolation. Piecewise-linear interpolation is the simplest piecewise-polynomial approximation consisting of joining a set of data points by a series of straight lines. However, the image reconstructed with piecewise-linear interpolation is not smooth, which seriously depresses the vision effect, because a disadvantage of linear function approximation is that there is likely no differentiability at the endpoints of the subintervals. Piecewise polynomial is also unavailable due to the derivative of the distinct points required. Cubic interpolation is the most common piecewise-polynomial approximation using cubic polynomials between each successive pair of nodes. A general cubic polynomial involves four constants, so there is sufficient flexibility in the cubic interpolation procedure to ensure that the interpolation is not only continuously differentiable on the interval, but also has a continuous second derivative. Consequently solution of cubic interpolation is smoother and higher quality, which can be employed to reconstruct the interface in high efficiency.

4. CONCLUSIONS

Real-time cooperative design is a good approach to integrate the advantage interdisciplinary professional resources to improve the integrated competitive capacity of a development, which is the desire to rapidly react to the diverse individuated requirements. The research proposes the agent-based intelligent control methodology and multi-lay net structure to solve the two main problems - simultaneous submission and transition delay, which are the main causes of simultaneous confliction. The methods have obviously improved the degree of intelligence and the in-phase capability of real-time cooperative design, consequently are helpful to integrate and optimize interdisciplinary professional resource.

REFERENCES

[1] J.SANDUSKY, “Infrastructure Management as Cooperative Work: Implications for Systems Design”, *International Journal of Computer Supported Cooperative Work*, Vol.12, No.1, 2003, pp.97-122.

[2] Kong Qingru, “Methodology for estimating the integrated core-competitiveness of multi-subject professionals”, *Journal of Hubei University of Technology*, Vol.27, No.3, 2007, pp.35~39.

[3] Harding, J.A, and Popplewell, K, “Enterprise design information: the key to improved competitive advantage”, *International Journal of Computer Integrated Manufacturing*, Vol.14, No.6, 2005, pp514-521.

[4] Erdeny, A, “Methodology for Intelligence Resource Integrating in Interdisciplinary Subjects”, *Journal of Computer Integrated Information Technology*, Vol.40, No.15, 2006, pp369-375.

Qingru Kong is a lecturer of Wuhan University of Technology. She graduated from Huazhong University of Science and Technology in 2003 with MS degree. Her research interests are in modern personal power management, which apply information technology into personal resource scheme to improve the efficiency of personal power management and lower the personal power cost. She has published over 10 papers in the field.

Multi Agent-Based Distributed Component Library System Architecture*

Wenfei Lan, Jiguang Lu
College of Computer Science, South-Central University for Nationalities
Wuhan, Hubei, 430070, China
Email: lanwenfei1@163.com

ABSTRACT

The component library is among the core technologies of component-based software development,is has become one of the focus in the field of software reuse.In this paper,by analyzing the current situation of component library and its reuse, a multi agent-based distributed component library system is proposed with agent technology. The architecture of this system is also presented. Via giving an example, we analyze the cross-library component retrieval process with the cooperation among agents of the system, therefor we can jump to conclusion that the system can meet the need of component retrieval of cross component library through the cooperation of multi-agent and has features of loose coupling as well as high flexibility.

Keywords: Multi-Agent, Component, Distributed Component Library, Component Retrieval, Facet

1. INTRODUCTION

Component-based software development (CBSD) can improve the efficiency of software development and quality as well as lower the cost of development and maintenance. So it has become one of the focus in the field of software engineering study. The component library is among the core technologies of CBSD.

Some enterprises and organizations have built private component library to support CBSD, such as Shanghai Component Library(SHCL)[1] owned by Shanghai Software Component Service Center (SSTC) and Jadebird Component Library(JBCL) owned by Beijing University[2][3]. Heterogeneous, generally distributed but locally integrated are features of component library. According to the application of software component in Shanghai, the total components of all enterprises are more than that of Shanghai Component Library, but it is difficult to put all of the components into Shanghai Component Library because of the intellectual property rights and business operation of the software components. Thus it has brought some problems listed below to the reuse of software components.

1) The component demanders do not know who has those components that he needs, and the component providers do not know their component demanders.
2) These enterprise component libraries are heterogeneous, because the component classification scheme is different from these libraries.

In this paper, we propose a multi agent-based distributed architecture solution of component library system by making use of the initiative and cooperativity of agent[4]. It could solve the problems mentioned above.

2. SOLUTION

In this paper, the fundamental thinking of architecture for distributed component library system is: deploying multi-agent on the node which has its own component library, such as Publishing-Agent(PAgent), Retrieval-Agent(RTAgent) and Request-Agent(RQAgent), while the physical distribution and implementation of the component library remain unchanged. These agents cooperate to meet user's retrieval request with its initiative and Bulletin Board System(BBS) which are comprised of public publishing area(PArea), retrieval area(RArea) and feedback area(FArea).

In the distributed component library system, any enterprise can register component to the publishing area of BBS through Publishing-Agent. Each enterprise can search a component which he wants but not at hand in the publishing area or sends its request to the searching area of BBS. Every enterprise can find the retrieval request from others by Retrieval-Agent initiatively and perform the retrieval transaction locally, and publish the retrieval result to the feedback area at last. Those who initiate the retrieval request can find the components in the feedback area through Request-Agent. Such multi-agent cooperation with BBS can solve the first problem mentioned in the introduction.

However, component libraries are heterogeneous because the component classification schemes of different enterprise component libraries are not the same. The faceted classification scheme[5][6] is adopted by most enterprise component libraries currently. The heterogeneous feature is embodied in the difference of faceted classification scheme. So we take the faceted classification scheme of Shanghai Component Library as the base scheme, and set a description adapter in each agent for a bilateral switching between Shanghai component classification scheme and other schemes. Thus we can solve the heterogeneous problem among different component libraries.

3. DISTRIBUTED COMPONENT LIBRARY SYSTEM ARCHITECTURE

The distributed component library system is comprised of enterprise component libraries systems(ECLS) which are distributed physically. Fig.1 shows the distributed component library system architecture that we proposed. Shanghai component library system(SHCLS) is the logic integrated entrance of the system.

Each agent in fig 1 can implement its own function with initiative and flexibility. Agents communicate with each other by means of board[7]. The public component information published by enterprises is shown in the publishing area of BBS. The component request information is shown in the request area and component retrieval results are stored in the feedback area. The Publishing-Agent is responsible for publishing the public component information to the publishing

* Supported by the Emphases Foundation for Shanghai Science and Technology Commission of China under Grant No.04DZ15022.

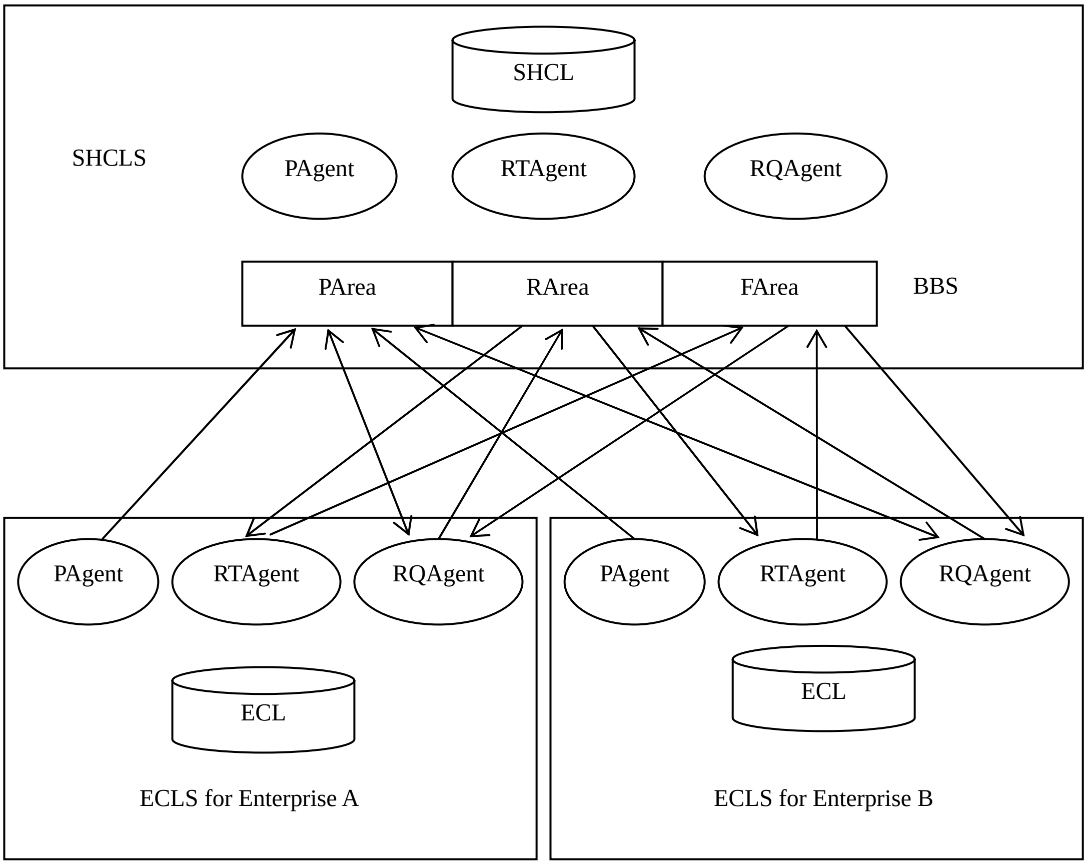


Fig.1. Mutil Agent-Based Distributed Component Library System Architecture

area. The Retrieval-Agent is responsible for getting requests from the request area and sending the retrieval results to the feedback area. The request agent is responsible for sending component requests to the request area and obtaining the retrieval results from feedback area or retrieving components from publishing area. Each agent should transform the component description into the facet scheme of Shanghai component library by a Component Description Adaptor when sending information, such as publishing information, retrieval information and feedback results, to BBS. Also these agents should transform the component description information into the local facet-based scheme after obtaining associated information from BBS.

Now we are going to analyze the cross-library retrieval process with the cooperation among agents. Let's take enterprise B for an example, there are two ways for the request and retrieval of the wanted components:

1) Enterprise A publish the component description information to the publishing area initiatively, and the Request-Agent of enterprise B retrieve the wanted component from the publishing area.
2) The Request-Agent of enterprise B sends its component request to the request area , the Retrieval-Agent of enterprise A gets the request from the request area and search in its own component library, then send the retrieval result to feedback area. At last the Request-Agent of enterprise B can obtain the wanted components from the feedback area.

Each enterprise could have such process of request, retrieval and obtaining components.

In the distributed component library system, those who own the components, such as enterprise A, could publish their component information to others by measure 1 or authorize the component to their partners. So it is very flexible and can protect intellectual property rights.On the other hand, the installation only consists of adding an agent to the enterprise component library which can interact with Shanghai Component Library. So the operation in the enterprise would not be influenced and there is no need for those enterprises to take any changes, which lower the cost of exchange components among enterprises.

In Fig.1 we simplify the architecture of every enterprise component library system; the full structure is shown in fig 2. The actual components and description information are stored in the component library. The component library management system(CLMS) is responsible for the management and maintenance to the components. In the process of enterprise development, at first they will search component they need in their own component library. If there are no matchable components, they will initiatively retrieve and get results from the publishing area with Request-Agent. If they still fail to find the wanted component, they will actively send retrieval request to the request area and get the results, which are sent to the feedback area by other enterprises, with Request-Agent. In addition, every enterprise can publish component information through Publishing-Agent and obtain retrieval request through Retrieval-Agent, search in local component library and send feedback results. From the above discussion we can see that the retrieval range is expanded to the whole distributed component library.

From the above analysis we can educe the conclusion that distributed component library system has the following functions:

1) Implements the association and retrieval between the heterogeneous component libraries.
2) Solves the problem of "blind connection" between component demanders and component providers.

Therefore, it can excellently meet the requirement of cross-library component retrieval and has features of loose coupling as well as high flexibility.

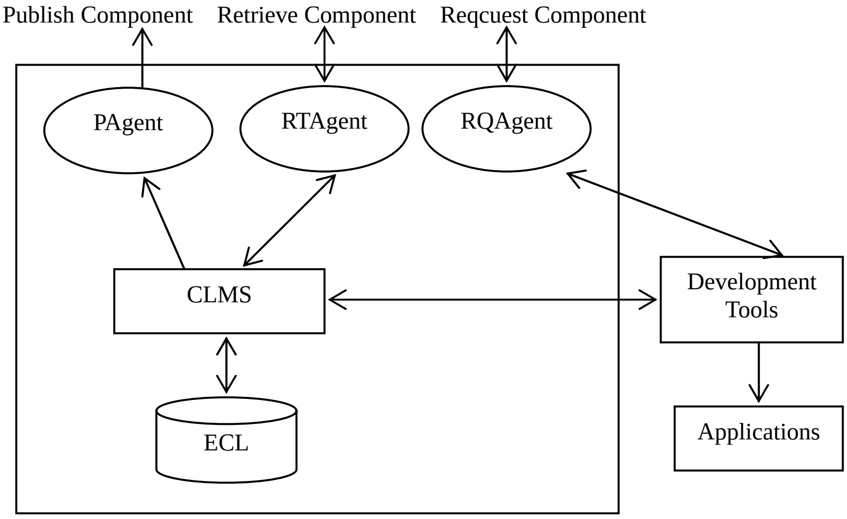


Fig.2. Enterprise Component Library System Architecture

4. CONCLUSIONS

In order to improve the retrieval range and reuse ratio[7], different enterprises should share their component assets and consequently bring the requirement of cross-library component retrieval. In this paper, by analyzing the current situation of component library and its reuse, we proposed a multi agent-based distributed component library system with agent technology[9][10], which can excellently meet the need of reusing component.The study is of theoretical importance and practical value to CBSD.

REFERENCES

[1] Yunjiao Xue,*Research on Intelligent Agent-based distributed compoment library technology:[Ph.D. Thesis]*. Shanghai:Fudan University,2006.

[2] Jichuan Chang, Keqin Li, et al,"Representing and Retrieving Reusable Software Components in JB (JadeBird) System,"in *Electronica Journal*,Vol.28,No.8, Aug 2000,pp.20~24.

[3] Jichuan Chang, Keqin Li, et al. R*epresenting and Retrieving Reusable Software Components.Computer Science*,Vol.26,No.5,May 1999,pp.45~49.

[4] Michael Wooldridge,*Multi-Agent System Introduction. Beijing: Publishing House Electronics Industry,* 2003.

[5] Yuanfeng Wang, Yong Zhang, et al, "Retrieving Components Based on Faceted Classification," in *Journal of Software*,Vol.13, No.8,Aug 2002,pp.1546~1551.

[6] Yuanfeng Wang, Xue Yunjiao, et al, "A Matching Model for Software Component Classified in Faceted Scheme ," in *Journal of Software*,Vol.14,No.3,Mar 2003, pp.401~408.

[7] XinJun Mao, *Agent-Oriented Software Development. Beijing*,Tsinghua University Press,2005.

[8] Yunjiao Xue, Qian Leqiu et al,"Research on the Knowledge Management in Component Retrieval by Intelligent Agent,"in *Computer Engineering and Application*, Vol.42, No.8,Mar 2006,pp.11~15.

[9] N.G. Shaw, A. Mian et al,"A Comprehensive Agent-based Architecture for Intelligent Information Retrieval in A Distributed Heterogeneous Environment,"in *Decision Support Systems,*Vol.32,No.4, Apr 2002,pp.30~35.

[10] Wanxia Yu, Weicun Zhang,"Synthetic Laboratory Management Information System in University Based on Agent," in *Computer and Modernization,* No.1, January 2006, pp.82~83.

Wenfei Lan is an associate professor and a graduate student supervisor of College of Computer Science, South-Central University for Nationalities, a candidate doctor of College of Computer, Huazhong University of Science and Technology. She graduated from South-Central University for Nationalities in 1989; from Huazhong University of Science and Technology in 1999 with specialty of computer softwore. She has published three books, over 30 Journal papers. Her research interests are in component technology, software reuse and object-oriented methodology.

A Discrete Part Manufacture Scheduling Framework Based on Multi-agent *

Zhanjie Wang, Xian Li, Ju Tian
Department of Computer Engineering, Dalian University of Technology,
Dalian, Liaoning, China
Email: wangzhj@dlut.edu.cn, lixian.dlut@yahoo.com.cn

ABSTRACT

According to the disadvantages of arrange schedule by hand in a discrete part produce company, this paper introduces a manufacture scheduling framework based on multi-agent. In this framework, parts which are going to be processed and devices which are used are regarded as different kinds of agents, every agent is able to calculate reason and decide, different kinds of agents are able to communicate and negotiate with each other. Based on this, the structure and regulations of single agent are designed, the negotiation mechanism among agents are defined. Under the negotiation mechanism, parts which are going to be processed can be assigned to devices intelligently, and the processing order can be changed temporarily according to the priority of parts. Finally, the result schedules which are generated by the framework using parts manufacture data of the company are given. The results show that the framework proposed is able to arrange the procession of parts reasonably considering of the processing order, processing cost, deposit cost, and is able to reach a higher device utilization rate and save a lot of hand work.

Keywords: Discrete Processing, Manufacture, Multi-Agent, Scheduling, Negotiation Mechanism.

1. INTRODUCTION

The research objective of this paper is a discrete part manufacture company whose produce mode is based on orders. Usually the sizes of orders are small and various, thus the manufacture are charactered as changeful steps and agile management. The old scheduling function of the company is so rough that it only able to arrange steps by hand. This paper strives to develop a new and efficiency scheduling method to resolve this problem.

Multi-agent scheduling method syncretizes many present-day theories and technical production, with its distribution attribute complicated scheduling problem can be decomposed to many simple sub-problems to resolve, and using its intelligence attribute, it’s quite easier to fit the need of flexibleness, agility and rebuild ability of manufacture system. So this paper designed a manufacture scheduling framework based on multi-agent method. The structures and regulations of agents are designed and intelligent functions are introduced to meet the needs of scheduling objective. The expecting scheduling methods are given to deal with unexpected events during the production. In the end of this paper, the framework is realized and used in the scheduling of company, and the results are analyzed.

2. THE DESIGN OF MULTI-AGENT SCHEDULING FRAMEWORK FOR DISCRETE MANUFACTURE

2.1 The Abstract of Scheduling Problem

The part manufacture company produces parts according to orders and its productions are discrete parts. An order is regarded as a set of products with same time restrictions, different orders have different time restriction. In practical production, a product may comprise several components; a component may appear as parent item with several components as its sons, as shown in Fig.1. And every component will be completed in several working procedures.

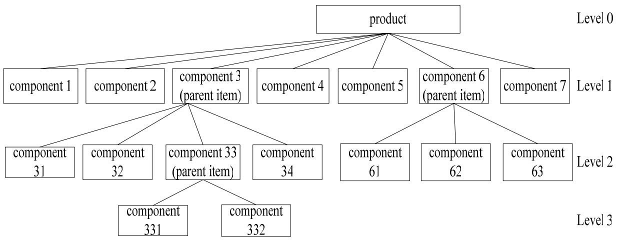


Fig.1. An example of BOM tree

In this framework, task represents component which is going to be processed, step represents a working procedure of a component, and device represents a machine which is used to execute a step. So the scheduling problem comes down to a job shop scheduling as follows.

1) There are n tasks, and each task consists of m steps.
2) Each task has been given processing order and processing time.
3) There are more than one devices can afford service for one step.

According to the requirement of product, the basic restriction conditions are [1]:

1) Only one step of a task is allowed to be processed at once.
2) Each device can process only one task at once.
3) Steps can’t be preempted.

These three conditions should be satisfied at the same time.

The conditions when a new step is allowed to be processed are:

1) None of the other steps of the task is in processing.
2) All of the steps before this step of the task are finished.
3) At least one of the devices which the step needed is available.

These three conditions also should be satisfied at the same time.

2.2 The Design of Multi-Agent Framework

According to Fig.1, assembly work is necessary when components are completed. This paper concerns the produce of components, thus the assemble time is not included in scheduling process, but is considered of. To make the system simple and intelligible, the transport time between devices and install and uninstall time on devices of part are included in the processing time of steps without compute separateness.

* This paper is partially supported by AMD-SUN United Lab.

Under the presumptions above, a hybrid multi-agent scheduling framework is designed based on the analysis in 2.1[2]. As shown in Fig.2, there are two subsystems: task subsystem consists of task management agent (TMA) and task agent (TA) and resource subsystem consists of resource management agent (RMA), resource agent (RA).

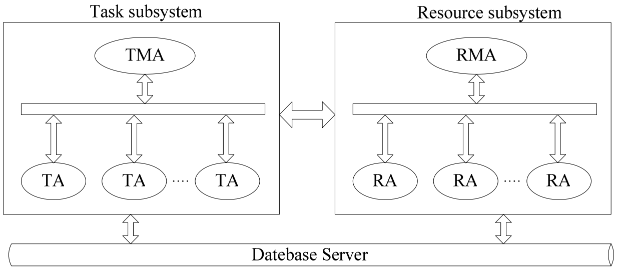


Fig.2. Hybrid MAS framework

Resource agent represents a device and it is created and managed by RMA. Task agent represents a task and it is created and managed by TMA. The manage relationship between TMA and TA, RMA and RA embodies the hierarchical of the framework. RA and TA are able to communicate with each other without management agents. They can make decisions by themselves according to their regulations and evaluation algorithms, and with their knowledge they are able to resolve some global problems independently. TMA and RMA have a broad view. They can control the system according to the global view and have capabilities to deal with some dynamic disturbance conditions. The communicate mechanism embodies the heterarchical of the framework.

A local area network is used as an infrastructure of the scheduling system in order to communicate among agents. Database is used as a blackboard to store the work in process information, device and task information.

This framework supports both transverse and lengthways negotiation which embodies hierarchical one and heterarchical one’s advantages contemporarily. These agents construct a flexible and adaptive system where agents can be replaced or added freely. The system configuration can be continuously changed to accommodate changing requirements.

3. IMPLEMENTATION OF MAS

3.1 The Structure and Negotiation Mechanism of System

According to the framework designed in section 2.2, the use case diagram of proposed MAS is shown in Fig.3.

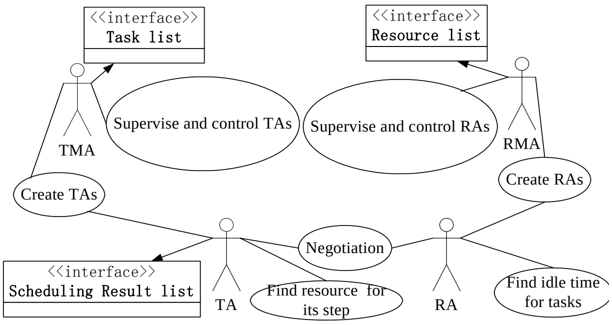


Fig.3. The use case diagram of MAS

The interfaces in Fig.3 are used to associate with database server. The negotiation mechanism of reference [2] is adopted for agents to communicate with each other. A analyze chart of the negotiation mechanism is shown in Fig.4.

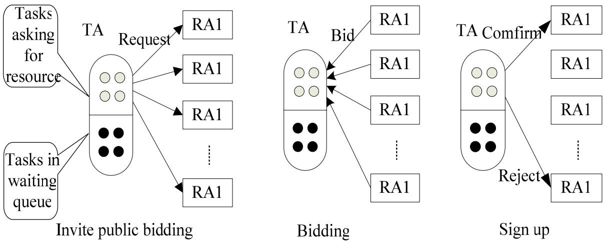


Fig.4. Negotiation processes

(1) Invite public bidding. TMA creates and starts TA when a new task arrives at the system. TA starts its first step and sends service request to RMA. After RMA receives the request, it selects m devices that can offer the service TA needs and sends request messages to RAs which represent these m devices.
(2) Bidding. RA selects the most suitable task in its requesting queue to offer service. RA computes the bid price, and then sends a proposal bid to the selected TA.
(3) Sign up. TA collects all the proposals and selects the best proposal. Then TA sends confirm message to the best RA and reject messages to the other RAs. RA who receives the confirm information assigns device for TA and notifies TA to start next step. RAs who receive reject messages delete TA from their buffer and begin to arrange other TAs in its buffer.

Each time TA starts its next step, it will check whether all its steps are finished, if true, it will notify TMA to check whether the arrange result is reasonable from global view and end itself. If TA receives no offer, bad offer or breakdown message in the designed time, TMA will deal with these problems from the global view.

Several negotiation processes progress in parallel when arranging time for several tasks. Agents make decision using evaluation functions during negotiation and arrange processing time for steps reasonably considering many factors.

3.2 Evaluation Functions of Agents

Evaluation functions are used to enhance the intelligence and autonomy ability of agents and decrease the negotiation times between agents. This paper adopts the price computing function of RA and bid evaluation function of TA in reference [2]. A new task selection method of RA is used according to practical production of company.

Task selection method of RA. When a lot of tasks are in there scheduling negotiation processes, there may be several tasks compete for a same device. Then the RA which represents the device they are competing for will decide which one to process first according to task selection method.

First it will compute the priority of the step of tasks considering of the earliest start processing time, the latest finish time, the processing duration time, the step number of this step and the deposit cost. Here defines some variables as follows:

o_j : Step *j* of task *i*;

$Lfinish_j$: The latest finish time of o_j ;

$pdur_j$: The processing duration time of o_j ;

$Stepno_j$: Step number of o_j ;

$cost$: The deposit cost of task *i*;

$durtime$: The due time of task *i*;

N : The quantity of steps of task *i*.

$level_i$: The level depth of task *i* in product.

$Estart$ is the finish time of last step of this task.

The latest finish time of step i of task is calculated using Eq. (1).

$$Lfinish_i = duetime - \sum_{j>i}^{N} pdur_j \qquad (1)$$

The priority of step *j* of task i is calculated according to Eq. (2).

$$priority_{ij} = Lfinish_j + pdur_j * 0.6 - Stepno_j * 0.3 + cost_j - level_i * 0.4 \quad (2)$$

Eq. (2) takes due time, remanent processing time, and the deposit cost of task and the depth into account. It is because according to E/T scheduling, tasks should not be finished too late to meet the due time, and it also should not be finished too early to increase the deposit cost. According to practical production, components with high level need to be assembled and combined with other components, so the level of components is taken into account in priority's computing. The higher the lever is, the higher the priority is, and thus the earlier the task will be selected and processed to reduce the waiting time of assembly.

3.3 Exception Scheduling

In practical production of company unpredictable situations such as accept pressing order, rejected part, device failure may occur. The proposed MAS is not only able to scheduling under normal situations, but also able to deal with unexpected incidents betimes when such events appear.

Urgent scheduling strategy. The urgent events maybe find urgent task or some task failed to be completed in time and so on. Here urgent task means task which is about to reach its due time because of some unexpected reasons, such as it comes from a pressing order or processing be delayed as the device which assigned to it was down.

If urgent tasks arrivals, the system only needs to deal with the urgent task without changing the original scheduling strategy. The urgent task scheduling strategy is as follow.

Step 1: Assign the task a special flag or higher priority which will exist during its lift time.

Step 2: If $T\_duedate\text{-}T\_process > T\_limit$, then take the task into the under scheduling or will be scheduled task lists to participate in the normal scheduling process, the urgent case has the prior privilege under same condition.

Step 3: Else appoint some resource agent to process this task according to knowledge repository. The appointed resource agent puts the task in its booking task queue directly and processes it firstly.

Machinery failure scheduling strategy. If device conks and the current processing part are damaged, the product must be reproduced, then a new task is created which is considered as urgent case by system and would be processed under urgency scheduling strategy. Agents will continue processing the damaged work piece under normal scheduling strategy and consult scheduling with other devices. Other tasks in the failure machine's booking queue will quit and reschedule under normal scheduling strategy.

Accept pressing order. If pressing order which demanding due time is signed during scheduling, then assign a urgent symbol to these parts and process under urgent scheduling strategy.

Order cancelled scheduling strategy. When an order is temporarily canceled, the system will stop the scheduling of all the tasks in this order and delete them from the MAS. And the delete operation will not impact on other tasks' scheduling process.

Task delay scheduling strategy. When certain Task Agent predicts the task would delay, it will increase its priority to be processed under urgent scheduling strategy to make sure it can meet the due time.

Indeed, during the scheduling procession there could be many other unpredictable, sudden accidents occurred, for example, client change the order request, material can not arrive on time and etc. These situations could follow the similar solution strategy as above.

4. EXPERIMENTAL EVALUATION

The scheduling experiments carried out in this paper uses one hundred part processing tasks of the company. Every task has different working procedures and the restriction of arts and crafts, processing time, the restriction of task are the same as practical production.

Schedules for 35 tasks and the devices assignment are shown as Fig.5 and Fig.6.

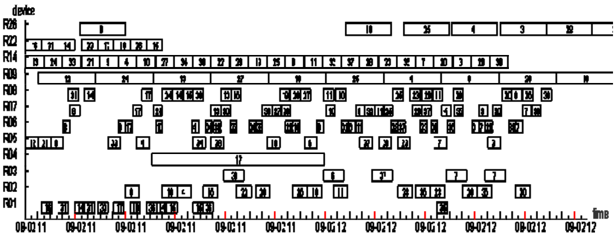


Fig.5. Device-Oriented Gantt chart

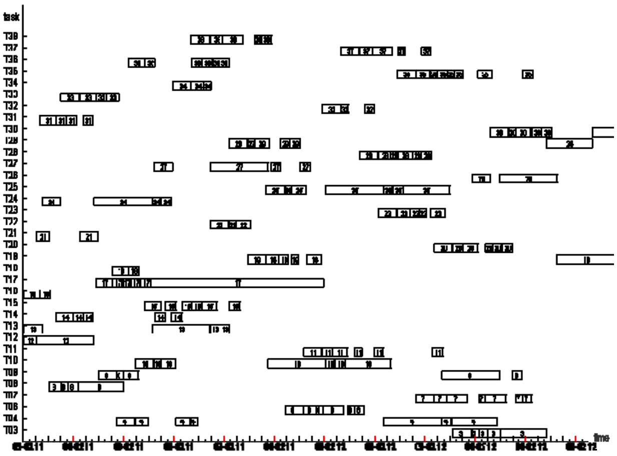


Fig.6. Task-Oriented Gantt chart

The evaluation parameters used in this paper are:

(1) Total processing time *makespan* is the time section between the start processing time and the end processing time of a set of tasks.

(2) Average processing time t_{ave} .

$$t_{ave} = \frac{1}{N}\sum_{j=1}^{N} t_j \qquad (3)$$

Here t_i is the processing time of task *j*, *N* is the quantity of tasks.

(3) Average device utilization U_{ave} .

$$U_{ave} = \frac{1}{M}\sum_{i=1}^{M} \frac{t_{ri}}{makespan} \qquad (4)$$

Here t_{ri} is the sum of processing time of device *i*, *M* is the quantity of devices.

The graph of average device utilization and average processing time with task number gained by the experiments is as shown in Fig.7.

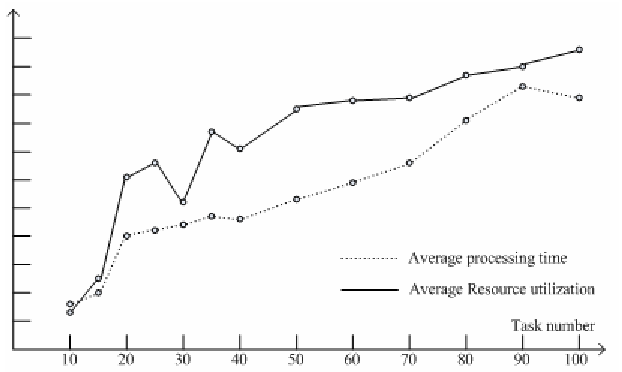


Fig.7. The result of experiments

We can see from Fig.7 that along with the increase of task number, the average device utilization and average processing time are increasing step by step. In practical production, the parts to be processed are not always use some devices concentratively, some devices are hardly used during the total processing procedure, but there existence influence the average device utilization of the experiment, so the curve of average device utilization is a little undulation, but the total increasing trend is not influenced. According to E/T scheduling, the premature end of a task may lead to increase of deposit cost, so prolong processing time of tasks without postponed the deliver time can reduce the cost of production and gain more profit.

We also conducted many experiments to prove that the proposed MAS is able to handle exception situations such as new task arrival, resource breakdown, old order be canceled. The results show that in these exception cases, the MAS can make decisions timely and correctly thus the influence on total performance of scheduling will be reduced to the minimal limit.

5. CONCLUSIONS

This paper designed a discrete part manufacture scheduling framework based on multi-agent. Experimental results show that compared to the old manual scheduling method, the proposed system is able to arrange processing order for tasks reasonably and efficiency considering of many factors, the device utilization is increasing along with the increasing of task number. So the cost of production is reduced, and some manual cockamamie works and unreasonable results are avoided. Because of many productions have working procedures and restrictions, the system designed in this paper is universal for discrete processing companies.

REFERENCES

[1] Li yan, *Influences of Process Planning on Scheduling Results*, China Mechanical Engineering,2000.

[2] Zhanjie Wang, Yanbo Liu, “A Multi-Agent agile scheduling system for job-shop problem. ISDA’06. Published” by *IEEE Computer Society*, page(s) 679 (EI).

[3] Zhanjie Wang, Yuan Wang, “Multi-agent Dynamic Scheduling Control Framework,” *Micro Computer Information* (EI), 2005, 21(1), pp.175~177.

[4] Jin hong, “An Integrated Design Method of task Priority,” *Journal of software,* Vol.14, No.13, 2003, pp.376~382.

[5] Boccalatte A, Gozzi A, Paolucci M, “A Multi-Agent System for Dynamic Just-In-time Manufacturing Production Scheduling” [J]. *IEEE International Conference*, 2004, vol.6, pp.5548~5553.

Zhanjie Wang is a Professor and a Director of AMD64-SUN united Lab of Computer Engineering Department, Dalian University of Technology. He has published a book named “64 bit Micro processor and programming”. His research interests are in distributed parallel processing and network security.

Consistency Management in Mobile Agent Systems

Guoling Hu
School of Information Science and Engineering, East China University of Science and Technology
Shanghai, 200237, P.R. China
Email: guoling.hu@gmail.com

ABSTRACT

The mobility of mobile agents has the potential to provide a convenient, efficient and robust programming paradigm for distributed applications. But it also creates new patterns of interactions and new conditions for distributed algorithms that are traditionally not issues for them, particular the supporting for the sophisticated coordination and close cooperation among different agents. In this paper, we focus on formalization and enforcing correctness properties of the coordination and interaction among agents. A formal agent model and its execution model were firstly presented. In the model, each agent is associated with a shadow agent to coordinate and enforce the serialization of its execution. Secondly, we proposed a formal model of agent execution control and its correctness criteria.

Keywords: Mobile Agent, Transaction Processing, Concurrency Control

1. INTRODUCTION

Mobile agents are programs that can move through a network under their own control, migrating from host to host and interacting with other agents and resources. While these mobile, autonomous agents have the potential to provide a convenient, efficient and robust programming paradigm for distribute applications, the mobility of these entities also creates new patterns of interactions and new conditions for distributed algorithms and applications that are traditionally not issues for them. The fundamental building blocks of many traditional distributed algorithms rely on assumptions, such as data location, message passing and static network topology. The movement of both clients and servers in mobile agent systems breaks these fundamental assumptions. Simply forcing traditional solutions into mobile agent systems changes the dynamic nature of their environments by imposing restrictions, such as limiting agent movement. As a result, new effective and efficient techniques for solving distributed problems are required to target mobile agent systems.

Many current researches on agent-based models [1,2,3] focus on aspects of intelligence and security issues and few researches related to the concurrency control problem of agents. In this paper, we use mobile agents to model long-live computations which own local resources and access data at remote systems. Besides their own data, agents access two types of data. Firstly, they access data owned by other cooperative mobile agents. Secondly, agents query and modify data that belong to remote systems which may include persistent data stores such as database. It is important to make sure that the concurrent execution of multiple agents does not violate the consistency of the data accessed while agents interacting with other agents and resources. In the paper, we firstly present a formal agent model and its execution model. In the model, each agent is associated with a shadow agent to coordinate and enforce the serialization of its execution. The shadow agent is itself an agent created by host environments and must be moved together with their associated transaction agent through the network. Secondly, we proposed a formal model of agent execution control and its serialization criteria.

The remainder of the paper is organized as follows: Section 2 presents the formal model of the mobile agent system. Section 3 presents the execution control in a mobile agent environment. Finally in section 4 we draw our conclusions.

2. FORMAL TRANSACTION AGENT MODEL

2.1 The Formal Agent Model

An agent is an active object encapsulating the state, behavior and location of a computation. It can be formally modeled as a 6-tuples array (AID, D, M, L, T, SAID). Where AID is the unique identifier of the agent; D is a set of values indicate its states; M indicates the behavior of an agent. It is a set of triples <MID, Type, State> where MID identifies a method of the agent, Type indicates the type of the method. In the model, we differentiate two types of methods: local and remote methods. Local methods are methods that only manipulate agent's own states. Remote methods are methods accessing resources of other mobile agents or resources of the remote system such as databases where the agent is transmitted. State={RD, SB, R, C, A} indicates the state of the method. As shown in fig. 1, the method may have four kinds of state: Ready (RD), Submitted (SB), Running (R), Commit (C) and Aborted (A). When a method is invoked, it enters into submitted state and moves to the ready state when it is actually ready for execution. In terms of remote methods, the shadow of an agent may delay the ready state of a method to wait for another agent or system resources to ensure correctness. When the execution starts, the method enters into the running state and remains in this state until its completion. Upon completion, the method may be either committed or aborted. If committed, the results of the method become permanent and the method enters into the committed state. If aborted, the results of the method are not recorded, and the method has no effect on data.

L indicates agent's location, which is the context of its execution; T indicates agent's type. Its value can be TA or SA. There are two types of agents populating in the mobile agent system:

- Transaction Agent (TA): The only agent that actively interacting with the data of other agents and remote resources.
- Shadow Agent (SA): An agent that associating with TA to coordinate its execution. Their main responsibilities are to coordinate with each other to enforce the serialization of the executions of their associated transaction agent. They are created by host environments and must be moved together with their associated TA through the network. Local variables of an SA include the order of method invocations, the history of the execution of the remote methods of its corresponding TA and the execution state of each method.

- SAID is the AID of the shadow agent associated with the transaction agent. For shadow agent, its value is NULL. If the associated shadow agent of the transaction agent has not created, the value for this item is also NULL.

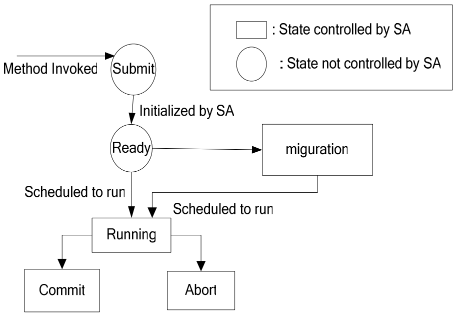


Fig.1. States and controls of the transaction agent method

2.2 The Execution Model of Transaction Agents

The execution model of TA is depicted in fig. 2.The host environments on each site are responsible for hosting the mobile agents and provide APIs to access resources. They coordinate the executions of TA and SA as they move through the network to ensure the correctness of the interleaved execution of agents.

Each host environment has two parts:

1 DGTM is distributed transaction manager. It located on top of each remote system. Each DGTM coordinates the submission of agents to it's system. Each DGTM receives database operation requests from the various shadow agents, schedules them to control concurrency and inter-agent synchronization and submits them to the corresponding LTMs.

2 LTM is pre-existing local transaction managers at each database system. It controls the execution of the basic database operations submitted to ensure the ACID(Atomic, Consistent, Isolated and Durable) properties of local database operations.

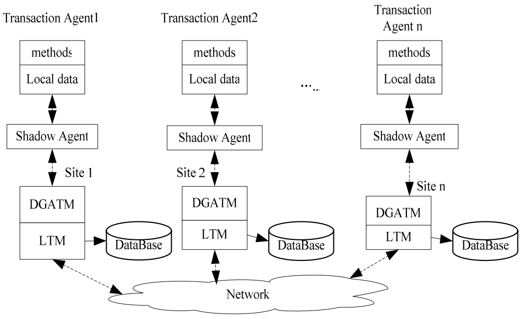


Fig.2. Execution Model of the Transaction Agent

3. AGENT EXECUTION CONTROL

Transaction agents may interact with each other when accomplishing a common goal. Moreover, the execution of a transaction may interfere with the execution of other transactions when they both access same data concurrently. Attributing to agents the properties of traditional transactions is too restrictive. Since agents can see the intermediate results of execution of other agents by accessing their local data, the execution of each agent is not isolated from the execution of other agents. Furthermore, the execution of an agent is not necessarily to be atomic, since while a method of the agent may be aborted, some other method may be accepted. In this section, we define what is a correct execution of an agent as well as a correct execution of a number of concurrently executing agents.

3.1 Well Structured Agent Execution

The interaction between two transaction agents can be controlled by explicitly defining contact points in the execution of a transaction agent where other transaction agents are allowed to observe its partial results.

Definition 1(Contact Point) A contact point *CP* of an transaction agent *TA* is a triple $(CP_{start}, CP_{end}, RM)$ where $RM=\{(A_i , M_j)\}$ is a set of pairs of transaction agents and their remote methods, $CP_{start}=(M_s, States)$ and $CP_{end}=(M_e, States)$ are pairs of methods and their corresponding controllable states of *TA* which allow members of *RM* to be executed between CP_{start} and CP_{end} of TA.

When methods of agents interact with each other, their execution order may affect the final state of data. For example, if they access the same data item and one of them update it. Two methods conflict if the order of their execution affects the final state of data. The execution of an agent is actually a partially-ordered sequence of method executions and contact point events.

Definition 2(Agent Execution) An agent execution *AE* of a transaction agent *TA* is a tuple $(E, <)$ where *E* is a set of events; $<$ is a partial order of events on *E*.

An event is a contact point in the execution of the agent or a pair (M, S) where *M* is a method of the agent. *S* is a state of *M*. The partial order < is such that for all conflict methods M_i and M_j, either $(M_i, R) < (M_j, R)$ or $(M_j, R) < (M_i, R)$, where *R* stands for the running state of the method. Thus, the execution of an agent is a partially-ordered sequence of method executions and contact point events.

An interval of an agent execution *AE* is defined as a subsequence of *AE* that includes exact events between CP_{start} and CP_{end} of a contact point in *AE*. A method of another agent interleaves with an interval if it starts its execution before an event e_1 and finishes the execution after another event e_2 of the interval.

The partial order must respect the structural dependencies among the methods of an agent as expressed in the following definition.

Definition 3 (Well Structured Agent Execution) An execution of an agent is well-structured in terms of a set *D* of structural dependencies if the partial order < of its events does not violate any of the dependencies in *D*.

3.2 Schedule Correctness

A schedule is an interleaved execution of methods of a set of agents. Formally:

Definition 4 (Schedule) A schedule $(E, <_c)$ is a executions $AE_i=(E_i, <_i)$ $(1 \leqslant i \leqslant n)$ of a set of agents $\{A_1, A_2 ,\ldots,A_n\}$ where $E= \cup E_i$ and $<_c$ is a partial order such that:

(1) if for any events e_k and $e_j \in E_i$, if $e_k <_i e_j$ then $e_k <_c s_j$

(2) for any two conflict methods $M_i \in E_i$ and $M_j \in E_j$ $(1 \leqslant i, j \leqslant n)$ of different agents , either $(M_i, R) <_c (M_j, R)$ or $(M_j, R) <_c (M_i, R)$.

The first condition states that, the interleaved execution of a set of agents preserves the execution order of each of the agents. The second condition imposes a relative order between the non-compatible methods of two different agent executions.

A method M_2 of a transaction agent is directly depends on a method M_1 if $(M_1,R) <_c (M_2,R)$.The depends on relationship is defined as the transitive closure of the directly depends on relationship.

The projection of a schedule S on the local data of a remote system DB_i is the schedule excluding from S all but the basic methods on data of DB_i. Formally:

Definition 5 (Projection of a schedule) A projection of a schedule $S=(E, <_c)$ on data of a remote system DB_i is the schedule $(E_p, <_p)$ where $E_p \subseteq E$ and includes only events operating on the data of DB_i. $<_p$ is a partial order such that if for any events e_k and $e_j \in E_p$, if $e_k <_c e_j$ then $e_k <_p s_j$.

To enforce that the concurrent execution of agents does not violate the consistency of data, the corresponding schedule must be conflict-serializable, that is, conflict equivalent to a serial schedule. Two schedules are conflict equivalent; if they consist of same events and the order of conflicting operations are the same.

Definition 6 (Relatively Serial Schedule) A schedule $(E, <_c)$ is a relatively serial schedule if,

(1) The execution of each transaction agent maintains data consistency if executed alone;
(2) For all pairs of agents T_i and T_j, if a method M of T_i is interleaved with a interval of T_j starting with contact point(CP_{start}, CP_{end}, RM), then at least one of the following is true: (a) M does not depend on any event of the interval and no event of the interval depends on M. (b) $(T_i, M) \in RM$.

Definition 7 (correct schedule) A schedule is correct if (1) all the execution of its agents are well-structured, and (2) it is conflict equivalent to a relatively serial schedule.

As the following theorem shows, to ensure that a schedule is conflict equivalent to some serial schedule, it suffices to ensure that its projections are conflict serializable with consistent orders:

Theorem 1. If each remote system projection of a schedule $S=(E, <_c)$ is conflict equivalent to a relatively serial schedule, and the order $<_c$ consistent with the serialization orders assumed by each projection, S is conflict equivalent to a relatively serial schedule.

Proof. Suppose $S=(E, <_c)$ is not a conflict equivalent to a relatively serial schedule $SC=(E, <_{rs})$, then there exist conflict events e_i and $e_j \in E$ operating on data A of DB_k, $e_i <_c e_j$ but $e_j <_{rs} e_i$. Suppose the projection of S on DB_k is $S_k=(E_k, <_K)$, then e_i and $e_j \in E_k$ and $e_i <_k e_j$. Since S_k is conflict equivalent to SC and there is $e_j <_{rs} e_i$ in SC, then there should have $e_j <_k e_i$ in S_k, a contradiction.

4. CONCLUSIONS

In this paper, a formal execution model of mobile agents and the scheme for ensuring correctness of the concurrent execution of agents were presented. In the model, the structure of an agent is defined through dependencies between the execution states of its methods. To enforce the structural properties of a mobile agent and the control of the interaction of a mobile agent with other agents and remote resources, each transaction agent is attached with a shadow agent. The shadow agent maintains all states of all methods of its transaction agent and accomplishes the coordination job through a well-defined small set of primitives.

REFERENCES

[1] Zwierko A., Kotulski Z, “Mobile Agents: preserving privacy and anonymity,” *Proc. Intern. Workshop on Intelligent Media Communicative Intelligence*, Warszawa, September 2004.

[2] Sander T., Tschudin Ch. F, “Protecting mobile agents against malicious hosts,” in *Mobile Agents and Security*, volume 1419 of LNCS, Springer-Verlag, 1998.

[3] Vigna G, “ Cryptographic traces for mobile agents,” in *Mobile Agents and Security*, volume 1419 of LNCS, Springer-Verlag,1998.

[4] P. Attie, M. Singh, A. Sheth, and M. Rusinkiewicz, “Specifying and Enforcing Intertask Dependencies,” in *Proceedings of the 9th International Conference on Very Large Database Systems*, pages 134-144, 1993.

[5] D. Georgakopoulos, M. F. Hornick, and A. P. Sheth, “An Overview of Workflow management: From Process Modeling to Workflow Automation Infrastructure,” *Distributed and Parallel Databases*, 3(2), 1995.

Guoling Hu is a lecturer in School of Information Science and Engineering, East China University of Science and Technology, She got doctoral degree from Huazhong University of Science and Technology in 1996 with specialty of computer science. Her current research interests are in distributed parallel processing, multi agent techniques, modern database technology.

Study and Implementation of the Power Dispatch Ticket System Based on the Multi-agent*

Sufang Chen
Industrial Center,Shenzhen Polytechnic,Shenzhen, Guangdong 518055,China
Email: Chensufang@ szpt.net

ABSTRACT

The existing automatic generation systems for the dispatch ticket of electric power grid are mostly off-line single agent systems, and they need to be improved in intellectualization, safe examination, and the execution of flow verification. According to this actuality, this paper proposes to establish operating instruction template and the maintainable mechanism to solve universalization problems of description, study and consequence of the operating instruction. Furthermore, the paper proposes to develop and implement a new intellectualized dispatch ticket system of electrical grid by using the multi-agent system. Such new system is maintainable for users. Integrating with the management information system and sharing real-time data with the SCADA system, this system includes kinds of applicable functions, such as automatic generation and manual generation of the ticket, simulation preview, execution and management.

Keywords: Muti-agent System, Electric Power System Dispatching, Operation Orders, Cognition Model, SCADA System, Security Proof

1. INTRODUCTION

The power control center is the key department of electric power’s production and running. Along with the rapid development of modern technology, the functions of dispatch automation system and dispatching communications system are more and more formidable. They complete many tasks, such as substation’s survey and control, energy collection. But the massive switching tickets’ executions cause dispatch personnel's cerebrum to be in the continuous and tense condition every day. Any slight negligence can bring about the great accidents. To reduce the accidents and guarantee the power system’s safe operation, the research of power switching sequence ticket system is significant and essential.

The power switching sequence ticket system based on the multi-agent body derives from dispatch personnel's rich experience, summarizes all kinds of operating rules, and has achieved accuracy, clearness and rapidness. On one hand, this extricates the dispatch personnel from the complex work, so they can focus on the in-depth problems in the power grid security and the economical running. On the other hand, it can make up the insufficiency caused by the inexperienced dispatch personnel in operation.

The switching sequence ticket system proposed in this paper, mainly uses the Client/Server pattern to establish the multi-user database, and may support the user sharing unified central database. The application software is of object-oriented design, developing and implementing a new intellectualized power switching sequence ticket system. The system is combined with the management information system, and shares real-time data with the SCADA system. The new system consists of the practical functions of automatically ordering tickets, manually ordering tickets, simulated preview, flow checking, executing and managing tickets. This system is maintainable for its users. After the operator issues the operation order by choosing the "the operation object + operation duty" on the software surface, the reasoning machine will start to carry on the consequence, and then the needed switching ticket will be generated automatically.

2. SYSTEM OVERALL STRUCTURE AND MAIN CHARACTERISTICS

The intelligent power dispatch command-sheet system is an automatic command-sheet generation system based on the real-time status of power grids, graphical mode, and expert system inference mode. The system is a new application in the artificial intelligence field. While power grids operate in an increasingly complex manner, higher security is required for power grids. Therefore, it is important to computerize the operations of power dispatch. In this case, this system offers an ideal solution. It enables the following functions:

- draw and edit power grid topologies,
- display the real-time status of power grids,
- generate item-by-item order tickets, synthesis order tickets and switching tickets automatically,
- revoke typical tickets and history tickets, manage system authority logs.

As shown in Fig. 1, the system consists of five modules: graphics management module, operation-sheet management module, sheet-generating rule management module, system management module, and database.

2.1 System Management Module

The system management module is composed of two parts, namely, the operation jurisdiction management and the user information. It carries on the management to the user. The system has three jurisdictions: system administrator, standard user and common user.

2.2 Graphics Management Module

The graphics management module is in charge of planning, modifying and managing the main-line diagram, completing the switch between substations and refreshing the equipments’ state on real-time.

2.3 Database Management Module

The database management module is primarily in charge of managing the five databases of the system: the substation information database, the knowledge library of rules and regulations, the real-time dynamic database, the user jurisdiction database and the switching tickets database.

1) Substation Information Database

The substation information database is used for storing the comprehensive information of substations’ equipments,

*This project is supported by Shenzhen Science and Technology Bureau.

including the equipments' basic parameters, states, topology connection information, *etc.*. These parameters constitute all the known data, which the system needs to create switching sequence ticket.

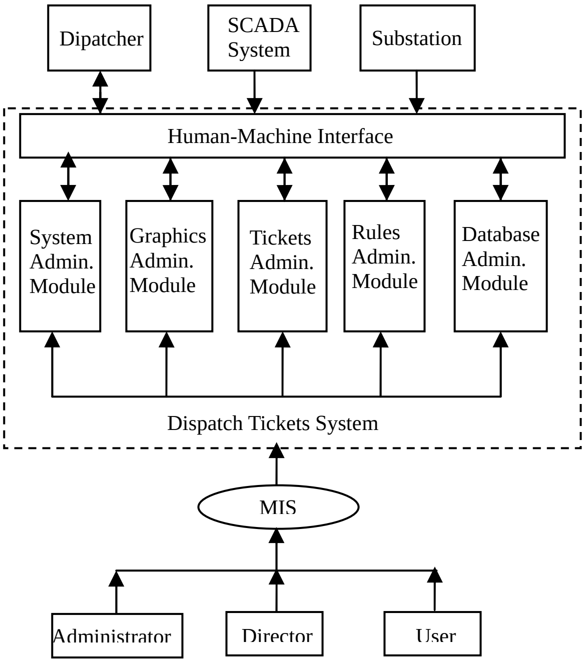


Fig.1. System architecture

2) Rules and Regulations Knowledge Library
The system has constructed the operation knowledge database separately according to operating regulations knowledge of kinds of electrical equipments of substations, mainly including: the generatrix operation knowledge library, the two-volume transformer operation knowledge library, the three-volume transformer operation knowledge library, the circuitry operation knowledge library and the voltage transformer operation knowledge library, and so on.

3) Real-time Dynamic Database
The real-time dynamic database is used as a data stopover station. It is used for storing the real-time data, which is transmitted from the SCADA system.

4) User Jurisdiction Database
The user jurisdiction database is used for storing user names and passwords of all system administrators and standard users. The user can add, delete or revise the datum in the database through the human-machine interface.

5) Switching Tickets Database
The operation ticket database is used for storing the operation ticket. After creating the switching ticket, if choosing saving, the ticket will be stored into the database. Apart from the operation sentences, the human name, equipment type and saving time are also saved into the database. The tickets stored into the database have two purposes, one as typical tickets, the other as document files.

2.4 Tickets Management Module

The tickets management module is in charge of kinds of work of switching ticket management, including generating, revising, saving, inquiry, deleting, preview, printing preview and printing. Switching tickets are generated via three ways: automatically ordering tickets, manually ordering tickets and transferring typical tickets.

2.5 Multi-agent System of Switching Tickets

The multi-agent system of switching ticket, the real-time SCADA system and the MIS share the same data definition and the uniform data structure in the aspects of power grid primary equipment, secondary equipment, flow management, operation terminology description, and so on. The multi-agent system receives the SCADA real-time data timely to refresh the power grid's running status in the database. The database provides each agent with the same data as the primary data structure to guarantee that each agent runs effectively.

Fig. 2 shows the structure of multi-agent system of switching tickets.

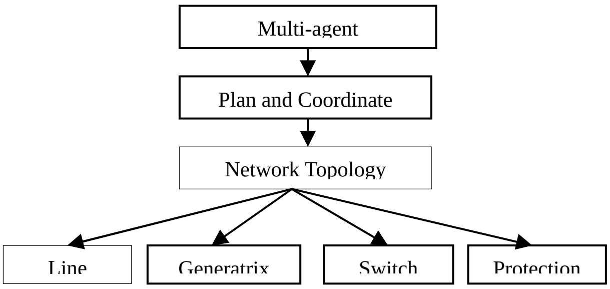


Fig.2. The Structure of Multi-agent System of Switching Tickets

3. THE COGNITION MODEL OF OPERATION TICKET

3.1 The Classification of Dispatching Order Tickets

The county dispatching order tickets for dispatching operation duty tickets, used in dispatching several even dozens of transformer substations generally, the running status and the logic relations which the dispatching order involves are much more complex than switching operation which has strict logical relations. According to the regulation of the power dispatching order tickets, it is mainly divided into two kinds: one kind is the synthesis order ticket, mainly used to the operation in a station, another kind is the roster order ticket , mainly used to the operation in the multi-station.

According to the criterion of the electric power order ticket, the dispatching order tickets are divided into two kinds: the regularity dispatching order tickets and the irregularity dispatching order tickets. For the former, under the operation duty explicit premise, the project's filling in order and its details have certain standards, therefore we can realize automatic production by carrying on the science classification of the operation and using the corresponding algorithm. The latter refers to the operation in one station and the operation in multi-station contact line. As to this kind of tickets, realizing the automatic production has certain order complexity because of using different devices in each operation even if under the same conditions.

3.2 References Multi-Interstation Operation Level Conceptual Model

When the working way of the electrical network is changed by operation, many stations may be affected and there is more than one operation way under operating rules. In order to determine the best operation way, the system should be unified with the real-time tidal current.

The operation goal of the county dispatching order tickets may be divided into three kinds according to equipments: (1) stop or give power of the generatrix, (2) stop or give power of the transimission line, (3) stop or give power of transformer. Because the three kinds of equipment are physically connected mutually, any kind of equipment stoping, the power transmission operation may causes other equipment lose electricity. To solving the problem, we must seek new power source for the equipment which has lost electricity, and the power may be in its factory or in others.

Therefore, we can establish a conceptual model for the factory interstation operation level by using the line, the generatrix and transformer to summarize the running status of the system. As shown in Fig. 3. Getting or losing electricity can be known from the running status of the line, the generatrix and the host change.

3.3 Only Stands the Operation Level Conceptual Model

Regarding a concrete substation, its sequence of operation is generally relatively fixed. Operating instructions use the pattern of “rule group + rule skeleton + rule body + rule sub-body”. The rule is classificated by operating equipments each time and then establishes the rule group. Corresponding to each operation duties of one substation which is showed in Fig. 3. we can draw each rule skeleton in the separate rule group. All the operating units in the rule body is the switch collection, which means transforming separately among the four switch status, The rule sub-body is used to describe the related equipments the node needed to complete a task. The rule sub-body may also carry on the segmentation again until to a practical movement equipment. Fig.4 shows the operation level conceptual model.

Equipment conditions in a factory station are generally divided into hot spare, cold standby and the overhaul. The concrete operation form generally transfers from one condition to another, which specifically considers the equipment changes of states as well as the operation sequence, to form the correct operation sequence and the operation terminology. As shown in Fig. 5, for a line overhaul rule model.

4. HIERARCHY-BASED KNOWLEDGE INFERENCE

4.1 The Basic Structure of Scheduling Operation Ticket Expert System

Gain and expression exert knowledge are bottlenecks to develop expert system. In the operation ticket automatic production system, knowledge expression of operating instruction is key point and difficulty in the performance history. Operating instruction in scheduling operation not only manifests the electrical network topology recessively, but also manifests the scheduling operation order,, both are the base of operating instruction design. The operating rules are the core of expert system, and complete the check automatically function with inference engine.

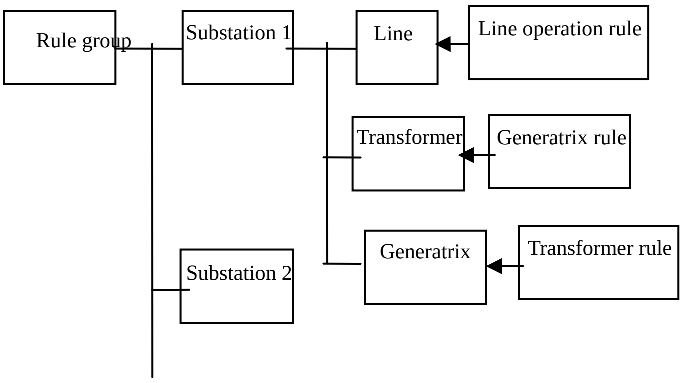


Fig.3. Inter-substation operation level conceptual model

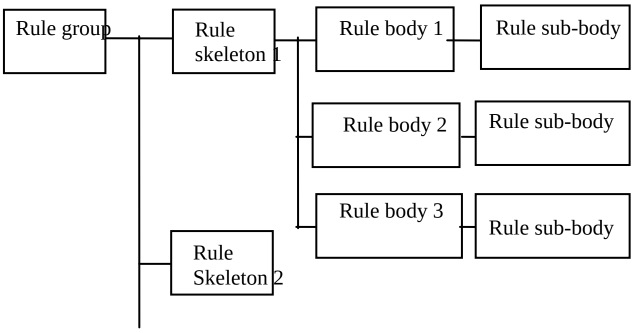


Fig.4. Mono-substation operation level conceptual model

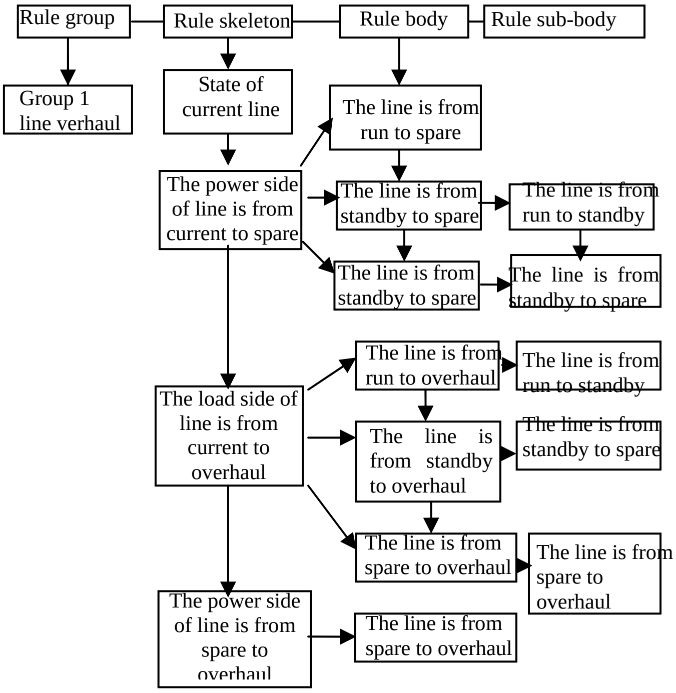

Rule group
Rule skeleton
Rule body
Rule sub-body
Group 1 line verhaul
State of current line
The line is from run to spare
The power side of line is from current to spare
The line is from standby to spare
The line is from run to standby
The line is from standby to spare
The line is from standby to spare
The line is from run to overhaul
The line is from run to standby
The load side of line is from current to overhaul
The line is from standby to overhaul
The line is from standby to spare
The line is from spare to overhaul
The line is from spare to overhaul
The power side of line is from spare to overhaul
The line is from spare to overhaul

Fig.5. Line overhaul rule model

Fig.6 is the basic structure of the operation ticket expert system, developed by this system. It makes the use of technology from expert system. In the expert system, dispatch personnel’s knowledge is devided into three big kinds: the network analysis status knowledge, the operating instruction knowledge and the regular knowledge against mistakes. The network topology knowledge saves the power topology sturcture. the operating rule knowledge is summarized and refined from the dispatch movement regulations, the operation rule and the dispatch personnel's experience.the regular knowledge against mistakes is similar to “five againsts” in the switching operation.

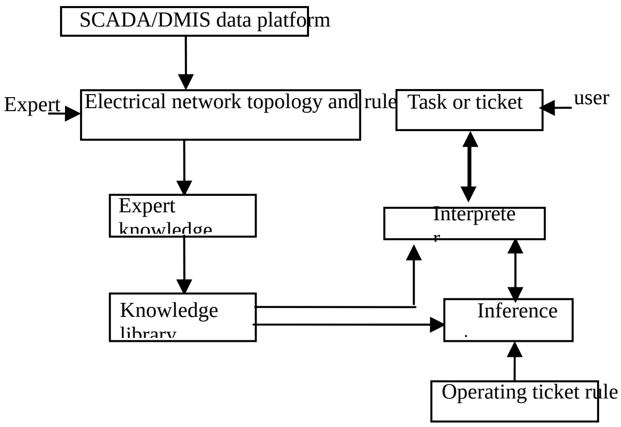


Fig.6. Basic structure of the operation tickets expert system

The system obtains the physical connection relations of electrical netwok, the switch and knife switch real-time state, real-time power flow information from the SCADA/DIMS data platform, and these information are used to renew the system database and the knowledge library as well as the network analysis status relations table needed by inference. The system obtains knowledge library needed by multi-station operation inference by using the neural network and the adjacency power matrix to indicate.

4.2 Hierarchy-based Knowledge Inference

A power grid topology consists of three layers: The first layer is the backbone, that is, the power plants and substations that are connected to each other through the lines of different levels of voltage; the second layer is the plant & substation layer that consists of power plants and the devices inside substations; the third layer is the device layer, that is, the relationship between devices. The power grid topology is hierarchical and the operations of power gird dispatch must follow strict operation sequence. Therefore, a hierarchy of “rule group + rule skeleton + rule body + rule sub-body” is used for knowledge inference, as shown in Figure 4. One rule group corresponds to one independent operation rule unit. A rule skeleton is used to describe the hierarchical relationship between nodes in an inferred network. One rule skeleton may correspond to multiple rule bodes because nodes are in different initial states. A rule body is used to describe the detailed inference rule or calculation rule of nodes. Each rule body corresponds to one rule sub-body. A rule sub-body is used to describe the related devices in the node to be operated for completing a task.

The rule description of the hierarchy is easy to understand, thus providing a natural and interactive expression means for knowledge acquisition. The operation task corresponding to a rule body of simple operations is basically indivisible. After some complex operations are divided, the rule body and its rule sub-body can be shared if the corresponding operation task is consistent with the rule body. In this way, the knowledge unit can be used for multiple purposes and correct operation sequence is ensured. Fig.7 shows the software implementation of this system.

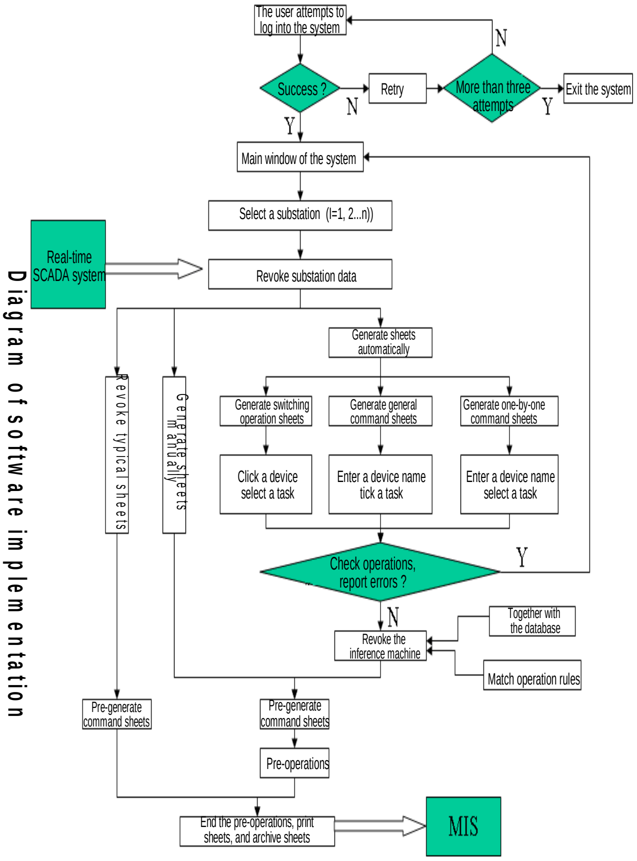

The user attempts to log into the system
Success ?
N
Retry
More than three attempts
N
Y
Exit the system
Y
Main window of the system
Select a substation (I=1, 2...n))
Real-time SCADA system
Revoke substation data
Diagram of software implementation
Revoke typical sheets
Generate sheets manually
Generate sheets automatically
Generate switching operation sheets
Generate general command sheets
Generate one-by-one command sheets
Click a device select a task
Enter a device name tick a task
Enter a device name select a task
Check operations, report errors ?
Y
N
Together with the database
Revoke the inference machine
Match operation rules
Pre-generate command sheets
Pre-generate command sheets
Pre-operations
End the pre-operations, print sheets, and archive sheets
MIS

Fig.7. Diagram of software implementation

4.3. Production Strategy Based On Intelligent Network Topology Relation

In order that guarantee the operation ticket automatic production system can show a ticket correctly, the correct electrical network topology must be established to describe the electrical network equipment reciprocity and the network active status. The related table must also established between the switch, the knife switch, the transformer, the generatrix, the contact line, the distribution wire and the switching equipment. Various tables realize bidirectional lamination search through connection field. For example, through the ”correlation equipment” of the knife switch table, we can find its upper switches generatrix, transformer and so on . The definition about the network topology relation has several good points, such as complete information, serching work nimble. Therefore, it can enhance the system’s running rate enormously and improve the system’s function.

After drawing up a good diagram, by clicking the chart revolving the chart we can carry on the real-time condition revolution work about every chart in the current station. The procedure may explain the real-time information which comes from the scene RTU through the serial port receive and the CDT terms of agreement. After renovating the real-time condition information, you may click on the chart automatic chart topology and colored chart, the procedure will produce the current topological information automatically, analysis the gains and losses of the electric network, and then label colors in electrical network graph.

5. SYSTEM REALIZATION AND APPLICATION

This system uses VC++, SQL server as the development tools. By using the database technology and the object-oriented programming technology , it achieves the electrical network scheduling operation ticket expert system. The system operation is divided into the automatic production of the item-by-item order tickets, the synthesis order tickets and the switching tickets. The automatic ticket generation is the core part of this system. Different form of operation tickets has different inference principle.

As for the switching ticket, what we should do is to select relative operation equipment and then to choose the relative operation order of the equipment. And for the synthesis order ticket and the item-by-item order ticket, we only need to input the operation order in the corresponding dialogue box. After inference, they all can transfer into the edit box of the correlation ticket form, then the automatic produced sentences will be displayed on it. Of course, we can do some amending work for the sentences.

After the operation ticket has been produced, any forms can carry on the simulation preview. The entire current operating process can be seen intuitively in the electrical network graph and the possible error operation can be discovered promptly by the simulation preview. The intelligence ticket interface is shown in Fig. 8. At present, the system is running in the power distribution center in Xinmi of Henan province.

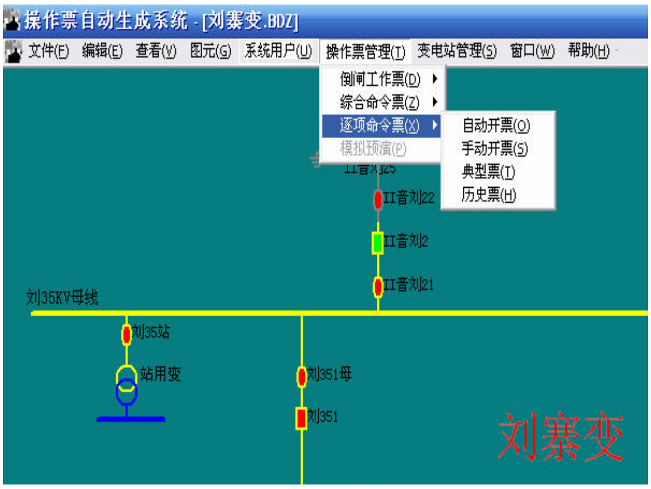


Fig.8. The intelligence ticket interface

6. CONCLUSIONS

The paper studies and implements power grid switching sequence ticket system based on the multi-agent. In the practical application, the efficiency of ordering ticket is remarkably enhanced by automatically ordering ticket and the direct-viewing simulated preview on the graphics. The practice has proved that the power grid switching sequence ticket system based on the multi-agent obviously surpasses the single agent in the aspects below: the ability of solving actual problem, the whole cooperation control and the information sharing.

REFERENCES

[1] J. Zhang, Y. L. Zhu, D.Li, "Situation ang prospects of order sheet expert system research," *Information on Electric Power*, vol. 1, pp. 61-64, Jan. 2002.

[2] L. Tang , B. M. Zang, H. B. Sun,et al. "General congnitive models of power network in operating command expert system," *Automation of Electric Power Systems*, vol. 22, pp. 6-9,21,Nov. 2001.

[3] M. L. Lin , G. Li , L. J. Qian ,et al. "A general intelligent system for electric power network operation order sheet, " *Relay*, vol2. 1,pp. 39-42,Dec. 2001.

[4] J. L. Lu, R. Li , Y. Liu,et al. "Command Tickets of Local Power Network Based on Dynamic Data Exchange,".*Journal of North China Electric Power University*, vol. 2,pp. 25-27,Feb. 2002.

[5] W. J. Liu , L. J. Qian, W. X. Liu ,et al. "Research and Practice of Order Sheet Generation System in Power Load," *Dispatch.Electric Power Science and Engineering*, vol. 3,pp. 34-38,Mar. 2003.

[6] W. Zhang, G. Chen , G. W. Zhang ,et al. "A Survey of Dispatching Operating Order System based on Expert System in NCC, " *Electric Power*, vol. 12,pp. 80-82,Dec.2003.

[7] X. M. Li , W. Chen , P. Wang , "Design and Implementation of a Muti-Substation Calling up On-line Order Forming System, " *Automation of Electric Power Systems*, vol. 5,pp. 71-74, May. 2004.

[8] L. C. Wang, Y. L. Jiao, "Order sheet expert system with better maintainability and generality for power network operation, "*Electric Power Automation Equipment, vol.* 2,pp. 44-47,Feb. 2004.

[9] S. Su, X. J. Zeng ,et a1. "Research on versatilegeneration method of switching orders for substations, " *Power System Technology*, vol. 14,pp. 14-18,Jul.2004.

[10] Q. Liu , M. Zhou , G. Y. Li et al. "Power system dispatching order generating system with calculation and analysis functions, " *Powe System Technology*, vol. 7,pp. 68-73,Jul. 2005.

[11] Q. Y. Li, W. Guo, "Identification on connection mode of intelligent dispatching operation instruction system," *Jiangsu electrical engineering*, vol. 25, pp. 54-56, Jan. 2006.

Su-Fang Chen received her B.S. and M.S. degrees in electrical engineering from Wuhan University of Technology in 1986 and 1989. She has been a professor in Shenzhen Polytechnic since 2001, and before that she worked at Hubei Power Dispatching Center for twelve years as an engineer. Her research interests include intelligent control , expert system and its applications in power systems. She has led over 6 national and provincial researching projects and published over 10 papers in international and domestic publications.

Holon Based Self-Organization Evolution in MAS

Jian Gao, Wei Zhang
School of Computer science and technology, Yantai University, Yantai, 264005, China
Email: hnr-1@163.com

ABSTRACT

With the growing usage of the world-wide ICT networks, agent technologies and multi-agent systems are attracting more and more attention, as they perform well in environments that are not necessarily well-structured and fully knowledge. For an effective system behavior, we need actual structure and organization. But the organization of a multi-agent system is difficult to be specified at design time in the face of a changing environment. In this paper, we propose a self-organization method based on holon organization evolution in multi-agent systems that allows dynamic self-organization during run-time. It satisfies the need of software design in which the open and dynamic environment must be faced with.

Keywords: holon organization evolution, self-organization, multi-agent systems

1. INTRODUCTION

A multi-agent system (MAS) consists of a collection of individual agents, each of which displays a certain degree of autonomy with respect to its actions and perception of a domain. Overall computation is achieved by autonomous computation within agents and by communication among them. The capability of the whole MAS is an emergent functionality that may surpass the capabilities of all individual agents [1, 2]. An extremely useful feature reducing the complexity for the designer of MAS is that an overall task can be broken down into a varied sub-task, each of which can be solved by a specific agent problem solver.

Jennings notes that "the development of robust and scalable software systems requires autonomous agents that can complete their objectives while situated in a dynamic and uncertain environment, that can engage in rich, high-level social interactions, and that can operate within flexible organizational structures"[3]. Organizational structure is superincumbent and prior-confirmed in traditional method[10,11], which is difficult to realize in the face of an open and changing environment. Thereby study of self-organization is important.

Self-organization usually refers to the fact that a system's structure or organization appears without explicit control or constraints from outside of the system[4]. The system's dynamics modifies its environment, and the modifications of the external environment influence in turn the system, but without disturbing the internal mechanisms leading to organization. The system evolves dynamically either in time and space.

Several approaches to organizing large groups of agents utilize emergent or bottom up techniques [12,13] for self-organization. While there are certainly situations in which such methods are appropriate, time constraints may not allow the self-organization processes to unfold.

In order to adapt open and changing environment, we presented a self-organization method based on holon organization evolution in this paper. It allows dynamic reorganization during run-time. Section 2 introduces holon and holon organization, self-organization based on organization evolution in multi-agent systems is presented in section 3, section 4 is conclusion.

2. HOLON AND HOLON ORGANIZATION

Multi-agent systems represent a new problem solving paradigm, where the difficult specification at design time of how a problem should be solved, is all well come by the interaction of the individual agents at run-time and the idea is that the solution of a given problem emerges from this interaction.

Divide and conquer is a widely accepted problem solving paradigm in computer science. Here, a centralized problem solving entity and a protocol are needed. The contract-net protocol [5] is a widely accepted problem-solving model in based on the divide and conquer model, where the centralized problem solving entity (called the manager) separates the overall task into sub-tasks. The manager uses a bidding procedure (a first price sealed bid auction) to find the most appropriate decentralized problem solver for each of the sub-task. The manager again does the integration of the solutions of the sub-task into an overall solution. This procedure can be recursively nested. This is a kind of superincumbent hierarchical Organization in traditional pattern.

2.1 Holon

A holon is a self-similar or fractal structure that is stable and coherent and that consists of several holons as sub-structures. Many distributed problems exhibit an inherent structure and we need to mirror this structure in the structure of the relationship between problem solvers. For this purpose in a holonic multi-agent systems, an agent that appears as a single entity to the outside world may be composed of many sub-agents, many sub-agents may decide that it is advantageous to join into the coherent structure of a super-agent and thus act as single entity. We call agents consisting of sub-agents with the same inherent structure holonic agents.

In some cases, one of the already existing agents is selected as the representative of the holon based on a fixed election procedure. In other cases a new agent is explicitly introduced to represent the holon during its lifetime. Representatives are called the head of the holon, the other agents in the holon are called body. In both cases, the representative agent represents the shared intentions of the holon and negotiates these intentions with the agents in the holon's environment as well as with the agents of the holon. Only the head communicates with the outside of the holon. The binding force that keeps head and body in a holon together can be seen as commitments.

The set H of all holons in MAS is defined recursively [6]:

For each agent a , $h = (\{a\}, \{a\},\varphi)\in H$, i.e. every instantiated agent constitutes an atomic holon, and h = (Head, Subholons, C)$\in$H, where Subholons$\in 2^H \backslash\varphi$is the set of holon that participate in h, Head $\subseteq$ Subholons is the non-empty set of holon that represent the holon to the environment and are responsible for coordinating the actions inside the holon. C $\subseteq$ Commitments defines the relationship inside the holon and is agreed on by all holons h' $\in$ Subholons at creation of the holon h. Given the holon h =(Head, {h_1, h_2,…, h_n},C) we call h_1, h_2,…, h_n the subholons of h, and h the superholon of h_1, h_2,…, h_n.

2.2 Holon Organization

This holon assumes that the subholons are fully autonomous agents with their predefined architecture and the superholon is just a new conceptual entity whose properties are made up by the properties of the subholons. Fig.1. displays this constellation. In this case no agent has to give up its autonomy, and the superholon is realized exclusively through cooperation among the subholons. The representation of a holon as a set of autonomous agents is in a sense just another way of looking at a traditional multiagent system. This is formally described as holon h = ({A_1,A_2,A_3,A_4}, { A_1,A_2,A_3,A_4},$C_{autonomous}$).

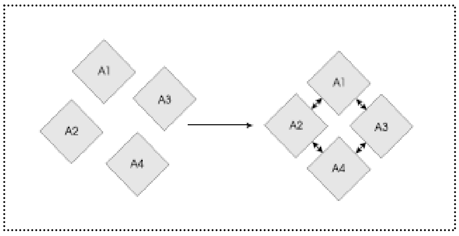


Fig.1. A holon as a set of autonomous agents

The other extreme of the design spectrum terminates the participating sub-agents and creates a new agent as the union of the sub-agents with capabilities that subsume the functionalities of the sub-agents (see Fig. 2.). In this case the merging agents completely give up their autonomy but they may be re-invoked when the superholon is terminated. Naturally, this is a new atomic holon h = ({A}, {A},Cmerge).

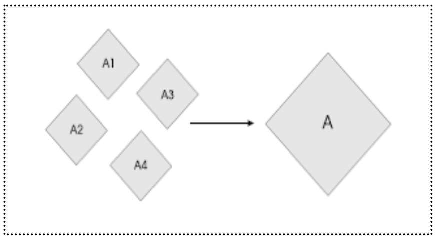


Fig.2. Several agents merge into one

The realization of this approach assumes that the procedure of merging holons leads to the creation of a new agent. For agents of the same kind with an explicit representation of goals and beliefs (e.g., BDI agents) merging can be achieved by creating an agent with the union of the sub-agents' beliefs and goals provided consistency. But for a heterogeneous set of agents this can be intractable and in either case may not be very desirable.

We consider a hybrid way of forming a holon, where agents give up only part of their autonomy to the superholon (see Fig. 3.). From a software engineering point of view it is advisable to allow only for a single head that represents the superholon to the rest of the agent population (to reduce coordination effort). Its competence may range from purely administrative tasks to the authority to give directives to other subholons. Furthermore, the head may have the authority to plan and negotiate for the holon on the basis of its subholons' plans and goals, and even to remove some subholons or to incorporate new subholons. Figure 3 visualizes this approach with an example resulting in a holon h = ({A1}, {A1,A2,A3,A4},Cassociation).

There may be several ways to determine the head. A new agent or one of the members of the holon takes the role of the head and gains the additional functionality. Or, an election procedure is needed to promote one of the agents to leadership. Depending on the application domain, the competence of the representative may vary: the resulting structure can range from a loosely moderated association to a authoritative, hierarchical structure. However, the members of the superholon are always represented as agents, hence we do not lose the capability to solve problems in a distributed fashion.

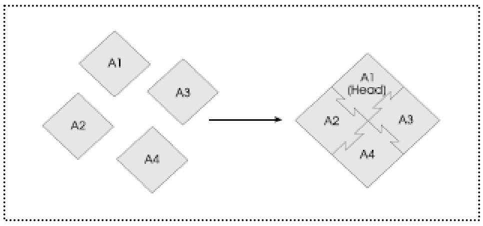


Fig.3. A holon as a moderated association

This approach allows for an explicit modeling of holons, a flexible formation of holonic associations and a scalable degree of autonomy of the participating agents that are subject to negotiation and make up the commitments association of the superholon.

3. SELF-ORGANIZATION EVOLUTION

Evolutionary algorithm is a kind of random searching method of modeling nature evolutionary mechanism. It has many advantages. It is especially suitable for incomplete information problem and complex problem, and successfully used in many domains.

In economics, *Coase* explains the size and forming of organization using "transaction cost" theory [7]. He thinks that the reason of its existent is organization reduces the superfluity cost of attaching to goods and service trade-off. Bases on the theory, Leung proposed a kind of autonomous mechanism for forming suitable scale organization in classifying systems [8]. In this paper, we introduce an organizational evolutionary algorithm (OEA) based on holon organization. The population consists of holon organization, three organization evolution operator (dividing operator, annexing operator, cooperating operator) act on the population for evolving population. In theory, OEA convergences to global optimization, which can be proved like literature [9].

3.1 Organization Evolution Operator

According to the specific application domain, a holonic fathomable utility function F is confirmed. Population members (holon) are accepted by inviting public bidding and the original population is formed by selecting mechanism: $\underline{F} \le F_{holon} \le \overline{F}$, where $\underline{F}$ and $\overline{F}$ are confirmed utility threshold in advance and express lowest and highest utility separately.

3.1.1 Dividing Operator

The condition of dividing organization *org* is that average utility of the organization *org* is smaller than $\underline{F}$ and the scale of the organization *org* is lager.

$$\frac{F_{org}}{|org|} \langle \underline{F} \quad \text{and} \quad U_{(0,1)} \langle \frac{|org|}{N_0} \qquad (1)$$

F_{org} is the utility of the organization *org*, ︱*org*︱expresses the number of the member in the organization *org*, U $_{(0,1)}$ expresses a random number in [0,1], N_0 expresses the scale of population.

If a father generation org_p satisfies（1）, then it is divided into two filial generation organizations org_{c1} and org_{c2}: $\frac{1}{3}\frac{|org_p|}{3} \sim \frac{2}{3}|org_p|$ members form a filial generation organization org_{c1} by selecting from org_p randomly, the others form org_{c2}. Finally, org_p is deleted from the current population; org_{c1} and org_{c2} join next generation population.

In fact, dividing operator limits the size of organization and sends some organizations into next generation population directly, which is propitious to maintain diversity of population and ensure global optimization.

3.1.2 Annexing Operator

We assume that two father generation organizations are org_{p1} $=\{x_1,x_2,\ldots,x_M\}$ and $org_{p2}=\{y_1,y_2,\ldots,y_N\}$。If $F_{org_{p1}} \ge F_{org_{p2}}$, then org_{p1} annexes org_{p2} and a filial generation organization $org_c=\{z_1,z_2,\ldots,z_{M+k}\}$ is produced. $z_i=x_i,(i=1,\ldots,M)$ and

$$z_{M+j}=\begin{cases} y_j & \text{if：} F_{y_j} \ge \dfrac{F_{org_{p2}}}{|y_j|} \\ y_j & \text{otherwise : when } :U(0,1) < Rs \\ \phi & \text{when } :U(0,1) \ge Rs \end{cases}$$

$$j=1,\cdots,k \qquad (2)$$

The *Rs* is a confirmed threshold in advance. If utility of org_{p2}'s member is greater or equals to average utility of org_{p2}, then the member is held; otherwise, it is held or deleted by random strategy. Finally, org_{p1} and org_{p2} are deleted from the current population, org_c joins next generation population.

Annexing operator make the best of high-utility members in organizations to increase the organizational utility. Its effect is equal to local optimization, which is propitious to accelerate the optimal speed. On the other hand, it selects most members in the annexed organization randomly for holding the population diversity.

3.1.3 Cooperating Operator

We assume that two father generation organizations are org_{p1} $=\{x_1,x_2,\ldots,x_M\}$ and $org_{p2}=\{y_1,y_2,\ldots,y_N\}$. Selecting $h_1 \in org_{p1}$ randomly: if $F_{h_1} \ge \dfrac{F_{org_{p1}}}{|org_{p1}|}$, and selecting $h2 \in org_{p2}$ randomly: if $F_{h_2} \ge \dfrac{F_{org_{p2}}}{|org_{p2}|}$, then two filial generation organization org_{c1} and org_{c2} are produced by cooperating.

$$org_{c1}=\begin{cases}\{x_1,\cdots,x_{i-1},h_2,\cdots,x_M\} & \text{if } \exists x_i,\ F_{x_i}<F_{h_2} \\ org_{p_1} & \text{otherwise}\end{cases} \qquad (3)$$

$$org_{c2}=\begin{cases}\{y_1,\cdots,y_{i-1},h_1,\cdots,y_N\} & \text{if } \exists y_i,\ F_{y_i}<F_{h_1} \\ org_{p_2} & \text{otherwise}\end{cases} \qquad (4)$$

Finally, org_{p1} and org_{p2} are deleted from the current population; org_{c1} and org_{c2} join next generation population.

Cooperating operator makes use of high-utility members in organizations to increase utility together by cooperation.

3.2 Organization Evolution Algorithm

The procedure of an organization evolution algorithm can be described as:

Step$_1$：Population's members are accepted by inviting public bidding, and the original population P_0 consists of N_0 members; $t \leftarrow 0$

Step$_2$：Reaching the end condition, then export outcome and close the procedure; otherwise, go to Step$_3$

Step$_3$：For each organization in P_t, if the dividing condition is true, then executes dividing.

Step$_4$：Selecting two father generation organizations org $_{p1}$ and org $_{p2}$ in P_t randomly, then execute annexing or cooperating.

Step$_5$：If the number of P_t's members<N_0 , then accept new members by inviting public bidding until the number=N_0.

Step$_6$：t←t+1, go to Step$_2$.

4. SIMULATING EXPERIMENTS

We use four traditional functions to test OEA performance, and compare the OEA performance with OGA/Q's[8]. F1 is single apex function, F2~F4 are multi-apex functions. In these experiments N_0=168, *RS*=0.78, n=30 in F1~F3, and n=100 in F4. In Table 1, the results are the average values of eighty times experiments.

F1: min $f(x)=\sum_{i=1}^{n-1}\left[100\left(x_{i+1}-x_i^2\right)^2+\left(x_i-1\right)^2\right]$;

$S=[-5,10]^n$; $f_{\min}=0$.

F2: min $f(x)=\sum_{i=1}^{n}(-x_i\sin(\sqrt{|x_i|}))$;

$S=[-500,500]^n$; $f_{\min}=-12569.5$.

F3: min $f(x)=\frac{\pi}{n}\{10\sin^2(\pi y_1)+\sum_{i=1}^{n-1}(y_i-1)^2$

$\times[1+10\sin^2(\pi y_{i+1})]+(y_n-1)^2\}+\sum_{i=1}^{n}u(x_i,10,100,4)$

$$u(x_i,a,k,m)=\begin{cases}k(x_i-a)^m, & x_i>a \\ 0, & -a \le x_i \le a \\ k(-x_i-a)^2, & x_i<-a\end{cases}$$

$y_i=1+\frac{1}{4}(x_i+1)$; $S=[-50,50]^n$; $f_{\min}=0$.

F4:

min $f(x)=-\sum_{i=1}^{n}\sin(x_i)\sin^2(\frac{i\times x_i^2}{\pi})$;

$S=[0,\pi]^n$; $f_{\min}$ less than –99.60.

Table 1. Comparing of OEA performance with OGA/Q's

f	Average value		Standard square error	
	OEA	OGA/Q	OEA	OGA/Q
F1	4.512E-05	6.896E-03	1.836E-04	1.952E-03
F2	-12569.48	-125769.42	8.746E-10	6.468E-04
F3	1.033E-30	6.012E-06	5.898E-30	1.161E-06
F4	-99.5313	-92.86	2.702E-02	2.628E-02

5. CONCLUSIONS

This paper presents self-organization based on holon organization evolution in multi-agent systems, whose advantage is threefold. First, the model preserves compatibility with standard multi-agent systems by addressing every holon as an agent, whether this agent represents a set of agents or not. The complexity of a group of agents is encapsulated into a holon represented by its head, the number of agents involved in the holon becomes irrelevant for other agents communicating with it. The second, holonic multi-agent systems are one way to introduce recursion into the modelling of multi-agent systems, which has been proven to be a powerful mechanism in software design. The third, self-organization not only incarnates inhere characteristics of the nature, being and human society, but also satisfies the need of software design in which the opening and dynamic environment must be faced with.

REFERENCES

[1] G. Weiss. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT Press, 1999.

[2] M. Wooldridge. "An Introduction to Multiagent Systems". *John Whiley & Sons*, 2002.

[3] N.R. Jennings. "Agent-based computing: Promise and perils". *In Proceedings of the 16th International Joint Conference on Arti_cial Intelligence (IJCAI-99)*, pages 1429-1436, 1999.

[4] S. Camazine, J.-L. Deneubourg, Nigel R. F., J. Sneyd, G. Theraulaz, and E. Bonabeau. "Self-Organization in Biological System". *Princeton Studies in Complexity*. Princeton University Press, 2001.

[5] R. G. Smith. "The contract net: A formalism for the control of distributed problem solving". *In Proceedings of the Fifth International Joint Conference on Artificial Intelligence (IJCAI-77)*, page 472, 1977.

[6] K. Fischer, M. Schillo, J. Siekmann. "Holonic Multiagent Systems: A Foundation for the Organisation of Multiagent Systems", *Lecture Notes in Computer Science*, 2003 - Springer

[7] R. H. Coase. *The firm, the Market, and the Law*. Chicago: University of Chicago Press,1988.

[8] Y. W. Leung, Y. Wang. "An orthogonal genetic algorithm with quantization for global numerical optimization". *IEEE Transaction on Evolutionary Computation*, 2001, 5(1):41:53.

[9] Liu Jing, Zhong Wei-Cai, Liu Fang , Jiao Li-Cheng. "An Organizational Evolutionary algorithm for Constrained and Unconstrained Optimization Problems". *Chinese Joural of Computers*, 2004, 27(2):157-167(in Chinese)

[10] V. Lesser, K. Decker, T. Wagner, N. Carver, A. Garvey, B. Horling, D. Neiman, R. Podorozhny, M. NagendraPrasad, A. Raja, R. Vincent, P. Xuan, and X.Q. Zhang. "Evolution of the GPGP/TAEMS Domain-Independent Coordination Framework". *Autonomous Agents and Multi-Agent Systems*, 9(1):87–143, July 2004.

[11] Bryan Horling, RogerMailler, and Victor Lesser. "A Case Study of Organizational Effects in a Distributed Sensor Network". *In Proceedings of the International Conference on Intelligent Agent Technology (IAT 2004)*, Beijing, China, September 2004.

[12] Mark Sims, Claudia Goldman, and Victor Lesser. "Self-organization through bottom-up coalition formation". *In Proceedings of Second International Joint Conference on Autonomous Agents and MultiAgent Systemsr*. In Proc. AAMAS-2003. ACM Press, 2003.

[13] Arnoud Visser, Gregor Pavlin, Sicco P. van Gosliga, Marinus Maris. Self-organization of multi-agent systems. 2005, http://www.science.uva.nl/research/ias/

Jian Gao, born in 1961. Master, associate professor. His current research interests include multi agent system, intelligent optimal.

Distributed Database and Data Mining

Grid-based Approaches for Distributed Data Mining Applications

Lamine M.Aouad, Nhien-An Le-Khac and Tahar M.Kechadi
Univercity College, Dublin-Ireland
{lamine.aouad, an.le-khac, tahar.kechadi}@ucd.ie

ABSTRACT

The data mining field is an important source of large scale applications and datasets which are getting more and more common. In this paper, we present performance evaluations of basic large scale data mining applications on an experimental grid environment. We test the scalability of a new clustering algorithm and a grid-based frequent itemsets generation using a grid workflow. We also compare the performance analysis of the workflow execution to simulated models and give measurements of the overhead related to the workflow engine and the underlying grid environment.

Keywords: Data Mining, Grid Computing, Grid Workflow, Overheads, Grid'5000.

1. INTRODUCTION

The grid has emerged as an important area in distributed and parallel computing and an increasing amount of scientific communities are using grid facilities, which are becoming mature, to share, manage and process large scale datasets and applications. The grid can be seen as a computational and large scale support, and even as a high performance support in some cases. Recently, many high-level grid-based knowledge discovery and data mining systems have been proposed [8], [14], [20], etc. In the grid, data mining applications need efficient and well-adapted approaches able to generate an efficient global knowledge based on distributed local mining and knowledge models.

In this paper we test new grid-based approaches for two important tasks in data mining, namely the clustering and the globally large itemsets generation, on a configurable and controllable grid infrastructures which allow to monitor and reproduce a variety of experimental conditions. The grid-based approaches for data mining applications are motivated by the inherent distributed nature and by the challenge of developing scaling solutions taking into account the constraints related to analyzing massive distributed datasets.

The rest of the paper is organized as follows, the next section briefly surveys some distributed and grid-based efforts in the data mining area. Then, section 3 describe the tested data mining applications. In section 4, we present the grid platform and the experimentation tool. Section 5 shows experimental results and evaluations and highlights directions for future work. Finally, section 6 concludes the paper.

2. RELATED WORK

Many research works addressed this area over the last few years. Most of them were specially dedicated to parallel implementations of widely used algorithms and techniques on clusters or high performance computers using standard message passing interfaces such as MPI or PVM [2][7][9][10][11][12][13][19][23], etc. These research works are related to the k-means clustering algorithm or its variants, some density based clustering algorithms (DBSCAN, BIRCH, etc.), association rules mining, among others. Most of the existing parallel approaches in the literature need either multiple synchronization constraints between processes or a global view of the dataset at some stages, or both.

Furthermore, there are only few works in distributed approaches using different computational sites and domains, with different architectures, dynamic resources, etc. and many of the existing techniques are based on algorithms which were developed for parallel systems. Typically, distributed algorithms act by producing local models followed by the generation of a global model by aggregating the local results. They are based on the global reduction, one or many times, of the so-called sufficient local statistics. Some works are presented in [15][16][17][18][24], mostly related to some clustering algorithms or association rules mining.

Furthermore, some grid-based projects and frameworks already exist or are being proposed in this area; Knowledge Grid [8], Discovery Net [14], ADMIRE [20], etc. These tools aim to offer a high level abstractions and techniques for distributed data mining and knowledge extraction from data repositories or warehouses available on the grid. It uses basic grid mechanisms, mainly provided by existing grid environment, to build their specific knowledge discovery services. Details about these systems is out of the purpose of this paper. Besides, beyond their architecture design, the data analysis and treatment policies, the data integration or placement approaches, the underlying middleware and tools, etc. the grid approach needs efficient and well-adapted algorithms. This is the motivation of this work.

3. DATA MINING APPLICATIONS

A. Distributed clustering

The key idea of this algorithm is to perform local clustering with a relatively high number of clusters, which are referred as subclusters, or an optimal number of clusters found by using an approximation technique, and to merge them at the global level according to an increasing variance criterion which require a very limited communication overhead. In this algorithm, all local clustering processes are independent from each other and the global aggregation can be done independently, from and at any initial local process. The merging process of the local subclusters at the global level exploits locality in the feature space, i.e. the most promising candidates to form a global cluster are subclusters that are the closest in the feature space, including subclusters from the same site.

An important aspect of this algorithm is that the merging is logical, i.e each local process can generate correspondences (or labeling) between local subclusters, without necessarily reconstructing the overall clustering output, i.e. without communicating subclusters. That is because the only bookkeeping needed from the other sites are centers, sizes and variances. The aggregation is then defined as a labeling process between local subclusters in each participating site. This can be followed by a broadcast of the results if needed. A perturbation process is activated if the merging action is no longer applied. A number of candidates (user defined parameter) are collected for each global cluster from its border. Then, the process moves these candidates by trying the closest ones and with respect to the gain in the variance criterion when moving them from the neighboring global clusters. Details about the algorithm can be found in [6].

B. Distributed Frequent Itemsets Mining

Discovering frequent patterns is another crucial task in data mining. It is a core step of association rules discovering. Many existing algorithms, both sequential and parallel, are related to the Apriori algorithm [3]. This algorithm exploits the observation that all subsets of frequent itemsets are frequent themselves. We propose here grid-based implementations based on this algorithm. Our approach is different from the existing distributed approaches since the grid-based implementations imply some constraints related to the underlying middleware and tools, and the communication and synchronization overheads, which are sometimes excessive and consequently highly suitable to avoid. The local generation of the candidate sets is in the Apriori manner, only the requested globally large k-itemsets are computed, and usually only the maximum required. This greatly reduce the communication costs and avoid multiple synchronization phases since only local pruning is preferred in intermediate steps. Details about this algorithm can be found in [5]

4. EXPERIMENTATION TOOLS

A. The Grid’5000 Platform

Grid’5000 [4] [21] is a large scale environment for grid research. It aims to provide a reconfiguration and monitoring instrument to investigate grid issues under real and controllable conditions. Grid’5000 is not a production grid, it is an instrument that can be configured to work as a real testbed at a wide area scale. Grid’5000 is composed of nine geographically distributed clusters. A set of control and monitoring tools allow users to make reservation, configure or reconfigure a specific owner environment (with a given stack of software and/or operating systems, etc.), make deployment, and run experiment and measurements.

B. Workflow Management in the Grid

Several significant research works have been conducted in recent years to automate the workflow activities using advanced workflow management tools in the grid. In the following, we briefly present the Condor system which includes the DAGMan workflow manager.

1) *The Condor system:* is a batch system providing a job management mechanism, resources monitoring and management, some scheduling functionalities and priority schemes, and grid-based and grid-enabled functionalities [22]. The Condor system provide a ClassAds mechanism for matching resource requests and offers, a checkpointing and migration mechanisms, and a job management capabilities to across the grid.
2) *The DAGMan meta-scheduler:* is a directed acyclic graph representation manager which allows to express dependencies between Condor jobs. It allows user to list the jobs to be done with constraints on the order through several description files for the DAG and the jobs within the task graph. It also provides fault-tolerant capabilities allowing to resume a workflow where it is left off. However, the scripting language required by DAGMan is quite heavy since every job in the dag has to have its own condor submit description file.

5. EXPERIMENTAL EVALUATIONS AND MODELS

A. Computational Resources

The computational resources are composed of 470 nodes distributed over five sites of the Grid’5000 platform. Local clusters are interconnected via Gigabit Ethernet, and the different locations by RENATER through VLANs using MPLS at level 2. Average bandwidths and latencies measured (using the NetPerf tool [1]) between sites range between 12.71 Mb/s and 106.63 Mb/s for bandwidth, and between 5 and 28 *ms* for latency. In local sites the average bandwidth is about 941 Mb/s and 0.07 *ms* for latency. The average times observed during the tests are used to compute the estimated computing times in the modelization section below. These measurements are reported in [5].

B. Tests

The tests are separated in two parts: experiments and modelization. The reason is that important overheads related to the workflow engine was noticed in the first set of tests. This will be highlighted in the rest of the paper by giving an estimation of these overheads. The modelization section is then intended to give boundary results and evaluation of the proposed algorithms instead of connecting them to the grid tools performance. We will also briefly discuss this issue.

For both applications, synthetic datasets are generated. In the clustering task, the data is a set of random Gaussian distributions. For the frequent itemsets mining, synthetic transactions from different sizes were generated.

For frequent itemsets generation, the global datasets are composed of 4×10^6 transactions distributed over 200 processes (with about 20000 transactions each). Two versions of the presented algorithm were implemented; the first version computes only the maximum required frequent itemset (a user defined parameter), and the second computes all frequent itemsets until the required size. However, this latter version uses interesting relationships between locally large and globally large itemsets to generate smaller sets of candidates at each communication step. Both versions can be used to reach maximal frequent itemset, i.e. until there is no more candidates. In this test, the globally large 4-itemsets were generated. The computing time for the proposed version was 521 minutes whereas it was 687 minutes for the second version. This corresponds to a better performance by up to 25% of the computing time. In fact, we noticed that, in addition of the multiple communication and synchronization steps in the second version, the remote support computation is quite computationally expensive, representing on average up to 13% of its whole computing time presented before. Since this process is repeated $k - 1$ times to reach the size k, this will decrease the performance comparing to the proposed version for large sizes. Actually, even with a small support for a large number of transactions, which means that local processes for the first step in the first version can be very computationally expensive, the gain is important for the proposed approach.We expect that the performance gain will be more important in grid platforms with less communication performance.

For the clustering task, the global dataset is composed of 5 × 10^7 samples distributed over 200 processes. The initial local clustering uses k-means with 20 subclusters in each process. For the merging step, the constraint parameter is set up as twice the highest individual subclusters variance. This step includes subclusters perturbation, once the merging is no longer applied, but no additional communications are required for the computation of the new statistics at this level. The average execution time for the global clustering is 1050 seconds, including the initial communications for the input data files and all the job preparation and scheduling steps. The actual computing time for the clustering (the maximum computed locally) and the merging processes represents approximately about 2% of the whole workflow execution time. We will discuss this further in the following paragraphs.

Then, considering the results previously shown, we estimate the overheads. For the frequent itemsets generation, the model presented here consider the execution times of the pieces of software needed in each process, namely different versions of Apriori generation depending on the size of the itemset, supports computation and different versions of globally large itemsets generation depending on whether a local itemset of candidates is generated or not, and the estimation of the communication times using measurements for bandwidths and latencies obtained by the NetPerf network performance benchmark. For the version that generates only the maximum requested frequent itemsets, if the application is intended to be executed in p processes, the three stages corresponding to the Apriori generation, supports computation for remote processes and the local globally large itemsets generation, are then parallel activities. The overall time execution is the sum of the maximum of the execution times at each stage. In the version generating all the frequent itemsets until a requested size or a maximal one, if the application is intended to be executed in p processes and k is the intended size, then there is 2k −1 stages of parallel activities starting by locally large 2−itemsets.

The estimation of the communication times is based on the size of the real remote supports requests files sent in the experiment tests described in the previous paragraphs. The overall estimation time of both versions is respectively 424 and 518 minutes for k = 4, with same sizes and local support as in the experimentation tests since the performance of the Apriori generation, remote supports computation and all intermediate processes can vary greatly depending on these parameters. The proposed version gives better estimated time by 18.2% compared to the version computing all globally large itemsets during the computing process. Actually, as quoted before this can be generated at the end of the overall process in the proposed version. On the other hand, this corresponds to an overhead between 19% to 22% respectively compared to the execution on the grid.

In the case of distributed clustering, the estimated execution time is the sum of the maximum of the execution times of local clustering processes, and the merging process. The estimation of this time is about 19 seconds on average. On the other hand, according to bandwidth measurements, the worst communication overhead case gives an estimation of 0.52 second. The sum represents less than two percent of the whole execution time presented in the previous paragraphs. Also, the communication overhead for the aggregation step of our algorithm is very small compared to the computation times since the only statistics about local clustering transmitted to the aggregation process are centers, sizes and variances. The communication cost of these statistics is insignificant. Thus, 98% of the whole computing time, on average, is considered as overheads from different levels for this grid workflow implementation.

C. Discussion

Our implemented algorithms, within a workflow environment for the grid, considerably reduce data communication and task synchronization which can be the most critical in term of execution efficiency since we consider these properties as the most suitable for grid-based applications. Still, some constraints on the experimental grid middleware make realistic expectations difficult to achieve and lead to relatively poor performance and scalability even with fast and independent activities. Job preparation and submission latencies are of prime importance in this case and are the first sources of efficiency loss.

Nevertheless, the comparison between different versions, especially for frequent item sets generation and the simulated results show the effectiveness of the presented algorithms, and at least give good targets for future evaluations. Other versions of both algorithms are currently under evaluation considering a hierarchy of processes according to the size of the local datasets and their locations which can be more suitable for the grid. As for execution overheads, we notice the important gap between the execution times and the bounds given by models which can be seen as ideal execution times. The severity of the clustering case may come from the fact that the local clustering is not that computationally expensive, and thus the parallel execution section containing these tasks cannot improve the overall execution time. The frequent itemsets generation gives less important overheads (up to 22%). This shows a great correlation between the size of the jobs in the workflow and its performance independently from the size of any parallel loops.

It is hard to estimate an ideal performance model for workflow applications on the grid since a hierarchy of overheads is involved and traditional parallel metrics are no longer applied. While it is mostly a complex scheduling issues for complex workflow applications, it seems to be more related to the grid tool structure and implementation in our case since the jobs were equivalent, almost in parallel loops which not exceeds the number of processors available, with a very limited communications and synchronizations, and done under deployed environments, i.e. controlled and stable conditions. The aim here is not to explain the nature and reasons of such performance looses but to accentuate this open issue that still receives small consideration from the community through the implementation of real-world grid intended applications.

6. CONCLUSIONS

In this paper we presented grid-based implementations of basic data mining applications. We evoked the need of well adapted distributed and grid-based algorithmic approaches. We proposed lightweight distributed algorithms, for the clustering problem and for frequent itemsets generation which are two fundamental applications in data mining. Both of the proposed algorithms have very limited communication overheads. The models and the comparison between different versions of the proposed algorithms show the effectiveness of the proposed approaches and attest that grid implementations have an essential need to exchange a few data and avoids synchronization, even with more computationally expensive tasks. However, realistic performance hopes are difficult to achieve since the gap between the simulated models and experiments reach 98% of the obtained execution time for the clustering case. This gap is less severe with more computationally expensive jobs for frequent itemsets generation, which seems to be more suitable for workflow programming since the preparation overheads are overlapped. Finally, more analysis of these overheads is planed to give more details about the efforts that should both middleware developers and algorithms designers focus on.

ACKNOWLEDGMENTS

Experiments presented in this paper were carried out using the Grid'5000 experimental testbed, an initiative from the French Ministry of Research through the ACI GRID incentive action, INRIA, CNRS and RENATER and other contributing partners, which the authors would like to thank.

REFERENCES

[1]. Netperf: "A Network Performance Benchmark," Technical report, *Information Networks Division*. Hewlett-Packard Labs.

[2]. R. Agrawal and J. C. Shafer. "Parallel mining of association rules," *IEEE Trans. On Knowledge And Data Engineering*, 8:962–969, 1996.

[3]. Rakesh Agrawal and Ramakrishnan Srikant. "Fast algorithms for mining association rules," In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 1994.

[4]. F. Cappello & all. Grid'5000: "A Large Scale, Reconfigurable, Controlable and Monitorable Grid Platform," In *6th IEEE/ACM Int. Workshop on Grid Computing, Grid'2005*, 2005.

[5]. L. M. Aouad, N-A. Le-Khac, and T. M. Kechadi. "Distributed data mining on grid platforms," Technical report, University College Dublin, 2007.

[6]. L. M. Aouad, N-A. Le-Khac, and T. M. Kechadi. "Variance-based clustering technique for distributed data mining applications," Technical report, University College Dublin, 2007.

[7]. M. Z. Ashrafi, D. Taniar, and K. A. Smith. Odam: "An optimized distributed association rule mining algorithm," *IEEE Distributed Systems Online*, 5(3), 2004.

[8]. M. Cannataro, A. Congiusta, A. Pugliese, D. Talia, and P. Trunfio. "Distributed Data Mining on Grids: Services, Tools, and Applications," *IEEE Transaction on System, Man, and Cybernetics*, 34(6), Dec 2004.

[9]. D. W. Cheung, J. Han, V. T. Ng, A. W. Fu, and Y. Fu. "A fast distributed algorithm for mining association rules," In *PDIS: Int. Conf. on Parallel and Distributed Information Systems*, 1996.

[10]. I. S. Dhillon and D. Modha. "A Data-Clustering Algorithm on Distributed Memory Multiprocessors," In *Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD*, 1999.

[11]. M. Ester, H.-P Kriegel, J. Sander, and X. Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In *2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1996.

[12]. A. Garg, A. Mangla, V. Bhatnagar, and N. Gupta. PBIRCH: "A Scalable Parallel Clustering algorithm for Incremental Data," In *10th International Database Engineering and Applications Symposium, IDEAS'06*, 2006.

[13]. H. Geng, X. Deng, and H. Ali. "A New Clustering Algorithm Using Message Passing and its Applications in Analyzing Microarray Data," In *ICMLA '05: Proc. of the 4th Int. Conf. on Machine Learning and Applications (ICMLA'05)*, 2005.

[14]. V. M. Ghanem, Y. M. Kohler, A. J. Sayed, and P. Wendel. Discovery Net: "Towards a Grid of Knowledge Discovery," In *Eight Int. Conf. on Knowledge Discovery and Data Mining*, 2002.

[15]. E. Januzaj, H-P. Kriegel, and M. Pfeifle. "Towards Effective and Efficient Distributed Clustering," In *Int. Workshop on Clustering Large Data Sets, 3rd Int. Conf. on Data Mining, ICDM*, 2003.

[16]. E. Januzaj, H-P. Kriegel, and M. Pfeifle. DBDC: "Density-Based Distributed Clustering," In *9th Int. Conf. on Extending Database Technology, EDBT*, 2004.

[17]. E. Januzaj, H-P. Kriegel, and M. Pfeifle. "Scalable Density-Based Distributed Clustering," In *8th European Conference on Principles and Practice Discovery in Databases PKDD*, 2004.

[18]. R. Jin, A. Goswani, and G. Agrawal. "Fast and Exact Out-of-Core and Distributed K-Means Clustering," *Knowledge and Information Systems*, 10, July 2006.

[19]. M. N. Joshi. "Parallel K-Means Algorithm on Distributed Memory Multiprocessors," Technical report, University of Minnesota, 2003.

[20]. N-A. Le-Khac, M. T. Kechadi, and J. Carthy. "Admire framework: Distributed data mining on data grid platforms," In *first Int. Conf. on Software and Data Technologies, ICSOFT*, 2006.

[21]. B. Quetier and F. Cappello. "A survey of Grid research tools: simulators, emulators and real life platforms," In *IMACS 2005, 17th World Congress on Scientific Computation, Applied Mathematics and Simulation*, 2005.

[22]. D. Thain, T. Tannenbaum, and M. Livny. "Distributed Computing in Practice: The Condor Experience," *Concurrency and Computation: Practice and Experience*, 2004.

[23]. X. Xu, J. Jager, and H.-P. Kriegel. "A Fast Parallel Clustering Algorithm for Large Spatial Databases," *Journal of Data Mining and Knowledge Discovery*, 3, 1999.

[24]. B. Zhang and G. Forman. "Distributed Data Clustering Can be Efficientand Exact," Technical report, Hewlett-Packard Labs, 2000.

A New Data Mining Approach Based on Rough Set

Shangping Dai, Tian He, Xiangming Xie
Computer science of Central China Normal University, Wuhan 430079 P.R.China
Email: hhiouhh@yahoo.com.cn

ABSTRACT

In this work, a new and an efficient RS is used to perform the task of predicting breast cancer recurrence. The rough sets based approach was applied to obtain risk factors for the breast cancer recurrence. The set of selected attributes, which ensured high quality of the classification, was obtained. This so-called "recurrence" is the return of cancer after initial treatment. It causes considerable distress for breast cancer survivors. To design interventions to breast cancer recurrence, we must know what survivors are potential to recurrence, what factors that influence the likelihood that breast cancer may return. Various published algorithms have been applied to breast cancer datasets, but rough set is a fairly new intelligent technique that applies to predict breast cancer recurrence. We analyze Ljubljana Breast Cancer Dataset, firstly, obtain lower and upper approximations and calculate the accuracy and quality of the classification. The high values of the quality of classification and accuracy prove that the attributes selected can well approximate the classification. Rough sets approach for generation of all reducts that contains minimal number of attributes and rules are introduced. Finally, rules based-on rough set are generated and the proposed rules are promising for improving prediction accuracy.

Keywords: Rough Set, Lower and Upper Approximation, Attribute Reduction, Core, Breast Cancer Recurrence, Decision Rules

1. INTRODUCTION

and in the choice of treatment. Data mining usually mea Breast cancer is very common worldwide, with 800,000 new cases diagnosed each year [1], and it is a multi-step disease, a malignant tumor that develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Although there are recognized factors that increase the risk of breast cancer, its causes are still unknown and thus there is no way of preventing it [2]. Consequently there has been a considerable amount of research focused on breast cancer, and death rates have fallen over the last 10 years. Successful application of data mining to cancer patient data can result in new knowledge which can assist in cancer diagnosis ns the methodologies and tools for the efficient new knowledge discovery from databases. It is also a form of knowledge discovery essential for solving problems in a specific domain. For instance, the data mining approaches are applied in the filed of medical care recently (Abbass, 2002; Chou, Lee, & Shao, 2004; Ohmann, Moustakis, Yang, & Lang, 1996; Pendharkar, Rodger, Yaverbaum, Herman, & Benner, 1999).

Initial treatment (surgery to remove the tumor and any lymph nodes to which the cancer may have spread) is usually complemented by chemotherapy, radiation therapy or hormonal therapy to reduce the risk of cancer recurring in the future [8]. The health care team will make every effort to remove all traces of a breast tumor during surgery. Many patients may never have a recurrence, but breast cancer may still recur in some patients. A recurrence can happen months or years after the original diagnosis and treatment. For example, even though a breast cancer tumor may appear small and localized, it may be aggressive and may have spread beyond the breast; this spread cannot always be detected by current methods [3]. This aggressiveness, as well as other factors, can lead to breast cancer recurrence. What's the worst is that a diagnosis of recurrent cancer is more devastating or psychologically difficult for a woman than her initial breast cancer diagnosis. So physician must analyze data relates to the recurrence of breast cancer among patients according to medical factors. Generally, the purpose of all the related researches is identically to predict a cancer recurrence. And at the same time, for the needs of improving the prediction accuracy in breast cancer recurrence, more and more researchers have tried to apply artificial intelligence related approaches for breast cancer prediction. The task here was to ascertain whether individuals suffered a recurrence of breast cancer, based on nine medical variables.

Recently, various published algorithms have been applied to build a computer-aided analysis system in medical field [4, 5]. The most commonly used algorithms are neural networks, Bayesian classifier [12], genetic algorithms, decision trees, and fuzzy theory [6]. Rough set theory [7] is a fairly new intelligent technique that has been applied to the medical domain and is used for the discovery of data dependencies, evaluates the importance of attributes, discovers the patterns of data, reduces all redundant objects and attributes, and seeks the minimum subset of attributes. Moreover, it is being used for the extraction of rules from databases. One advantage of the rough set is the creation of readable if-then rules. Such rules have a potential to reveal new patterns in the data material; furthermore, it also collectively functions as a classifier for unseen data sets. Unlike other soft computing methods, rough set analysis requires no external parameters and uses only the information presented in the given data. It is shown to be an interesting and powerful theory, and it has been used previously in attribute selection, rule induction, classification, multi-agent systems, medical diagnosis and other application domains[9~11].

This paper applies a new and efficient rough set-based methodology to analyze the Ljubljana Breast Cancer Dataset (Zwitter and Soklic, 1988), a set of 286 instances of real patient data with a binary outcome (Recurrence/ No Recurrence) and 9 possible predictive attributes. The objective of using rough set theory is to identify the set of most relevant attributes leading to the recurrence, and to generate decision rules based on these attributes so that preventive actions can be taken.

The remainder of this paper is organized as follows. Section2 describes the methodology of Rough Set Theory. Section3 describes date and uses ROSETTA (Predki & Wilk, 1999) software to analyze the data and discuss the experimental results. Scetion4, the conclusion of the paper is summarized and the directions for future research are described.

2. PRELIMINARY

Rough sets theory was first introduced by Pawlak in the 1980's [7] Reduct and core are the two important concepts in rough sets theory. Based on Pawlak's book [7], we explain the basic concepts in rough sets theory in the following.

Definition1. Information table :an information table is defined as S = (U,Q,V), where U is the universe consisting of a finite set of objects, Q is a finite set of the attributes and V is a set of values$=\bigcup_{q\in Q} V_q$ where Vq is a value of the attribute q.

Definition2. Lower and upper approximation: For some objects in an information table cannot be exactly distinguished given the set of attributes, they could be roughly (approximately) distinguished. This idea gives rise to the development that defines a set by a pair of sets, i.e., lower and upper approximations. Let $A \subseteq Q$ and $X \subseteq U$, the A-lower approximation, denoted by $\underline{AX}$, of the set X and the A-upper approximation, denoted by $\overline{AX}$, of the set X are defined as follows:

$$\underline{AX} = \{x \in U : [x]_A \subseteq X\}$$

$$\overline{AX} = \{x \in U : [x]_A \cap X \neq \varnothing\}$$

These definitions state that objects $x \in \underline{AX}$ belongs certainly to X, while objects $x \in \overline{AX}$ could belong to X.

Definition3. The accuracy and quality of approximation of classification:
Given the approximation of a classification, the accuracy of approximation of classification Y by A (or accuracy of classification in short) is defined as follows:

$$\alpha A(Y) = \frac{\sum_{i=1}^{n} |\underline{AY_i}|}{\sum_{i=1}^{n} |\overline{AY_i}|}$$

Where |X| denotes the cardinality of a set X. the quality of approximation of classification Y by A (or quality of classification in short) is defined as follows:

$$\rho_A(Y) = \frac{\sum_{i=1}^{n} |\underline{AY_i}|}{|U|}$$

Definition 4 Reduct: given a classification task mapping a set of variables C to a set of labeling D, a reduct is defined as any $R \subseteq C$ such that $\gamma(C, D) = \gamma(R, D)$. The set of all reducts of A is denoted Red (A). Attribute reduction involves removing attributes that have no significance to the classification at hand. It is obvious that a dataset may have more than one attribute reduct set.

Definition 5 Core: an attribute Cj∈C is a core attribute in C with respect to D is the intersection of all reducts. It is defined as follows:

$$Core(C) = \bigcap_{Ri \in Red(B)} Red_i, \quad \text{i=1, 2……}$$

Definition 6 Decision rules: Given an information system, RST can generate decision rules for objects of known classes, or predict classes to which new objects belong. Assume that Q = C ∪ D and $C \cap D = \varnothing$ where C is the set of condition attributes, and D represents the set of decision attributes. Let the d-elementary sets in S be denoted by Yj (j = 1… n) and called decision classes. The syntax of a decision rule can be expressed as follows:

If (conjunction of conditions) then (disjunction of decisions).

3. EXPERIMENTS AND RESULTS

3.1 Problem Description

For our application we chose the problem of predicting recurrence in breast cancer, a well-characterized problem in breast cancer prognosis. We used the Ljubljana Breast Cancer Dataset (Zwitter and Soklic, 1988), a set of 286 instances of real patient data with a binary outcome (Recurrence/ No Recurrence) and 9 possible predictive attributes. This dataset has been used in the past for several machines learning projects (Michalski et al., 1986; Clark and Niblett, 1987).These attributes and their values are given in Table 1

Table 1. Attributes and values of breast cancer

No	attributes	value
1	Class	no-recurrence-events, recurrence-events
2	Age	10-19,20-29,30-39,40-49,50-59,60-69, 70-79,80-89,90-99
3	Menopause	it40, ge40, premeno
4	Tumor Size	0-4,5-9,10-14,15-19,20-24,25-29,30-34, 35-39,40-44,45-49,50-54,55-59
5	deg-malig	1,2,3
6	Inv nodes	0-2,3-5,6-8,9-11,12-14,15-17,18-20, 21-23,24-26,27-29,30-32,33-35,36-39
7	Node Caps	yes, no
8	Breast Quadrant	left-up, left-low, right-up, right-low, central
9	Breast	left, right
10	Irradiat	yes, no

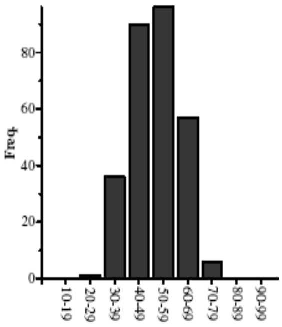
For brevity, we omit the explanation of the attribute and value and describe it only necessary later.

Since all records can be partitioned into two classes, and nearly70.3% of records（that is 201）are no-recurrence-events, 29.7% of records (that is 85) are recurrence. Recurrence is the return of cancer after initial treatment and it is devastating and due to death, so it is necessary to predict the recurrence and take some useful steps to prevent it as soon as possible.

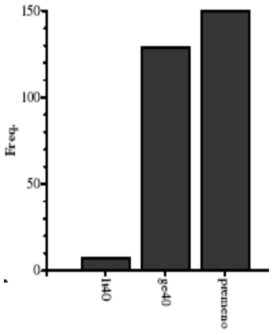
Fig.1. Lists the patient population characteristic of each attribute, from which we can know distributing of population on every values.

3.2 Compute Lower and Upper Approximations

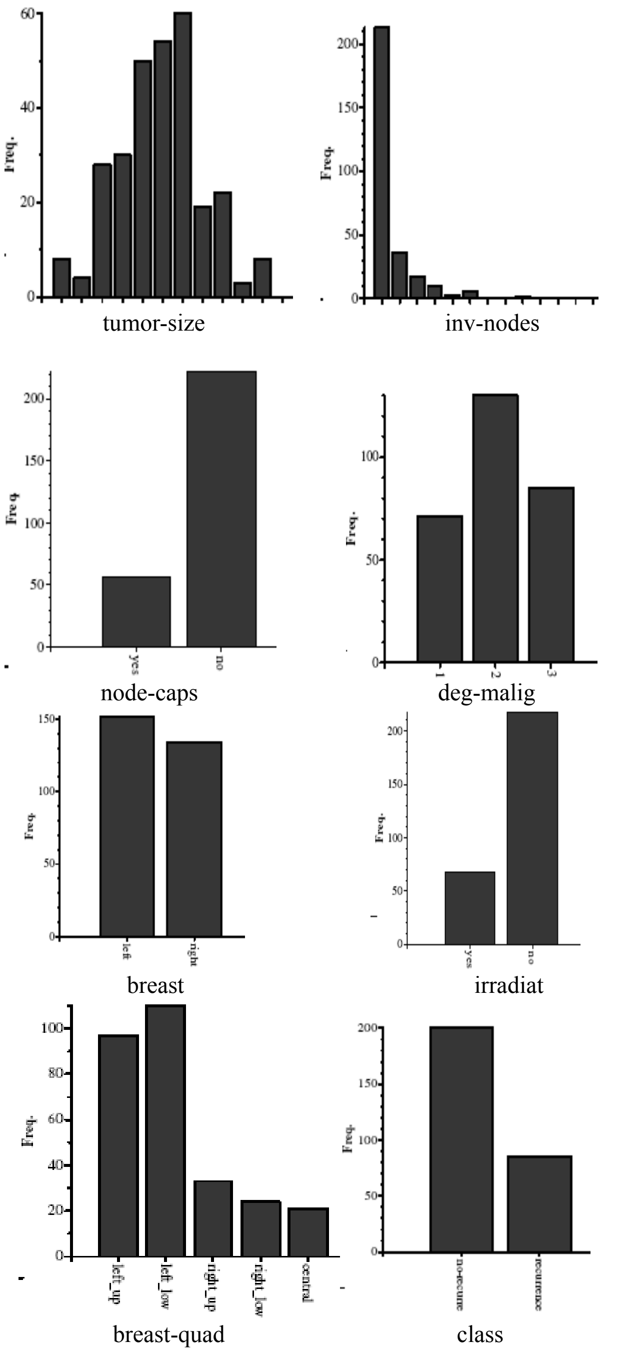
We use ROSETTA (Predki & Wilk, 1999) software to analyze the data in section3, it will be direct to calculate the lower and upper approximations of a certain set, if equivalences classes of condition and decision attributes are given.



age



menopause



tumor-size inv-nodes

node-caps deg-malig

breast irradiat

breast-quad class

Fig.1. The patient's population distribution of each attribute of each value

Table 2. Give the number of records of lower and upper approximation

	No. of records	No. of lower approx	No. of upper approx	Class Accuracy
No-recurrence	201	198	205	0.9659
Recurrence	85	81	89	0.9101

On the basis of table2, we compute the accuracy of the classification (198+81)/ (205+89) =0.9490 and the quality of the classification (198+81)/ (201+85) =0.9755.In general, high values of the quality of classification and accuracy mean that the attributes selected can well approximate the classification. Low values suggest that the set of attributes may be inadequately chosen.

3.3 Reduct And Selection Of Attributes

As described earlier, RST is well suited to identify the most significant attributes by computing reducts and cores. The computation of the reducts and the core of the condition attributes from a decision table is a way of selecting relevant features. Some condition values may be unnecessary in a decision rule produced directly from the database. Such values can then be eliminated to create a more comprehensible minimal rule preserving essential information. A decision table may have more than one reduct. Anyone of them can be used to replace the original table. Finding all the reducts from a decision table is NP-Hard [19]. Fortunately, in many real applications, it is usually not necessary to find all of them, one is sufficient enough.

In traditional rough set models, a prevalent method to get the core is to construct a decision matrix first, then search all the entries in the decision matrix to find all those entries with only one attribute. If they entry in the decision matrix contains only one attribute, that attribute is a core attribute. This method is not efficient and it is not realistic to contract a decision matrix for large database. So, we adopt a new function to compute the core attribute based on the cardinality computation. The reduct algorithm based on the computation of core function is given as follows:

```
Algorithm: Reduct (minimal number of attributes)
Input: a decision table.
Output: Red: a set of minimum attribute subsets.
(1) Compute the core attribute of table DT (i.e. Core)
(1.1) Set Core= { }
(1.2) For each attribute Ci∈C do
{if Card(C,Ci+D)≠Card(C-Ci) then Core=Core∪C}
(2) Red→{ } Do
(3) DT→Red
(4) Loop x∈(C-Red)
(5) if  γR∪{x}(D)> γDT(D)
(6) DT=R∪{x}
(7) R→DT
(8) Until (γR(D)= γC(D)
(9) Return
```

Following the above steps, we can obtain attributes reduction and core:

$$Core(C) = \bigcap_{R_i \in Red(B)} Red_i, \quad i=1, 2\ldots\ldots$$

Intersecting all reducts, we derive core as {tumor-size, inv-nodes, deg-malig,} The quality of classification using only core attributes is 0.9321 which is close to 0.9755that uses all attributes. In contrast to the original 9 condition attributes, the CORE contains only 3 attributes while still achieving good quality of classification.

3.4 Compute Lower And Upper Approximations

Rules represent dependencies in the dataset, and represent extracted knowledge, which can be used when classifying new objects not in the original information system. When the reducts were found, the job of creating definite rules for the value of the decision attribute of the information system was practically done. To transform a reduct into a rule, one only has to bind the condition attribute values of the object class from which the reduct originated to the corresponding attributes of the reduct. Then, to complete the rule, a decision part comprising the resulting part of the rule is added. This is done in the same way as for the condition attributes.

To classify objects, which has never been seen before, rules generated from a training set will be used. These rules represent the actual classifier. This classifier is used to predict to which classes new objects are attached.

Table 3. Tumor-size and recurrence

Tumor-size	*≤29*	*>29*
P(Recurrence)	0.24	0.36
No. of records	160	112

x has Tumour ⩽29 mm⇒x will have Non-Recurrence
x has Tumour >29 mm⇒x will have Recurrence

Table 4. Number of inv-nodes and recurrence

No. of inv-nodes	*≤2*	*>2*
P(Recurrence)	0.20	0.55
No. of records	205	67

x has Number of Involved Nodes ⩽ 2 ⇒ x will have Non-Recurrence.
x has Number of Involved Nodes>2⇒x will have Recurrence.

Table 5. Histological Grade and Recurrence

Hist.grade	*≤2*	*3*
P(recurrence)	0.19	0.53
No. of records	193	79

x has Histological Grade⩽2⇒x will have Non-Recurrence
x has Histological Grade >2⇒x will have Recurrence

To induce a set of decision rules, we use LEM2 (Grzymala-Busse, 1992) algorithm that generates the minimum set of rules, i.e., the set does not contain any redundant rules. In order to discover more important rules in each class, we focus on the rule whose strength is great.

Table 6. Decision rules

	Decision rules
1	if breast = left∧inv-nodes > 2.5 then recurrence else no-recurrence
2	if deg-malig > 2.5∧inv-nodes > 2.5 then class recurrence else no-recurrence

3	if deg-malig > 2.5∧inv-nodes > 2.5 then class recurrence
	if deg-malig> 2.5∧inv-nodes < 2.5∧(tumor-size = 30-34∨ tumor-size= 25-29∨tumor-size= 50-54) then class recurrence
	else no-recurrence

In this experiment data set is incomplete data, we assumed that all missing attribute values were lost, we compute the error rate is that 23.02 %(66err).

It will not benefit Decision Maker, if too many decision rules are produced. We chose the rules whose strength is great and take actions soon.

The exploration above shows that there are three risk factors: tumor-size, inv-nodes, deg-malig, women who have lymph node involvement are more likely to have a recurrence. In general, the larger the tumor, the greater the chance of recurrence, so does deg-malig, the deeper the degree, the greater the chance.

There are other methods for attribute (feature) extraction, such as the FUSINTER technique and VPRS, which may enhance the testability of Rough Set Theory (Beynon & Peel, 2001). The question about analysis of spatial data in spatial data mining and an optimally discrete attribute may a topic for future in depth research.

4. CONCLUSIONS

This paper provides an alternative approach by using rough set-based mining approach to discover the useful decision rules automatically from the breast cancer database. From the results of the above analysis, the following conclusions can be drawn :(1) How big the tumor-size is initially affects how likely it is that the cancer will return, with larger tumor more likely to recur. (2) What degree of malignancy-Grade 3 tumors predominately consists of cells that are highly abnormal, such abnormalities influences the chances of a recurrence. (3) How many inv-nodes –the axillary lymph nodes act as a primary site of drainage for the breast, they are a common site of early metastasis. If the cancer has spread to the lymph nodes – the collection of immune cells and supportive tissue located throughout the body that act as filters and rid the body of unwanted substances – then there is a higher chance of breast cancer recurrence.

By using the proposed rough set-based approach, we can eliminate the redundant attributes that have no significance to the classification at hand, and keep the ability of classification, accordingly, the quantity of calculate decrease. The rules obtained by proposed method have slightly higher accuracy, also, the rules extracted by the proposed approach are much simpler and easier to be understood. It indicates that the proposed method is a suitable approach for generate decision rules and thus it provides effective decision supports for solving the breast cancer recurrence prediction problem.

REFERENCES

[1] D.M. Parkin, P. Pisani, J. Ferlay, *Global cancer statistics*, Ca. Cancer. J. Clin. 49 (1999),pp33–64.

[2] http://www.patienthealthinternational.com/features/14566.aspx

[3] http://www.y-me.org/information/concerned_about_breast_cancer/recurrence.php

[4] Aboul Ella Hassanien,"Rough set approach for attribute reduction and rule generation: a case of patients with suspected breast cancer,"in *Journal of the American Society for Information Science and Technology (JASIST)* 55(11) (2004),pp954–962.

[5] K. Nico, T. Martin, H. Jan, V.E. Leon, Digit*al Mammography:Computational Imaging and Vision*, Kluwer Academic Publication,Hingham, MA, 1998.

[6] Aboul Ella Hasanien, Jafar M. Ali, N. Hajime,"Detection of speculated masses in mammograms based on fuzzy image processing,"in *Seventh International Conference on Artificial Intelligence and Soft Computing, ICAISC2004*, Zakopane, Poland, June 7–11, Lecture Notes in Artificial Intelligence, vol. 3070, Springer series, 2004, pp. 1002–1007.

[7] Z. Pawlak,"Rough Sets,"in *International Journal of Computer and Information Science 11*(1982),pp341–356.

[8] Ta-Cheng ChenTung-Chou Hsu, "A GAs based approach for mining breast cancer pattern," in *Expert Systems with Applications 3*0 (2006),pp674–681

[9] Jhieh-Yu Shyng, Fang-Kuo Wang,"Rough Set Theory in analyzing the attributes of combination values for the insurance market,"in *Expert Systems with Applications* 32

(2007) 56–64

[10] Hsu-Hao Yang, Tzu-Chiang Liu,"Applying rough sets to prevent customer complaints for IC packaging foundry,"in *Expert Systems with Applications* 32 (2007) 151–156

[11] AboulElla Hassanien,"Fuzzy rough sets hybrid scheme for breast cancer detection,"in *Image and Vision Computing* 25 (2007),pp172–183

[12] MATT WILLIAMS1 and JON WILLIAMSON2, "Combining Argumentation and Bayesian Nets for Breast Cancer Prognosis,"in *Journal of Logic, Language and Information* (2006) 15: 155–178C.W. Churchman, The Design of Inquiring Systems, New York: Basic Books Inc. Pub.,1971.

Tian He was born in1983, a postgraduate student in Computer science of Central China Normal University, research interests mainly including artificial intelligence, data mining and rough set, concept lattices.

Hydrological Data Mining Research Based On Time Sequence

Feng Xu, Zhijian Wang
College of Computer & Information Engineering,
Hohai University, Nanjing 210098
E-mail: njxufeng@163.com

ABSTRACT

A time sequence basic arithmetical groupware flexible integrating technology project based on J2EE platform, which is aiming at water conservancy area, and the rationality and feasibility of the project are also researched carefully. At the basic of theses work, a hydrological time sequence basic arithmetical groupware flexible integrating technology project is designed to realise the target of basic arithmetical groupware flexible integrating, and the rationality and feasibility of the project are validated. Consequently the software reusing level in the area is improved, which provides flood prevent information and flood control decision gist.

Keywords: Hydrological Data Sequence, Data Mining, EJB, J2EE, Flexible Integration

1. INTRODUCTION

With the wide spread of the Compute Information System and data acquiring technologies such as large scale storage technology, people accumulated large quantities of data, most of them presented as time sequence, in daily affair managements and science researches. And researchers always pay lots of attentions to them, which has become a hotspot for both theory and practical in research task. Especially in water conservancy field, there are large quantities time sequence data in the database. And the research for similarity of the hydrological time sequence has important meanings in flood forecasting, preventing and dispatching. So hydrological data mining emerged. And the purpose for time sequence similarity search, also called similarity query, is to find out the similar sequence in time sequence database and given sequence model or query similar sequence pairs in databases. So it is not only the useful data analysis tool but also the base for data mining application, such as clustering, classifying and association rules.

This paper focuses on the research and comparison for hydrological time sequence, including lot types of basic mining algorithms that encapsulated as EJB groupware, for example: data pretreatment, character pattern mining and comparability measurement. Then the most important step is designing an extensible and flexible integrating frame for its basic arithmetical groupware according to J2EE standards. On base of this frame, we integrate these groupware and then design a flexible integrating system for it, which faces water conservancy field.

2. BASIC ALGORITHMS FOR HYDROLOGICAL TIME SEQUENCE DATA MINING

2.1 Hydrological Time Sequence Data Mining

First of all, we introduce several definitions:

- Data Mining[1]: pick up knowledge people interested in from large scale databases. And this knowledge is connotative, unknown but latent useful. The distilled knowledge can be represented as Concepts, Rules, Regularities, and Patterns. So it is uncommonness manage process for pick up correct, original and latent useful groupware that finally can be understood by people.
- Hydrological Data Mining: application for data mining in hydrology field. It is a process for picking up crytic and useful hydrological information and knowledge from abundant, incomplete, noisy, blurry and random hydrology related data.
- Time Series[2]: refers to the set of observation results, which arranges accord to the time sequence. Base on the difference between the research issue and phenomena, we can get all kinds of time sequence. And the time sequence data occupied a relatively large proportion in today’s computer memory.
- Time Sequence Similarity Search[3]: also called similarity query, is an important task for time sequence data mining. And the purpose for this search is finding out the similar sequence in time sequence database and given sequence model or query similar sequence pairs in the databases.

The time sequence similarity search is not only the useful data analysis tool but also base for data mining application, such as clustering, classifying and association rules, etc. So this research has an important meaning for both academic and practical.

Because most of the hydrological data are represented as hydrological time sequence, hydrological data mining is mainly used for time sequence data mining. And it needs to closely combine with hydrological data analysis and management. It also has to completely apply compute software technology, database, data warehouse and other latest data processing technologies.

According to the practical and research needs for hydrology, hydrological time sequence data mining can be divided into similarity search, sequence pattern mining and period analysis as its function. And if divided as its continuous process, it can also be divided into three continuous mining processes: data pretreatment, character pattern mining and similarity measurement.

2.2 Basic Algorithms For Hydrological Time Sequence Data Mining

2.2.1 data pretreatment algorithms

In most cases, hydrological time sequence is incomplete, inconsistent and existing missing, noisy and data redundancy. But data pretreatment can improve on them and then help to enhance the precision and performance of the mining process. So data pretreatment[4] is an absolutely necessarily step in the data mining process. The pretreatment is a time-consuming work, which often takes 70% to 80% work of the total data mining task. And the hydrological time sequence data pretreatment work includes data integrating, data selection, data cleaning and data transformation and so on. Because of the equipment or man-made reasons, the gathered time sequence often exists missing and noisy. But data cleaning can “clean” the data through the processes of filling in the missing value, smoothing noisy data, identifying and deleting isolated points and then solve variances. So it won’t influence the accuracy of the query result.

Missing value, which is usually generated by the lapse of the collection or transmission equipment in the data gathering process and man-made omissions, is the missing of the time sequence. If we neglect this missing data or simply replace it with a static constant, for example “0”, it will bring the similarity query into chaos and lead to trustless output. So we have to fill in the missing before the query process. The most frequently adopted filling method is handwork, which fills the black with the average for the other time data of the sequence, the average for the data of other relative sequences or most possible value.

Noise is the random error or deviation of the measured value. It is usually worked by unexpected mistakes, such as the inaccuracy of the equipment, the computer or man-made input errors or mistakes happened in the transmission process. And the methods used for smooth the time sequence noisy includes filtering, moving averages, peak-valley-mean smoothing, median smoothing and binning and etc.

Data pretreatment can improve the quality of the data and then help to enhance the precision and performance of the mining process. So it is an absolutely necessarily step in the data mining process and we often use filling in missing data and smoothing noisy to realize it.

2.2.2 Character pattern mining algorithms

Sequence pattern mining[5] refers to the mining for patterns that are relatively time or other high frequency appearing patterns. And existed pattern discovery methods for time sequences mainly focused on static pattern, so the frequent measurement for it generated through scanning all records of it. However, the frequent local patterns in a period exist widely in practical. For example, pattern like “in the subsequence S, if A occurs, then B will occur in T” can provides better knowledge about hydrological phenomena evolutions.

A good time sequence character pattern mining method should have the following characters: First it can find the change pattern that can really represent time sequence. And this pattern not only has graphics sensibility cognition but also in favor of description and application in practical. Secondary, it should be efficient and simple, which make it applicable for magnanimity time sequence mining.

But ordinary methods only carry on changes on subsequences, which mapping subsequences to character space through Fourier transform, wavelet transform or piecewise average approximation. The standard for determinant is if the distances between the subsequences in the character space is smaller than the given threshold, which is used for lookup similar subsequences. And the key for the problem is the transform method and the adaptability for different hydrological factor processes. And we often use Piecewise Average Approximation, PAA in short, to realize reducing dimension.
Character pattern mining can find out the change pattern that really represents the time sequence change. And such pattern not only has graphics sensibility cognition but also in favor of description and application in practical.

2.2.3 Comparability measurement algorithms

Comparability search means finding out the most comparable sequence with the given one. And finding out all sequences comparable with the given one is called subsequence matching. So comparability search for hydrological sequence is finding out all types of comparable subsequences in it. And the hydrological data represented by it contains the evolution processes and tendency for the climate and the earth surface.

The key problem for hydrological time sequence comparability mining is confirming the measure of comparability, which means different hydrological factors should have different characters. And according to the characters for the hydrological process, confirm the measure of comparability fit it most is the important to the pledge for usefulness for the mining result.

We adopt Euclidian distance algorithm to judge the comparability for the changes of the two sequences. If the calculated Euclidean distance[6], represented as D, is smaller than the given threshold, using w to represent, then we can think the two sequence is comparable with each other.

Through comparability measurement and character for hydrological process, we can confirm its corresponding comparability measurement. And it is the important condition that ensures the usefulness for the mining result.

3. BASIC ARITHMETICAL GROUPWARE and ITS FLEXIBLE INTEGRATING

3.1 Corresponding Groupware Technology

Pattern design refers to virtual scheme used to solve the repeat occurring design problems. It is the convenient method for reuse for projects and programmers. It can separates classes, prevents them knowing too much about each other, so it can also prevent tight coupling. And it’s the description for the communication objects and classes that are customized for solve general design problem in particular instance.

Session Façade pattern, one of the seven design patterns, Sun Java Center use them to point out the business layer. We should treat with the complicated EJB, business encapsulation logical and data when the serve is compiling the service layer groupware. But the provide interface is very complicated; the service layer will become quite complicated too. However, this pattern can reduce the complicity and it is quite useful when needs to hide the details for object interaction of the business layer. Under such pattern, the business object, which provides business logical management service, is realized by session Bean. And the entity Bean is used to realize the object using for represent the sharing lasting memory view among several sessions. If the clients interact with the business objects directly, “tight” coupling will occur.

Service Locator, pattern used for draw out all JNDI, hind the initialization for context, lookup for EJB object and narrow operation. All clients of the system use the same code in order to reducing code complicity, providing consistence control and using Cache to improving performance. It also can extract complicity. Then it centralized treated the unified service access, provided by system clients, through encapsulating query and complicity of the create process. And it is useful for appending new business groupware, enhancing net performance and improving capability by Cache.

3.2 Ejb Algorithms Groupware

The manager groupware for hydrological time sequence data mining algorithm, which will be called as arithmetical manager in the following for short, is the key part for achieving groupware's flexible integrating. It is used for unified answering the request form clients and providing query and locating, appending, deleting and calling for the basic data mining arithmetical groupware. At the same time, it also provide access method, which used for accessing basic arithmetical groupware warehouse, historical database and real time database, and return relative result to the client. Different types of basic data mining algorithms, which used to realize different basic arithmetical groupware warehouses, provide relevant remote interface, which facilities the clients' dispatch, can expand data mining algorithms conveniently and flexibly. Because arithmetical manager, the session façade, encapsulated the basic mining algorithms, client can only interact with entity Bean through the arithmetical groupware session Bean, which is called by Session Façade. And we adopt strategy pattern, one of the object behavior patter for Java design pattern, to realize the flexible integrating for the arithmetical groupware.

3.2.1 Basic arithmetical groupware

1. Interface of Groupware Abstract Class

To realize the flexible integrating for the basic arithmetical groupware, according to JNDI Name, its given label name, the lookup and orientation, the system have to find relevant local Home interface for Session Bean, instant a remote interface with create(), and then call its specific business logical method. So we should define relevant abstract class interfaces, including implement class, Home interface and remote interface, for basic arithmetical groupware including data pretreatment, character pattern mining and comparability measurement.

Abstract Check Data Class Interface: includes an implement class called abstractCheckDataBean, a Home interface called abstractCheckDataHome and a Remote one called abstractCheckData. The following Fig.1 shows the class structure for implement class abstractCheckDataBean.

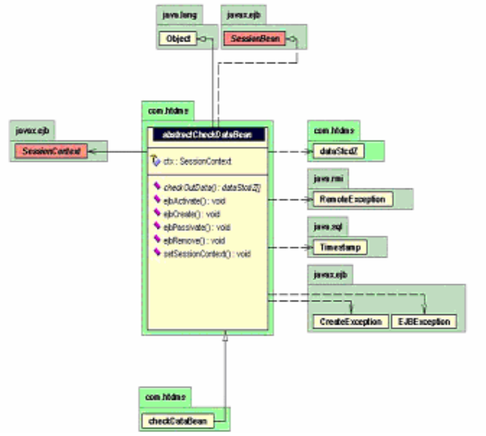


Fig.1. class structure for abstractCheckDataBean.

Abstract Character Pattern Mining Groupware Class Interface: includes an implement class called abstractCharacteristicModeBean, a Home interface called abstractCharacteristicModeHome and a Remote one called abstractCharacteristicMode. The above Fig.2 shows the class structure for implement class abstractCharacteristicModeBean.

Abstract Comparability Measurement Groupware Class Interface: includes an implement class called abstractSymmCompareBean, a Home interface called abstractSymmCompareHome and a Remote one called abstractSymmCompare. The following Fig.3 shows the class structure for implement class abstractSymmCompareBean.

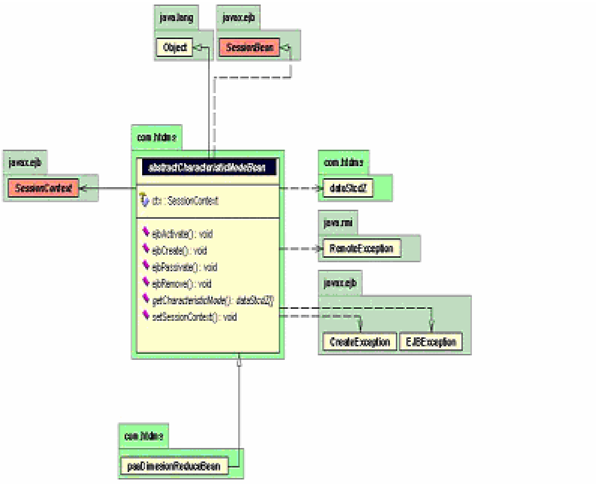


Fig.2. Class structure for abstractCharacteristicModeBean.

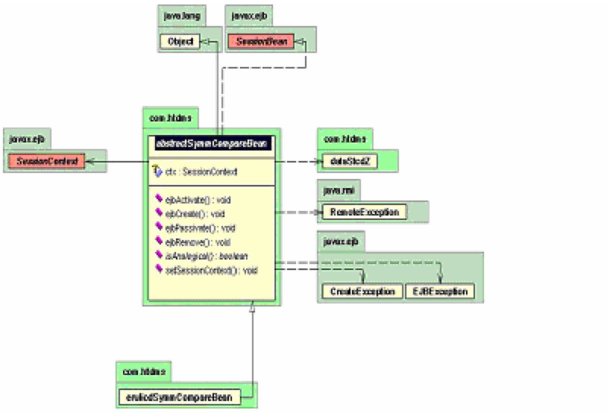


Fig.3. Class structure for abstractSymmCompareBean.

2. Criterion for groupware interface and realization for groupware

To realize the function for different basic arithmetical groupware, we should land in flexible integrating system and download above arithmetical groupware interface frame. Then program session Bean for it, which follows the following steps:

- Realization for basic arithmetical groupware session Bean has to inherit the implement class of relevant type abstract groupware class. And that is data pretreatment groupware class xxxxBean extends abstractCheckDataBean, character pattern mining groupware class xxxxBean extends abstractCharacteristicModeBean, comparability measurement groupware class xxxxBean extends abstractSymmCompareBean.

- Business logical method in basic arithmetical groupware session Bean must consistent with the one in the abstract groupware class interface. That is to say the data pretreatment groupware public com.htdms.dataStcdZ[] checkOutData(Timestamp beginDt, Timestamp endDt, com.htdms.dataStcdZ[] sourceData), character pattern mining groupware public com.htdms.dataStcdZ[] getCharacteristicMode(int k, dataStcdZ[] checkData), comparability measurement groupware public boolean isAnalogical(com.htdms.dataStcdZ[]

desData1,com.htdms.dataStcdZ[] desData2,double eThreshold).

- Some relevant modifications on ejb-jar.xml, the disposal description document of basic arithmetical groupware session Bean. I is the label <enterprise-beans>

<session>
……
</session>

And some modifications also should be done on the appointed Home interface of <home>…</home> and the appointed Remote interface of <remote >…</remote > that are both in </enterprise-beans>. All of these modifications are shown in the following table:

Groupware Type	**<home>…</ home>**	**<remote >…</remote >**
Data pretreatment	com.htdms.abstractCheckDataHome	com.htdms.abstractCheckData
Character Pattern Mining	com.htdms.abstractCharacteristicModeHome	com.htdms.abstractCharacteristicMode
Comparability Measurement	com.htdms.abstractSymmCompareHome	com.htdms.abstractSymmCompare

3.3 Flexible Integrating Frame

3.3.1 Frame structure

Integrate all discussion and research about flexible integrating technology for groupware in this paper, we design a flexible integrating frame for the integrating and opening basic time sequence mining algorithm, which is hydrological-oriented. And the frame is shown in Fig.4.

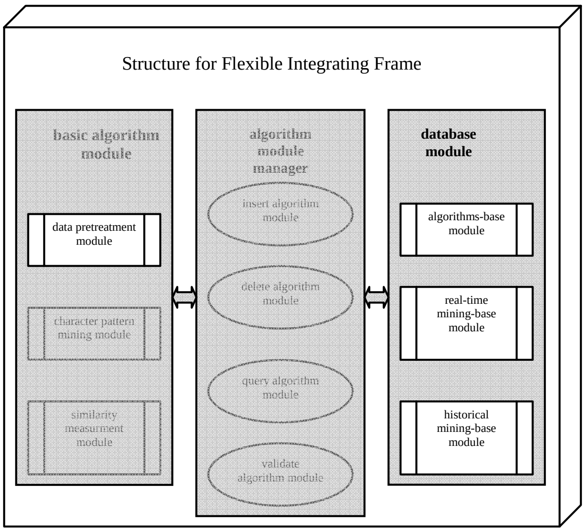

Structure for Flexible Integrating Frame
basic algorithm module
data pretreatment module
character pattern mining module
similarity measurment module
algorithm module manager
insert algorithm module
delete algorithm module
query algorithm module
validate algorithm module
database module
algorithms-base module
real-time mining-base module
historical mining-base module

Fig.4. Structure for Flexible integrating frame

From the figure we can find out that the frame contains the following parts:

- Basic Arithmetical Groupware: contains three basic type of arithmetical groupware: data pretreatment, character pattern mining and comparability measurement, which are responsible for the management and controlling for each type.

- Arithmetical Groupware Manager: key groupware for realize the flexible integrating, contains specific business logic including flexible integrating, management used for basic arithmetical groupware. Made up of four groupware: appending arithmetical, deleting arithmetical, querying arithmetical and validating arithmetical. Since executing the flexible integrating mechanism for the groupware, it realizes the expansibility for the groupware, which is plug and play and flexible calling.

- Database Groupware: contains three kinds of database groupware: arithmetical corresponding to the arithmetical warehouse, is used for manage information about all basic algorithm in the system; real time mining corresponding to real time warehouse, manages all need mining data in the real time warehouse; And historical mining corresponding to the historical warehouse, manages all need mining data in the historical warehouse.

3.3.2 Function and character for the frame

The main function for the flexible integrating frame is providing a flexible integrating frame for basic time sequence mining arithmetical groupware in hydrological field and integrating all types basic arithmetical groupware in a not only expandable but also plug and play frame, which can realize the “seamless” association among groupware. To achieve the appending function, all groupware according with the interface standards can be added into the frame and enriched the groupware. Achieving the deleting function for it, we can wipe off the unqualified groupware to prepare updating for the groupware in the future. Then comes to the calling function, client can call all functions of the algorithm to achieve the data mining goal through visible interface without knowing the details in the arithmetical groupware. And the querying and location function, which provides the services such as calling and deleting, is mainly run on the background. At last, the validating function is used for validate the client-designed arithmetical groupware to ensure its validity.

So the flexible integrating frame has following characters: accord with J2EE technology standards, expendable, retractility, plug and play, “seamless” association, providing normal interface in specific field, easy to maintenance, reusable and use simple.

4. DESIGNING AND REALIZATION FOR FLEXIBLE INTEGRATING GROUPWARE SYSTEM

This system based on J2EE platform and taking EJB groupware technology as its center adopts the application scheme with multilevel B/S pattern and provides a business logical level centered system structure. It also sharply decreases the difficulty for system developing, which facilities the researchers to devote themselves into the research of the business issues. So it is dependability and transplantable making it easier for the maintenance and upgrade. And through encapsulation and flexible integrating for hydrological time sequences basic algorithms, it is much easier for the evolvement and rebuilding the system frame. As a result, the period for developing a system becomes shorter, while its expansibility is better and charge for maintenance and upgrade is lower. It can select, combine and replace the arithmetical groupware or expand new function more easily. So it realizes the goal for plug and play, flexible integrating and software reusing. It supports the industry and enterprise flooding forecasting and dispatching in hydrological field comprehensively, which has a broad application foreground[9]. The system adopts four levels: represent level, controlling, business logical and data operating. They are mainly made up of arithmetical groupware manager, three database arithmetical groupware including basic, historical time sequence and real time sequence, and four function arithmetical groupware that is registering, deleting, group flexible integrating and downloading interface frame.

This system has already been tested on the base of hydrological time sequence data sets and the result shows that it can beautifully realize the requirement for the flexible integrating of the groupware. It has a good expansibility and a lower charge for maintenance and upgrade. It also can select, combine and replace the hydrological time sequences arithmetical groupware or expand new function of the groupware more easily. And it realizes the goal for plug and play, flexible integrating and software reusing. So it supports the industry and enterprise flooding forecasting and dispatching in hydrological field comprehensively.

5. CONCLUSIONS

This article researches and compares different algorithms for the comparability for hydrological time sequences, including data pretreatment, character pattern mining and comparability measurement. Then it proposes a technical scheme for solving the flexible integrating for time sequences algorithms. And it encapsulated them as EJB groupware, and then designs an extensible and flexible integrating frame for hydrological time sequence basic arithmetical groupware according to J2EE standards. On the base of the frame, it integrates these groupware and then designs a flexible integrating system for hydrological time sequence, facing water conservancy field. After experiment, we validated the validity and correctness for the scheme. So its dependability and transplantable making it easier for maintenance and upgrade. And through encapsulation and flexible integrating it is much easier for evolvement and rebuilding. As a result, the period for developing a system becomes shorter, while its expansibility is better and charge for maintenance and upgrade is lower. It can select, combine and replace the arithmetical groupware or expand new function easily. So it realizes the goal for plug and play, flexible integrating and software reusing. It supports industry and enterprise flooding forecasting and dispatching in hydrological field comprehensively, which has a broad application foreground.

REFERENCES

[1] Fayyad,U., Piatetsky-Shapiro, G.&Smyth,P. *From data mining to knowledge discovery*[M], AAAI Press/MIT Press. 1996.

[2] Provinelli, R.J,*Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of Time Series Events[J]*, Marquette University, Milwaukee, 1999.

[3] H.Mannila, H.Toivonen, and A.I. Verkamo,"Discovering Frequent episodes in Sequences[C],"in *Proc. Of KDD-95*, Montreal, Canada, Aug.1995.

[4] R.Agrawal, T.Imielinski,"Ming association rules sets of items in large database[C],"in *Proceedings of the ACM SIGMOD Conference on Management of ACM*, 1983.

[5] Dina.Goldin,Paris C.Kanellakis,*On Similarity Queries for Time-Series Data[J]:Constraint Specification and Implementation*, CP 1995.

[6] Christopher Alexander,*A Pattern Language[M]: Towns, Buildings, Construction*,Oxford University Press,1977.

[7] Linda G.DeMachiel,*Enterprise JavaBeans Specification[M]*, Version 2.0.Sun Microsystem,2001.

[8] Guijun Wang, Liz Ungar, Dan Klawitter,"Component Assembly for OO Distributed Systems[J],"in *IEEE Computer*,1999,32.

Incremental Mining Item Sets Based on H-Mine and XML *

Xingjie Feng Jing Zhang
College of Computer Science and technology
Civil Aviation University of China, Tianjin, China
Email: fxingjie@163.com

ABSTRACT

Efficient discovery of association rules in large database is a well studied problem and several approaches have been proposed. However, the previously proposed methods still encounter some performance bottlenecks when mining databases is updated, such as inserted and deleted. In this paper, we propose an incremental updating technique based on H-mine and xml, for the maintenance of association rules when new transaction data is added to a transaction database. A performance evaluation shows that our algorithm is available and scalable.

Keywords: Incremental, H-mine, XML, Association rules

1. INTRODUCTION

Mining frequent item sets is a key step in many data mining problems, such as association rule mining, sequential pattern mining, classification, and so on. Since the pioneering work in [3], the problem of efficiently generating frequent item sets has been an active research topic.

There have been many algorithms developed for fast mining of frequent patterns, which can be classified into two categories. The first category, candidate generation and-test approach, such as Apriori [4] as well as many subsequent studies is directly based on an anti-monotone *Apriori* property [4]. The *Apriori* algorithm achieves good reduction on the size of candidate sets. However, when there exist a large number of frequent patterns and or long patterns, candidate generation-and-test methods may still suffer from generating huge numbers of candidates and taking many scans of large databases for frequency checking.

Another category of methods, *pattern-growth methods*, such as *FP-growth* [9] and *TreeProjection* [2] have been proposed. A pattern-growth method uses the *Apriori* property. Instead of generating candidate sets, it recursively partitions the database into sub-databases according to the frequent patterns found and searches for local frequent patterns to assemble longer global ones.

Nevertheless, these proposed approaches may still encounter some problems. In real environment, real databases contain all the cases. Real data may be updated. For this situation, a simple method is to re-compute the frequent itemsets of the whole updated database. This is clearly inefficient because all the computations done initially for finding the old large itemsets are wasted.

This poses a new challenge: “Can we find a better method which is (1) with minimal re-computation when new transactions are added or deleted from the database, and (2) space requirement is small, even for very large databases?”

Our algorithm is based on H-mine [1] and XML[5]. The most difference is that we develop a simple memory-based hyper-structure, IH-struct instead of H-struct. Then we save the IH-struct as XML document on the hard-disk. When the new data sets inserted, we only need to update the XML document

to generate the new IH-struct by scanning the new data sets, not to handle the whole data sets. In the way, we can update the association rules incrementally. Experimental results show that, in many cases, H-mine has very limited and exactly predictable space overhead and is faster than memory-based Apriori and FP-growth.

The remaining of the paper is organized as follows. In Section 2, we briefly introduce the association rules algorithm H-mine and XML is introduced in Section 3. The incremental H-mine for updating databases is proposed in Section 4 and extensive performance evaluation is reported in Section 5. Section 6 concludes with a summary and some directions for future research.

2. H-Mine (Mem): Memory-based Hyper-Structure Mining

In this section, H-mine is introduced and in Section 3, the method is extended to handle incremental data mining.

We first define the problem of frequent pattern mining [1].

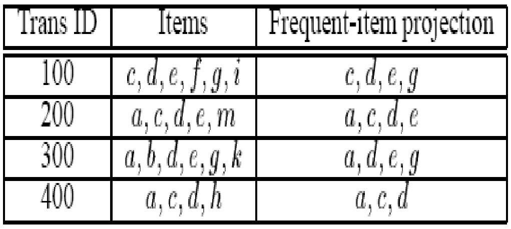
Definition 2.1 Let I={ $x_1,\ldots x_n$ } be a set of **items**. An **itemset** X is a subset of items, i.e.,
X $\subseteq$ I. For the sake of brevity, an itemset X= $\{x_1, x_2, \ldots x_m\}$ is also denoted as X= $x_1 x_2 \ldots x_m$.
A transaction T= (tid, X) is a 2-tuple, where tid is a transaction-id and X an itemset. A **transaction** T= (tid, X) is said to contain itemset Y if and only if Y $\subseteq$ X. A **transaction database** TDB is a set of transactions. The number of transaction in TDB containing itemset X is called the **support** of X, denoted as sup(X). Given a transaction TDB and a support **threshold** min_sup, an itemset X is **frequent pattern**, or a **pattern** in short, if and only if sup(X)≥min_sup.

The **problem of frequent pattern mining** is to *find the complete set of frequent patterns in a given transaction database with respect to a given support threshold.*
The general idea of H-mine is as follows.
Example 1 Let the first two column of Table1 be our running transaction database TDB. Let the minimum support threshold be min_sup=2

Following the *Apriori* property [3], only frequent items play roles in frequent patterns. By scanning TDB once, the complete set of frequent items{a:3, c:3, d:4, e:3, g:2} can be found and output. Let *freq(X)* be the set of frequent items in itemset. For the ease of explanation, the frequent-item projections of all the transactions of Table 1 are shown in the third column of the table.

*This work was supported by the National Natural Science Foundation of China （No. 60672174）

Table1. The transaction database TDB as our running example.



Trans ID	Items	Frequent-item projection
100	c, d, e, f, g, i	c, d, e, g
200	a, c, d, e, m	a, c, d, e
300	a, b, d, e, g, k	a, d, e, g
400	a, c, d, h	a, c, d

If the frequent-item projections of transactions in the database can be held in main memory, they can be organized as shown in Fig.1. Following the alphabetical order, all items in frequent-item projections are sorted according to the *F-list*.

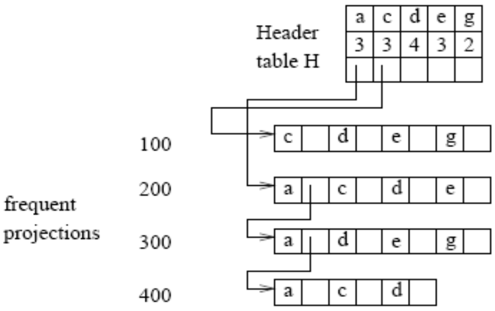


Fig.1. H -struct.

A *header table H* is created, where each frequent item entry has three fields: an *item-id*, a *support count*, and a *hyper-link*. When the frequent-item projections are loaded into memory, those with the same first item (in the order of *F-list*) are linked together by the hyper-links as a queue, and the entries in header table H act as the heads of the queues. The d-, e-, g-queues are empty since there is no frequent-item projection that begins with any of these items;

First, to find the set of frequent patterns containing item a. For doing this, an a-header table Ha is created, as show in Fig.2. In Ha, every frequent item, except for a, has an entry with the same three fields as H, i.e., item-id, support count and hyper-link. The support count in Ha records the local support of the corresponding item in the a-projected database. For example, item c appears twice in a-projected database, thus the support count in the entry c of Ha is 2.

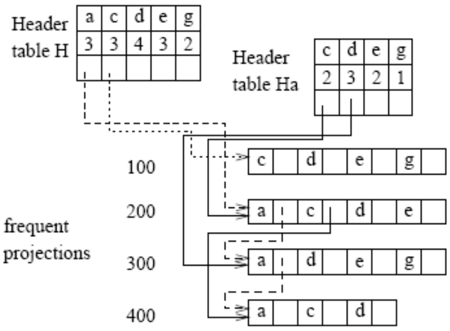


Fig.2. Header table Ha and ac-queue

By traversing the a-queue once, the set of locally frequent item in the a-projected database is found, which is {c:2, d:3, e:2}. g is not here since it is not locally frequent. This scan output the frequent patterns {ac: 2, ad: 3, ae: 2} and builds up links for Ha header as shown in Fig.2.

Similarly, the process continues for the ac-projected database by examining the c-queue in Ha, which can creates an ac-header table Hac, containing d and e.Since only item d:2 is locally frequent item in the ac-projected database, only acd:2 is output, and the search along this path completes.

Then the recursion backtracks to find patterns containing a and d but no c. Since the queue started from d in the Header table Ha, i.e., the ad-queue, links all frequent-item projections containing items and d (but excluding item c in the projection), one can get the complete ad-projected database by inserting frequent-item projections having item d in the ac-queue into the ad-queue. This involves one more traversal of the ac-queue. Each frequent-item projection in the ac-queue is appended to the queue of the next frequent item in the projection according to F-list.

Similarly, the mining goes on. It is easy to see that the above mining process finds the complete set of frequent patterns without duplication.

Notice also that the depth-first search for mining the first set of frequent patterns at any depth can be done in one database scan by constructing the header tables at all levels simultaneously.

3. XML (Extensible Markup Language)

Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML (ISO 8897).Originally designed to meet the challenges of large-scale electronic publishing. XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere.

DOM (XML Document Object) defines a standard way for accessing and manipulating XML documents. It is a platform and language-neutral interface that allows programs and scripts to dynamically access and update the content, structure, and style of a document.

The DOM presents an XML document as a tree-structure (a node tree), with the elements, attributes and text defined as nodes. e.g., Fig.3 as follows.

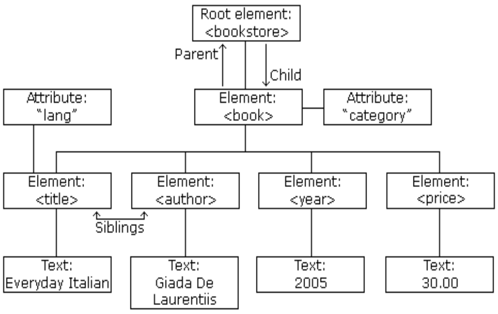


Fig.3. An XML Document Tree

4. Incremental H-Mine (IH-Mine): efficient mining handling database updated

In this section we develop an efficient method based on H-mine and XML for updating the association rules when database is updated. The database update effectively means addition of new transactions to the database, since in the data warehouse the data is not deleted frequently. Assuming that the two thresholds, minimum support and confidence, do not change, then the update problem can be reduced to finding the new set of frequent itemsets. After that, the new association rules can be computed.

The general idea of IH-mine is illustrated in the following example.

4.1 Create the header table IH(Incremental Header table)

In Example 1, the header table H only contains the frequent items, but in our IH-struct, it contains all the items since an item not frequent in the database TDB may be frequent when a Δ TDB is added. So we must keep it in the header table. The header table IH of Example1 is shown as follows:

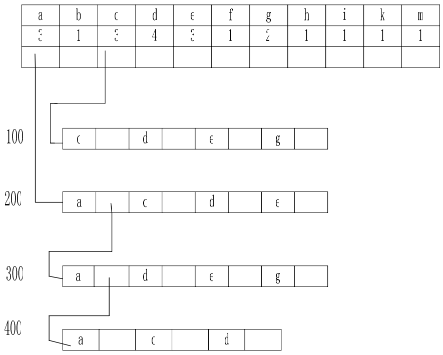


Fig.4. Header table IH for TDB

A header table IH is created, where each item entry has three fields: an item-id, a support count, and a hyper-link. When the item projections are loaded into memory, those with the same first item (in the order of F-list) are linked together by the hyper-links as a queue, and the entries in the header table IH act as the heads of the queues. We can build it by scanning the database twice.

4.2 Keep the IH as XML document on the hard-disk

In H-mine, the header table is in the (main) memory. So when Δ TDB is added, the header table of TDB can not be used. Here we keep the header table of TDB on the hard-disk by transferring it to an XML document since XML document has been a new standard for information representation and exchange, with the following progress.

First, a memory-based, tree structure T is proposed. It is proverbial that an XML document presented as a tree structure in the memory. So the XML document corresponding to the tree can be simply constructed. In this way we transfer the header table to XML document successfully.

Here, every item in the header table IH is a child of root. Each item node has two attributes, name and support respectively. The item node also has several children those with the first item (in the order of F-list). The children of an item node stand for the transactions in the F-list, also having one attribute, id. Finally, the leaves of T store the transactions. e.g. Fig.5 as follows:

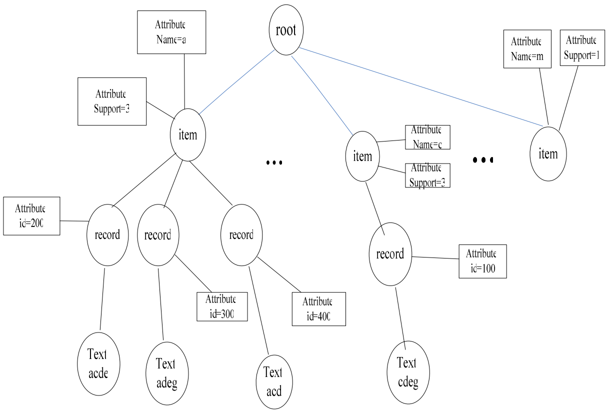
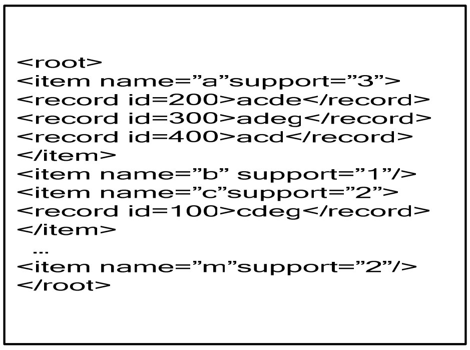


Fig.5. T of header table IH

The root has 11 children since there is 11 items in the IH. The first children of root standing for item a, we call it item node of a. Item node of a has three children since its F-list contains three transactions. The item of m has no children since its F-list is null. The corresponding XML document appears in Fig.6:



```
<root>
<item name="a"support="3">
<record id=200>acde</record>
<record id=300>adeg</record>
<record id=400>acd</record>
</item>
<item name="b" support="1"/>
<item name="c"support="2">
<record id=100>cdeg</record>
</item>
  ...
<item name="m"support="2"/>
</root>
```

Fig.6. An XML document of header table IH for TDB

Second, we prove that the operation on the IH and on the XML document is similar. The information in the first two lines of header table has been kept as the attributes (name and support) of item element in the XML document. The entries in the header table acting as heads of queue can be implemented by accessing first child of item node. We can also access siblings of item nodes one by one instead of accessing the queue. So we can say that the header tables IH and XML document have the same function.

The pseudo-code for transferring the header table to XML document appears as follows.

Algorithm 1 TransferToXML (IH)
Input: IH: header table
Output: an XML document

```
for each item i in IH
output i as an element to the XML document.
    if (i.link!=null)
      for each transaction t in a.queue
              output t as an sub-element
         of i to the XML document.
              end for
        end if
        end for
    return the XML document.
```

4.3 Handling database updated

As shown above, an XML document is created and can be saved on the hard-disk. When a Δ TDB is inserted, we only need to compute Δ TDB and update the XML document, not to handle the whole database.

Example 2 Let the first two column of Table2 be our running transaction database Δ TDB. Let the minimum support threshold be min_sup=2.

Table2. The transaction database Δ TDB as our running example

Trans ID	Items	Frequent-item projection
500	a, b, d, e, i	a, d, e
600	c, d, e, m	c, d, e, m
700	a, c, k, m	a, c, k, m
800	d, e, k	d, e, k

Based on H-mine, by scanning Δ TDB once, the complete set of frequent items
{a:2, c:2, d:3, e:2, k:2, m:2}can be found and output. But a, c, d, e have been computed in TDB and are all frequent. In the Δ TDB since the min_sup do not change and they are still frequent.

Clearly, the locally support of i and b is 1 and not frequent, but we should still consider them since in the whole database (TDB+ Δ TDB) these two items may be frequent, so the finally result is
{a:2, b:1, c:2, d:3, e:2, i:1, k:2, m:2}. The following alphabetical order is a-b-c-d-e-i-k-m.

As shown in Section 4.1 and Section 4.2, we build the header table IH and XML document for Δ TDB.

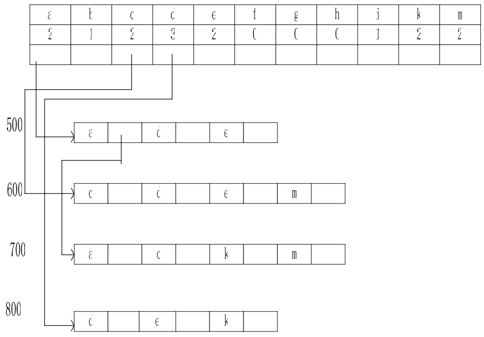


Fig.7. Header table IH for Δ TDB

```
<root>
<item name=” a” support=” 2” >
<record id=” 500” >ade</record>
<record id=” 700” >ackm</record>
</item>
<item name=” b” support=” 1” />
<item name=” c” support=” 2” />
<record id=” 600” >cdem</record>
</item>
<item name=” d” support=” 3” />
<record id=” 600” >dek</record>
</item>
…
<item name=” m”  support=” 2” />
</root>
```

Fig.8. An XML document for Δ TDB

Now we can incorporate two XML documents in Fig.6 and Fig.8 into one to build the new header table for (TDB+ Δ TDB), we can finish it in two steps.

Step one: **LoadToMemory**

It is very simple to load two documents into memory since DOM interface has been implemented by many programming language, such as Java and C#, here we use C#.net.

Step two: **ConstrctNewXML**

Actually, it is the operation on the tree in the memory.

The pseudo-code of whole process document appears as follows.

Algorithm 2 UpdateXML (F1, F2)

Input: F1: XML document for TDB, F2: XML document for Δ TDB

Output: new F1

```
    LoadToMemory
    for each item node n1of F1 and n2 of F2
        if (n1.name==n2.name)
      set n1.support=n1.support+n2.support
          for each child d of n2
              if(d not in the F1’s child list)
                  insert every children of F2
                  into F1’s child-list
          end for
      end if
  end for
  return F1
```

A new XML document for (TDB+ Δ TDB) is created. We can use it to mine frequent itemsets as show in H-mine by loading it into memory using DOM. Limited by space; we omit the specific process as well as a complexity analysis of the algorithm. The new XML document is shown as follows.

```
<root>
<item name=” a” support=” 5” >
<record id=” 200” >acde</record>
<record id=” 700” >acfkm</record>
<record id=” 300” >adeg</record>
<record id=” 500” >abdei</record>
<record id=” 400” >acd</record>
</item>
<item name=” b” support=” 2” />
<item name=” c” support=” 5” />
<record id=” 600” >cdek</record>
<record id=” 100” >cdeg</record>
</item>
<item name=” d” support=” 7” />
<record id=” 800” >dek</record>
</item>
…
<item name=” m”  support=” 3” />
</root>
```

Fig.9. New XML document

5. Performance Study

The most feature of our incremental H-mine algorithm is that the whole database (TDB+ Δ TDB) is not scanned. When a Δ TDB inserted, we only scan the Δ TDB once. Then updating the xml document of TDB, this process is handled in the memory, thereby reducing the I/O requirements drastically.

The experiment is performed on a 1.8GHz Pentium PC machine with 512megabytes main memory and 60G hard disk, running Microsoft Windows/NT. IH-mine and H-mine are implemented by us using C#. All reports of the runtime of IH-mine and H-mine include both the time of constructing H-struct and mining frequent-patterns. They also include both CPU time and I/O time.

To test the efficiency and scalability of the algorithms on mining very large databases, we use data set recorded during flight. It has 1,000,000 transactions, while there are 15 items per transaction.

Since the H-mine cannot achieves the incremental data mining, for the sack of the contract, suppose each group of data is updated 10 times. After every time updated, we must run H-mine on the new data set to get new result. Its time is 10 running time in all. Let the minimum support threshold be min_sup=0.1.Experiment result shown as follows.

From the Fig.10, we can see that the larger the database is, the faster the IH-mine is. That is because H-mine has to scan the whole database when Δ TDB is added**,** with the database updated; the process of scanning is slower and slower. While IH-mine only scans the Δ TDB, when the size of Δ TDB is not changed, the time of this process is almost a constant.

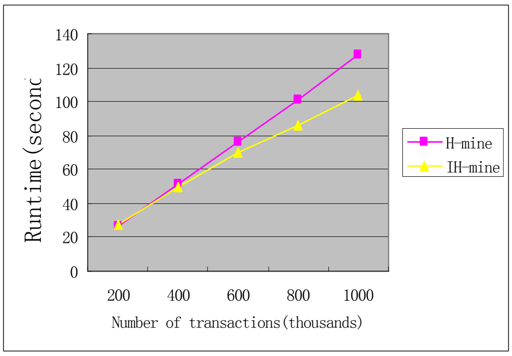


Fig.10. Experiment result

6. Discussion and Conclusions

In this paper, we have proposed a incremental frequent mining algorithm (IH-mine) based on H-mine and XML, which takes advantage of H-struct data structure and platform-independent of XML. As shown in our performance study, it performances well when database is updated.

The most difference between IH-mine and H-mine is that IH-mine keeps the header table on the hard-disk as XML document, while H-mine keeps it in the memory. When database updated, H-mine has to re-compute and cost is high.

Base on the above analysis, one can see that IH-mine represents a new, efficient and incremental mining method. Further, our algorithm is applicable to deletion of transactions.

REFERENCES

[1] Jian Pei, Jiawei Han and Hongjun LU.H-Mine: Hyper-Structure Mining of Frequent Patterns in Large Databases

[2] R.Agrwal, C.Aggarwal, and V.V.V.Prasad. A tree projection algorithm for generating of frequent itemsets. In J.of Parallel and Distributed Computing (Special Issue on High Performace Data Mining), 2000.

[3] R.Agrawal, T. Imielinski, and A. Swamin. Mining association rules between sets of items in large databases. In SIGMOD'93, Washington, D.C., May 1993.

[4] R.Agrawal and R. Srikant. Fast algorithms for mining association rules. In VLDB'94, pp. 487-499, Santiago, Chile, Sept.1994.

[5] http:// www.w3c.org

[6] J.Han, J.Pei and Y.Yin. Mining frequent patterns without candidate generation. In SIGMOD'00. [9]J. Han, Micheline Kamber.Data Mining Concepts and Techniques, China Machine Press,2002.

[7] S.Thomas,S.Bodagala,K.Alsabti,et al.An efficient algorithm for incremental updation of association rules in large database. Proceedings of 3rd Int. Conf. on Knowledge Discovery in Databases,1997.

[8] FENG Xingjie， HUANG Yalou. Research on the Inheritable Problem in Data Mining. Information and Control, 2005, 34(6):249-253.

[9] J. Han, J. Pei, and Y. Yin. Mining frequent patterns withoutcandidate generation. In SIGMOD'00, pages 1–12.

[10] J. Pei, J. Han, and R. Mao. CLOSET: An efficient algorithm for mining frequent closed itemsets. In Proc. 2000ACM-SIGMOD Int. Workshop Data Mining and KnowledgeDiscovery (DMKD'00), pages 11–20.

The Research of Electronic Commerce Systems Based on Web Server Log Mining

Lin Chen, Qingping Guo
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, China
Email: clcc169@sina.com

ABSTRACT

This paper discusses the necessity of electronic systems based on web server log mining and introduces the three steps of web mining firstly, then, the paper discovers how to mine the web server logs of electronic commerce web sites and how to utilize these mining results, as well as how to discover which content is the customer most interested in and the pattern of accessing by sequential pattern. By analyzing these data, the E-Commerce with pertinence can be developed and a mining algorithm is given.

Keywords: Web Log Mining, Data Mining, Electronic Commerce, Data Preprocessing, Mining Algorithm

1. ELECTRONIC COMERCE

Currently World Wide Web is developing rapidly in China. Some Experts forecast that the number of Chinese network users must be reach to 2 hundred million in 2010 and became the most network users in the world. Just like the American vice president Gore said: we will live in a digital earth in 21 century. Now, on-line shopping is being popularized day by day. With the environment of shopping is further improved, the market of on-line shopping will reach to 28,000 million Yuan in the near future, some experts forecasted. On-line shopping and e-bank will is accepted as a new live mode by more and more people. Electronic commerce is developed rapidly, but as one of the most important patterns in electronic commerce, the pattern of B2C is further fall behind the B2B ,an important cause to this status is that electronic commerce platform can not understand the electronic commerce, and just do some form work. Although there are thousands electronic commerce web sites, only very few could realize the function, but also say nothing of the latent function and takes full advantage of the electronic commerce. Some systems even just put some commodity on the web page and the backstage database is not existed. So it is very necessary to empolder B2C electronic commerce web site with compressive function which is suitable to Chinese situation.

The domestic market have already transformed from the seller market to the buyer market. Then, how does the network seller could stand to the invincible position? Besides to quality wins, we should make well use of the advanced technology web data mining to understand the mentality and habit of shopping to adjust the structure of web pages dynamic and improve the service, such as the interest of accessing, the frequency of accessing, the time of accessing, and so on, even the visitors themselves may be not know these kinds of information. Developing the electronic commerce with pertinency could satisfy the users' need. Holding the market pulse accurately and knows oneself and the other side, is undefeated in many battles. They hope what they see is individuality page and get the better service as far as the visitor are concerned.

2. WEB DATA MINING

2.1 The Significance of Web Mining

Data mining means extracting useful knowledge and information automatically from the mass of data. Web data mining is one of important branches in data mining. The rapid growth of Internet has pushed the research and the development of web usage mining even more into focus. It uses for solving some questions which are met in WWW. Web data mining and its application have become critical to the business world. It is mainly aims at between the web page content, the page structure and each kinds of the electronic commerce information ,using the mining method to discover the useful knowledge , helping to extract the information from WWW, improve web sites designing and develop electronic commerce effectively[1]. The web data mining to be allowed to divide into three kinds according to the objects of web mining: web content mining, web structure mining and web usage mining [2]. The object of web mining mainly includes web servers logs data, the electronic commerce information, the web page data and so on. The accessing information was recorded in the web server logs. These data include: the log of network server, the log of agent server, the log of browser, the users' synopsis, the information of registration and the users' information of session and the transaction .Such as users' IP address, visited URL, the date and time of accessing, the method of accessing (GET or POST), the results of accessing (successful, failed, error), the size of accessing information and so on. Because the users' all commercial activity and the browsing information were recorded in the web server log, So the web logs mining system could analyze the users' accessing pattern from server log or other data. Saying from the web site designer that, understanding the information of users' is extremely important. According to the users' paths of accessing and the process, the performance of the system is improved and the structure of web site is optimized, thus the rate of clicking is enhanced or used by market stratagem. Therefore it's essential to research the system of electronic commerce based on web data mining.

2.2 The Steps of Mining

The web usage mining is divided into three steps [3]: data preprocessing, pattern discovery, pattern analysis. They are displayed in Fig.1.

2.2.1 Data Preprocessing

The object of data preprocessing is provided by data gather, data gather mainly gathers various kinds of journal file, like the server, client server as well as agent server. There is a part of journal file has nothing to with the content or the form which will be mined, such as operating system and browser, the size of document and so on, therefore the irrelevant data should be removed before the mining work. Standard journal file should include three parts: source web site, goal web site and time stamp. The source web site and goal web site are a series of URL or IP addresses. The technology of data preprocessing mainly has: data cleaning, users identification, session identification, path completion and transform data.

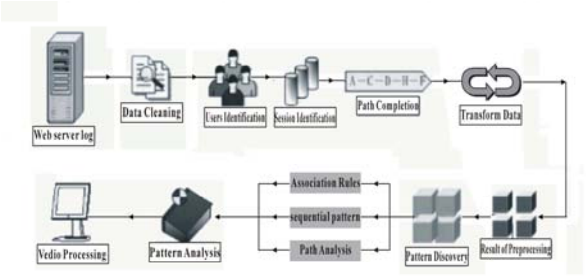


Fig.1. Steps of web mining

Data Cleaning:
There are data items which are has nothing to do with web mining in journal file, so these data items should be deleted in data cleaning. Generally speaking, HTML document is correlate to users' session in the journal file. When user accesses and downloads the web page, the pictures on the page is also downloaded, then, it is recorded in the log, but most of pictures are not asked for by the user initiatively, therefore, the user's behavior could be not judged according to the data items so the irrelevant items are deleted by judging extended name. For example, the graphic file whose extended names are gif, jpeg, jpg as well as the script document whose extended name is cgi. If only the web site contains the graph documents, the gif and jpeg files in the log should not be deleted. Because they are requested by users possibly. Also, web pages which are illegal request should be cleaned up.

Users Identification:
The work of the web usage mining system is to discover the user favorite accessing pattern, so how many users are accessing what information should be knew firstly. As a result of the existence of proxy server and firewall and users dynamic get IP address, the work of user recognition becomes very complex. The commonly solution includes:

(1) Different IP represents the different user;
(2) When IP is the same ,the judgment is according to user's operating system and browser;
(3) Uniting the access log and web site topology, complete user's browsing path. If there is no link between current request page and the requested web pages, we could think there are more than one user.
(4) Judge the user by Cookie, when the user request to web resource, the web server would produce only Cookie.

Session Identification:
It refers to a complete process of a user's request of web page. In the web server log whose time span is big, the user may access the web site more than one time. Session identification refers to identify every session, the simplest solution is establish a time value, when the request surpasses this time, a new session just be taken.

Path Completion:
The path completion's goal lies in completing web site log which there is no record to the user's request, then, getting the integrated pattern of accessing, only doing this, the user's significant access path can be distinguished correctly. If the client server uses buffer, it would be very difficult to identify the sequence of accessing. Under this situation, the server log would lose the accessing web page, but the log could be completed by forecast the lost web pages. For example, the user accesses the web page 1, then, he accesses the web page 3, however, there is no link between the web page 1 and web page 3, that there is one web page is lost in the path at least could be concluded.

Transform Data:
This is the last step of preprocessing before data mining. Formatting the related data according to the duty of mining, and make it meet the mining's need.

2.2.2 Pattern Discovery

Pattern discovery is mining effective, novel, latent, useful and finally understood by mining algorithm. Many methods and algorithms are used in pattern discovery, adopting which method is decided by the type and scale of question. Not every data mining method is suitable to knowledge discovery in electronic commerce, the related methods are association rules, sequential pattern, path analysis and so on. For instance, the browsing pattern could be discovered. The browsing pattern is a series of web pages which was accessed in a session. Also other patterns can be discovered through the web usage mining. For example, all web pages in a session are inspected by association rules, but their orders are not cared about. Association rules are used to discover which web pages are accessed in one session. The sequential pattern is used to discover the order in a period of time, namely the successively relations of data.

2.2.3 Pattern Analysis

The pattern was discovered once, it can be analyzed and utilized usually, and the pattern which the user is not interested in is deleted. For example, make a comparison to the customers' and the non-customers' browsing pattern of electronic commerce web site, the visitors of the web site are divided into the short-term visitors, the surveyor and the customers. The short-term visitors are filtrated through data preprocessing, then, the web log are divided into the customers' and the non-customers'. The pattern is discovered by analyzing both of web logs according to the specific request, finally, a comparison to the patterns are made, at last, the final result are intuitionist presented by video processing.

3. APPLICATION AND ALGORITHM

A great promotion is made in web site designing through web server log mining, such as decision of business, performance of system and so on. A example about how to make well use of the result of mining is given.

Suppose a electronic commerce web site's function structure is: homepage-> classification of goods-> detailed information of commodity-> leaving word and commentary -> transaction. The web server log should be preprocessed according to the step of data mining firstly, then, the pattern to the preprocessed object should be discovered. As result of pattern discovering, a web page database about the users' accessing is got and the sequence is consistent with the accessed web page. Pattern analysis means solving the practical problem by mining algorithm. For the convenience of operation, each web page is abstracted as node, the link among web pages is abstracted as line, then, what and how users have accessed can be illustrated by topology. From the topology, the accessed web pages' relativity and the large frequency of web pages are discovered. That means the most interested or the least commodity can be found.

As a result of the security, the addresses of web pages are marked by only web page. In order to save the space of storage, the pages are figured in number, then, the digital arrangement of many ways may be obtained. The associated numbers could be discovered by the association rules and sequential pattern, that is associated commodity, so it is easy to get the frequent access paths and the large one finally. The topology of web site could be designed or market stratagem is adjusted and what is desired is given according to the paths. For example, if the web pages of electronic commodities are accessed frequently, the merchant should enlarge the investment and increase the types and the advertisement of the electronic commodity. Also if some pages of commodities are accessed very few, those mean customers have little interest in these commodities. By association rules and sequential pattern, we discover the associated commodities, then, make market stratagem. If not only the association is very low, but also the frequency of accessing is low, the selling of this commodity may be stopped. The algorithm as following [4]:

The whole electronic commerce web site is regarded as completely graph. The homepage of the web site is marked as 1,the web page of the classification is 2,…,supposed there are n pages, the pages are marked as 1,2,3,…,n. Then, the path of accessing is expressed as (a, b, c, d, e),that means the order is a->b->c->d->e. In order to express many paths of users, the data is stored in adjacent matrix. Through adjacent matrix, it is easy to judge if there is link between the arbitrary web pages by users. If there is link from vi to vj by users, add one to the degree of A[i][j].

```
For ( int i=1;i<=n;i++)
For ( int j=1;j<=n;j++)
ID(A[j])= ID(A[j])+A[i,j];
OD(A[i])= OD(A[i])+A[i,j];
TD(A[i])= ID(A[j])+ OD(A[i]);
K=Max(TD(A[i]));     // discover the most frequent accessed web pages;
cout<<k;          //output the most frequent accessed web pages;
cout<<A[i,j];         //output the most frequent link by users;
```

The optimization path also could be discovered. There are may be more than one large frequency link by users, they are recorded as set w. Then:

```
W=[s];
For(int K=1;K<=n;K++)
If K not in w then
L[K]=A[s,K];p[K]=s;
For(int i=1;i<=n;i++)
For(int j=1;j<=n;j++)
Wm=0;
If K not in W and l[K]>wm then
m=k;wm=l[K];
W=w+[m];
If length(w)=n then
for(int i=1;i<=n;i++)
cout<<l[i];
if p(v)=z,p(z)=y,．．． p(u)=s,then output :p(s,v)=su．．． yzv
Else
For j=1 to n do
If m!=j then
If l[j]<a[m,j];p[j]=m;
If l[j]=a(m,j) then p[j]=m;
```

4. CONCLUSIONS

After the most frequency accessed web pages is got according to above algorithm, the least one also could be discovered. The discovered path is the sequence of accessing usually. The designing of web site and convenient for customers could be not only improved, the psychology and habit of shopping but also analyzed, then the electronic business effectively must be developed. For example, the hyperlink should be increased or deleted according the situation, and advertisement could be added or other market stratagems. These cause the B2C form of electronic commerce no longer stay at the surface, and develop the electronic commerce which is suit to the Chinese habit of shopping.

REFERENCES

[1] Zhang Yin, *The Research and Application of Data Mining[D]*, Bei Jing: Chinese Academy of Sciences, 1999.

[2] Gordon S.Linoff, Michael J.a.Berry, *Ming the web*, John Wiley &Sons, Inc, United States of America 2001.

[3] Margaret H.Dunham, *Data Mining[M]*, 2006 .

[4] Wang Chunxia Wang Yiran, “Mining Web Logs to Improve Website Organization[J],” *Microcomputer Information 2006*, 22(11-3), P218-220.

[5] Junjie Chen and Wei Liu, “Research for Web Usage Mining Model,” *International Conference on Computational Intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents*, Web Technologies and Internet Commerce [C], 2006 IEEE.

Lin Chen (1983-) is a master of School of Computer Science and Technology, Wuhan University of Technology, graduated from the Anqing University in 2005 with specialty of Computer Science and Technology, research interests are in data mining and electronic commerce

MCRM：Mining Classification Rules by Multiple Supports *

Rong Gu, Chunhua Ju
Computer Science Department
Zhejiang Gongshang University, Hangzhou, Zhejiang, China
Email: grtm117@126.com

ABSTRACT

The paper presents a multi-support algorithm MCRM of classification rule mining according to the huge business databases and the unevenly- distributing classification patterns. The algorithm integrates breadth- first strategies with depth-first ones for solving a memory-lacking problem and applies multi-support thresholds to settle the uneven distribution problem. The paper also introduces the method of expression and realization for the frequent classification item-set tree FCIST and the array-based threaded transaction forest ATTF.

Keywords: FCIST, ATTF, Classification Rule.

1. INTRODUCTION

The association rule mining is an important research topic of data mining, especially the multi-support association rule mining. The current researches on classification rules mainly focus on the mining of the unevenly distributed and huge databases.

The existing algorithms for classification rules mining support either multi-support thresholds or scalability of store memory. These algorithms mainly adopt the strategies of breadth-first, depth first, or both of them. However the breath first algorithms like Apriori [1] [2] [3] which are inefficient for dense datasets that contain long patterns.

The depth first algorithms such as FP-Growth [4] [5], H-Mine [6] and etc. do not scale to large sparse datasets and is time-wasteful in mining dense datasets. Some algorithms like OP [7] [8] putting breath first and depth first strategies together but using single support are not good for mining unevenly-distributed business databases. Although algorithms like CRM-PP [9] adopt multi-support method, they are not scalable to store space with the limitation of memory.

This paper presents a novel algorithm, multi-support algorithm MCRM, for mining classification rules in business databases, which is efficient on dense databases at all levels of support threshold, and scalable to very large databases.

Our contributions are as follows: First, we present frequent classification item-set tree for constructing frequent pattern sets, and adopt both breath first and depth first strategies, which enhances the build of frequent pattern tree. Second, we propose an array-based format for classification tree for pseudo projection with high efficiency and low memory cost, and adopt multi-support thresholds to mine more potential and effective association rules. Finally, we design and implement a multi-support algorithm MCRM of classification rules mining.

With the compare experiments with Apriori, FP-Growth and OP, MCRM has certain advantages in mining efficiency and scalability of the large business databases.

2. CONSTRUCTION OF FREQUENT CLASSIFICATION ITEM-SET TREE

2.1 Problem Descriptions

Definition 1 The classification database D=(O,I,C), in which O is the finite set of data objects, I is the set of data attributes, and C is the set of classification items.

The data attribute set also known as an item set, I={i1,i2,⋯,im}, in which ik is called data attribute. The data object set O={(tid1，t1)，⋯(tidn，tn)}, in which（tidk，tk）presents a data object or a transaction, and tidk is the identifier of the object or transaction, tk is the attribute set of classification objects, and tk I∪C, ‖tk ∩C ‖= 1. The classification item set C={c1，⋯，cs}, in which ck stands for a classification item.

According to the definition above, the database in table 1 can be described as:
I ={a,b,c,d,e,f,g,h,j,k,l,m,p,s};
C={ c1, c2};
O={(01,t1),(02,t2)⋯,(08,t8)} composed by 8 transactions, in which t_1= I∪C ={a,b,c,d,e,f,h,p,s,c_1}, t_2={a,c,e,f,g,j,l,p,c_2}.

Table 1. The Classification Database

Tid	Items	Class
01	a b c d e f h p s	c_1
02	a c e f g j l p	c_2
03	a b e j p	c_1
04	b e f m p	c_2
05	a e f m p s	c_2
06	a k l	c_1
07	a d e g f j p	c_2
08	b e k s	c_2

Definition 2 (Classification) pattern p⊆I is frequent if support(p) ⩾ minsupk, minsupk is the minimum support(threshold) of class ck.

Definition 3 If support (X ∪{ck} ,T) ⩾minsupk , and support (X ∪{ck} ,T) / support (X,T) ⩾minconf, X< I ,ck ∈ C , then X→{ck} is a classification rule.

2.2 Frequent Classification Item-set Tree

Frequent classification item sets can be represented by a tree, namely frequent classification item-set tree, abbreviated as FCIST, and in order to avoid repetitiveness, we impose an ordering on the items.

FCIST is an ordered tree, where each node is labeled by an item, and associated with a weight. The ordering of items labeling the nodes along any path (top down) and the ordering of items labeling children of any node (left to right) follow the

* Supported by National Natural Science Fund of China [70671094]

imposed ordering. Each frequent classification item set is represented by one and only path starting from the root and the weight of the ending node is the support of the item set. The null root corresponds to the empty item set. The weights associated with nodes need not be actually implemented.

Each node has its own classification projected transaction set (abbreviated as CPTS). CPTS consists of transactions that support the item set represented by the path starting from the root to the node. CPTS of the null root is the original database. CPTS of any node other than the null root is obtained by projecting transactions in CPTS of its parent node, according to the priori property. One CPTS is filtered if each transaction in the CPTS only maintains items that contribute to the further construction of descendants. In other words, filtered CPTS of a node only contains items that label the sibling of its parent node. Otherwise, the CPTS is unfiltered. Apparently, items in filtered CPTS are local frequent in its parent CPTS.

Each FCIST node is represented by[i,w1,⋯,ws]followed by its own CPTS, ws is the weight of item i belonged to the classification cs. Let π be the dictionary order, then the classification database in Table 1 can be represented as Fig.1 with the support thresholds minsup1＝2 and minsup2＝3. The path[,,]—[p,2,4]—[e,2,4]—[a,2,3] represents the classification item sets{a,e,p,c1} with support of 2 and{a,e,p,c2} with support of 3. The unfiltered CPTS of the root, namely priori classification database has 8 transactions, in which transactions 01, 02, 03, 04, 05, 07 support item p, and transactions 01 and 03 belong to classification c1, transactions 02, 04, 05 and 07 belong to classification c2. So the CPTS of [p,2,4]is composed by such 6 transactions. The CPTS of all the nodes except the root is filtered, only containing the brother nodes ahead.

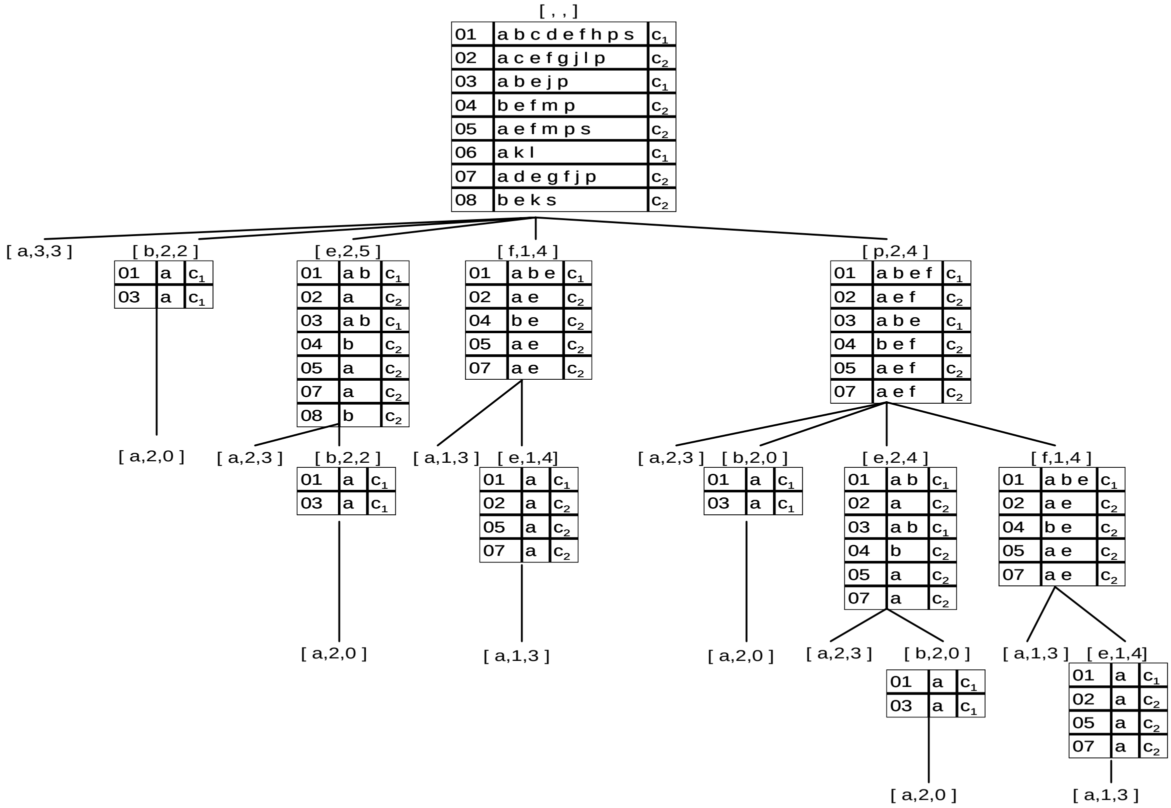


Fig.1. The FCIST in the example

2.3 Representing CPTS by ATTF and Pseudo-projecting

An array-based threaded transaction forest, ATTF, is adopted to represent CPTS because of its low memory spending. And two different pseudo-projecting methods are used to construct the FCIST. ATTF consists of two parts: an item list (IL), and a forest. Each local item in CPTS has an entry in the IL, with three fields: an item-id, multiple supports, and a pointer, namely e.item, e.conut and e.link respectively. And the multiple supports e.count can be divided into ||C|| different parts, called e.count (k) representing the support in classification ck. Entries in IL are ordered by the imposed ordering. Each transaction in the CPTS is represented by one and only one path in the forest. Each node in the forest is labeled by an array [i,w1,⋯,ws] where i is an item and wk is a count that is the number of transactions in classification ck represented by the path starting from the root ending at the node. Items labeling nodes along any path are sorted by the same ordering as IL. All nodes labeled by the same item are threaded by the entry in IL with the same item. ATTF is filtered if only local frequent items appear in ATTF, otherwise unfiltered.

For example, the filtered ATTF representation for the CPTS of the null root in Fig.1 is shown in Fig.2, where the path [a,3,3]-[b,2,0]-[e,2,0]-[p,2,0] represents transaction 01 and 03, [b,0,2]-[e,0,2]-[f,0,1]-[p,0,1] represents transaction 04, and so on. The third item in FIL is item e, with multi supports 2 and 5 in classifications c1 and c2 respectively. The pointer (arrow-headed broken line) links the nodes [e,2,0],[e,0,3] and [e,0,2] together.

From the ATTF representation of the CPTS of a parent node in FCIST, we can project its children's ATTFs either in a bottom up way or in a top down way. The FCIST in Fig.1 is constructed in a top down projecting way, so we will introduce this way in the paper.

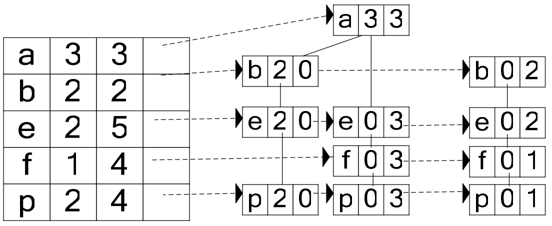


Fig.2. Representing CPTS by ATTF

In the top down way, the pseudo ATTF of a child CPTS consists of sub forest whose leaves are threaded together in its parent ATTF. Firstly, we choose the IL items one by one from the parent ATTF in the imposed ordering. Secondly, by traversing the sub forest threaded by the chosen IL, we can delimitate the CPTS by re-threading nodes in the sub forest, count the support of each item in the sub forest by re-calculating the count of each node according to the leaves' multi supports. For example, in Fig.3, the sub forest whose leaves, [f,0,3] and [f,0,1] are threaded by the entry of item f, compresses transactions that support item f. By traversing this sub forest, we get local multi supports of item a, b and e of [0,3],[0,1] and [0,4] respectively, and the count of first node label by a is changed from [3,3] to [0,3], and node [b,0,2] is adjusted into [b,0,1], node [e,0,2] into [e,0,1]. Therefore, the sub forest of the pseudo ATTF consists of two paths, [a,0,3]-[e,0,3] and [b,0,1]-[e,0,1]. The IL is {([a,0,3],ptr), ([b,0,1],ptr), ([e,0,4],ptr)}. This is a recursion process, e.g. the child ATTF of the node [e,0,4] in Fig.3can be pseudo projected as shown in Fig.4.

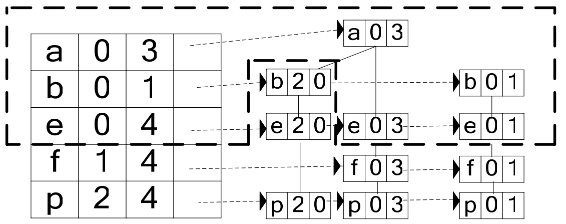


Fig.3. Top down pseudo projecing ATTF

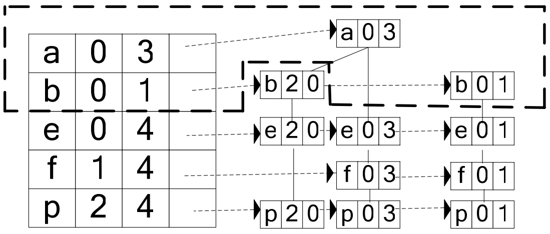


Fig.4. Child's ATTF ▯ top down▯

Such method of pseudo projection avoids recursively building projected transaction set, which is in the same number as frequent item sets. This method is not only space efficient in that no additional space is needed for any child ATTF, but the counting and projecting operation is also highly CPU-efficient. Especially we adopt different supports for classification items which makes the FCIST construction more efficient.

3. ALGORITHM MCRM

Now we present the multi-support algorithm of classification rules mining, abbreviated as MCRM, which integrates depth first and breadth first strategies, array-based threaded tree forest representation and filtered projection. First, MCRM creates a null node for the root of the FCIST, whose CPTS is the priori classification database. Second, MCRM calls BreadthFirst to grow the upper portion of FCIST by breadth first search until the reduced set of transactions can be held in a memory based structure. Third, DepthFirst is called to build the lower portion of FCIST by depth first search.

```
MSC（O, π,Minsup）
 create FCIST root R and let R.PTS=O;
  BreadthFirst (R,L, π,Minsup);
DepthFirst (T, π,Minsup);
```

BreadthFirst attaches counting vectors to all nodes at the current level L to accumulate local supports for items in the CPTS of each node. The counting vector has an element for the item of each sibling node that is before the node attached according to the imposed ordering π. We project the transaction t along the path from the root to nodes at the current level L and accumulate counting vectors. If a transaction can be projected to a level L node and contribute to its counting vector, it may also be projected to level L+1, therefore record it in D'. Otherwise it can be removed from further consideration. Then we create children for each node at the current level L for its local frequent items whose element in the counting vector has a value over the multi-support thresholds. The BreadthFisrt is a recursive procedure. We use the available free memory as parameter to control breadth first search process.

```
BreadthFirst(R,L, π,Minsup,D)
  for each node v at level L top down by π do
    CreateCountingVector(v);
    D'={};
    for each transaction t in D do
      ProjectAndCount(t,L,D')
    for each node v at level L do
      GenerateChildren(v);
If(NoMem(D')) then BreadthFirst(T,L+1, π, Minsup) ;
    else return(D');
```

If bread first projecting ends at level L, then DepthFirst is called to build the sub trees with roots of the leaves in level L.

```
DepthFirst (v, π,Minsup)
  for each node e in v.PTS.IL top down by π do
    if e.count(k)≥minsupk for some class ck in C then
create a child node d for v;
d.item= e.item;
d.weight(k)= e.cuont(k);
PseudoProj(v,d, π);
DepthFirst (d, π,Minsup);
```

In DepthFirst, first, the IL items are chosen by the imposed ordering; second, if the classification items of the IL item are frequent then creates the corresponding children nodes; third, calls the PseudoProj process. The PseudoProj process will adjust the weights and re-threading the sub trees to guarantee the children CPTS be contained in the parent CPTS.

4. PERFORMANCE EVALUATIONS

To evaluate the efficiency and effectiveness of our algorithm

MCRM, we have done experiments on the dataset Forest from UCI machine learning data warehousing by comparing with Apriori and FP-Growth on a 286MHz Pentium III PC with 512 MB main memory and 30 GB hard drive, running on Windows 2000 Professional.

The empirical results indicate that MCRM is one to three orders of magnitude more efficient than Apriori and FP-Growth. For example, in Fig.5, when the support thresholds are lower than 0.1%, MCRM is 2 to 12 more efficient than Apriori and 1.5 to 8 than FP-Growth. At the reasonable low support threshold of 0.05%, MCRM requires 16 seconds, whereas FP-Growth requires 31 seconds and Apriori requires 65 seconds. At the even lower support threshold of 0.02 %, MCRM requires 20, while FP-Growth requires 72 seconds and Apriori requires 155 seconds. The rankings of algorithms are MCRM＞FP-Growth＞ Apriori.

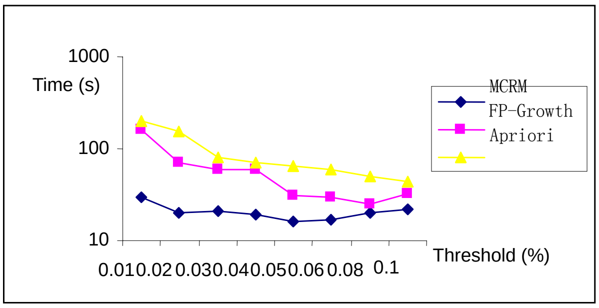


Fig.5. Performance comparison of frequent item set mining

5. CONCLUSIONS AND FUTURE WORK

The paper introduced a multi-support algorithm MCRM of classification rules mining, which adopts array-based threaded transaction forest method and pseudo projection to highly improve the efficiency of classification rules mining. The future work will be focus on the association rules mining in the business data streams [10].

REFERENCES

[1] R Agrawal, R Srikant., “Fast algorithms for mining association rules”, The 1994 VLDB, Santiago, Chile, 1994, pp. 487-499.

[2] S Brin , R Motwani , J Ullman et al, “Dynamic item set counting and implication rules for market basket analysis”, The 1997 ACMSIGMOD, Tucson , AZ , May 1997, pp. 255-264.

[3] Lu Jie, Zhang Zhijing, “An Improved Apriori Algorithm for Mining Association Rules”, *Microelectronics & Computer,* Beijing, China, volume 23, No.2, 2006, pp. 10-12.

[4] J Han, J Pei , Y Yin, “Mining frequent patterns without candidate generation”, The 2000 ACM2SIGMOD, Dallas, TX, 2000, pp. 1-12.

[5] Song Yuqing, Zhu Yuquan, Sun Zhihui, Chen Geng, “An algorithm and its updating algorithm based on FP-Tree for mining maximum frequent itemsets”, *Journal of Software*, China, volume 14, No.9, 2003, pp.1586-1592.

[6] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang,” H-Mine:Hyper-Structure Mining of Frequent Patterns in Large Databases”, Proc. 2001 Int. Conf. on Data Mining (ICDM'01)}, San Jose, CA, Nov. 2001, pp.441-448.

[7] Liu JQ, Pan YH, Wang K, Han J. Mining frequent item sets by opportunistic projection. In: Hand D, et al, eds. Proc. of the 8th ACM SIGKDD Int’l. Conf. on Knowledge Discovery and Data Mining. Alberta: ACM Press, 2002, pp.229-238.

[8] Liu Junqiang, Pan Yunhe, “An efficient algorithm for mining closed itemsets”, *Journal of Zhejiang University Science,* Zhejiang, China, Jan 2004, pp. 8-15.

[9] Liu Junqiang, Sun Xiaoying, Wang Xun, “Pseudo projection algorithm for mining of classification rules”, *Computer application & Software,* China, Sept 2003, pp. 8-10.

[10] Geoff Hulten, Laurie Spencer, Pendro Domingos, “Mining time-changing data streams”, *In Proc.ACM Int.Conf. on Knowledge Discovery and Data Mining*[C], San Francisco, California, 2001,pp.71-80.

Mining Frequent Patterns from Xml Data Based on Vertical Data

Shangping Dai, Xiangming Xie, Tian He
Computer science and technology, centeral china normal university
Wuhan, Hubei, China
Email: xqm621@163.com

ABSTRACT

Association rule mining is one of the most frequently discussed topics, and among which, frequent patterns are the most interesting. But now, nearly all the algorithm and relevant work are based on horizon data. We present a method based on vertical data and algorithm FD. It is proved both in theory and practice that this method is adoptable and effective than traditional one based on horizon data.

Keywords: XML document, Frequent patterns, DOM, Vertical data.

1. INTRODUCTION

Recently, XML is widely used as the de facto standard for data exchanging in internet. As more and more data is stored and represented in XML format, there have been increasingly more research efforts in mining XML data. Existing works on mining XML data include frequent substructure mining [5, 6], classification [7], association rule mining [1, 2, 3, 4], etc. Among that, the association rule mining is one of the most frequently discussed topics. The set of discovered frequent patterns can be useful in different XML-based applications. All the mining work is conducted in following method: first mapping xml data into relational database, then mining in relational database with existed algorithm. Now, nearly all the exchanged data in relational database is known as in the format of horizontal data .Alternatively, according to the characteristic of xml document, data can also be easily exchanged to vertical data format.

2. MAIN CONTRIBUTIONS

Many papers are about the improvement of algorithm on mining xml document, such as FP-growth. That is to say, nearly all the papers are based on horizontal data format. But now, vertical data format is also used widely, and frequent itemsets can also be mined efficiently in vertical data format. Can the mining problem be worked out from the first step? This paper will present the method of how xml data be exchanged to vertical data format and mining on it .Experiment shows that it is quite efficient and practical.

3. PRELIMINARIES

3.1 Vertical Data Format

In most cases, both Apriori and FP-growth algorithm mine frequent patterns from a set of transactions in TID-itemset format ({TID: itemset}), where TID is a transaction-id and itemset is the set of items bought in transaction TID. The above data structure can be like table 1:

Table 1. Horizontal data format

TID	List of item-IDs
T100	I1,I2,I5
T200	I2,I3,I4
T300	I2,I3

Alternatively, data can also be presented in item-TID-set format (that is, {item: TID-set}), where the item is an item name, and TID-set is the set of transaction identifiers containing the item. The above data structure can be like table 2:

Table 2. Vertical data format

itemset	TID-set
I1	T100
I2	T100,T200,T300
I3	T200,T300
I4	T200
I5	T100

3.2 Mining frequent patterns

In the work of data mining, especially association rules, the main problem can be reduced to mine frequent itemsets. So the way to discover the frequent itemsets is very important. In vertical data, mining can be performed by intersecting the TID-sets of every pair of frequent single items. Suppose the minimum support is 2, and because every single item is frequent in table 2,there are 3 intersections performed in total ,which lead to 6 nonempty 2-itemsets,2-itemsets,{I1,I2},{I1,I5},{I2.I3},{I2,I4},{I2,I5},{I3,I4}, except the itemset {I2,I3}, each contain only one transaction, thus they do not belong to the set of frequent 2-itemsets. Based on the Apriori property, a given k+1 itemset is candidate k+1 itemset only if each of its k-itemset subsets is frequent, so it is maximal frequent itemset.

4. MINING FREQUENT PATTERNS IN XML DOCUMENT

4.1 Mapping XML Data to Vertical Data

Methods [8, 9] have been invented to map xml data to relational data base, but nearly all are in the form of horizontal data. How can it be mapped to vertical database? First in mind, the mapped horizontal data can be exchanged to vertical data, but it is expensive, it needs to scan database for as many times as the number of itemsets. In xml document, itemset is leaf node, so to find it is easy. We could use java DOM to parse the xml document. The key problem is how to define TID. After analyzing massive xml DTD, it is found that to define the first element or element sets to the root element which is under the tag "+" as TID is most suitable, and it is thus called "super tag", because fragment is useful information only if it appears frequently. That is to say, when scanning, if the tag is "super tag", the TID adds one. Here is an example:

Information (goods, customer)$^{+}$
Goods (appliance, clothes)
Appliance (items)
Clothes (items)

Items (item) $^{+}$

It is best to define the goods, customer element sets as TID. Because the xml fragment goods and customer appear together. So we can find the relation of appliance and clothes separately, and between them.

4.2 Mining Vertical Data

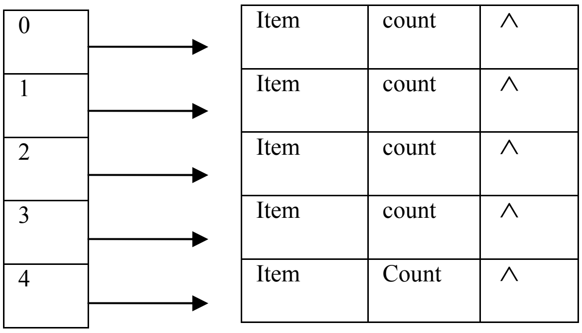
Referencing the Apriori algorithm, in general, association rule mining can be viewed as a two-seep process;

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min-sup.
2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

In the above two steps, because the second step is much less costly than the first, the overall performance of mining is determined by the first step. So how we can discovery the frequent itemsets is very important in the work.

Parsing xml document using java DOM, we could store the information in memory, using the method of hash table. In every tuple of hash table, it contains such information: item, count and TID. Count is the number of TID. and to store TID, we use hash chain pointer. It is shown in table 3.

Table 3. Hash array



0	Item	count	^
1	Item	count	^
2	Item	count	^
3	Item	count	^
4	Item	Count	^

When scanning xml document, if an item is not in hash table, then adds it or the count number adds one, and stores the TID. When the scanning step is over, all the work can be done in memory. It can be called FD algorithm.

Algorithm FD:

Input: xml document d and DTD D;

Output: frequent transaction T;

1. Scanning D, find out first element or element sets to the root element which is under the tag "+"

2. STR (d, Tp); //parse xml document using DOM, store every leaf node to hash table;

3. EXT (Tp, Th); //extract hash table Tp, put the result Th;

Step description:

STR (d, Tp)

① when "super tag" occurs, TID adds one;

② for each leaf node, get the content;

③ if the content has already existed in hash table, store the TID, count adds one, else create the item;

EXT (Tp, Th)

The specification of the algorithm is like Apriori algorithm, the main difference lies in the link of TID list

Algorithm of joining two candidates:

Input: two candidate *L* 2item sets X and Y, and their ctid-list X.tids and Y.tids;

Output: new candidate (*L* + 1) 2item set *C* and its tid2list;

```
Begin
   C=X [1] X [2]…X [L] Y [L];
   C.support= 0;
   I= 1;J = 1; K = 0;
   while I ≤|X| and L ≤|Y | do {
      if X. tids [I] == Y. tids [J] then {
         K+ + ; C.support ort ++;
         C.tids [K]= x.tids [I];
         }
      else if X.tids [I] < Y,tids [J] then I ++;
      else J + + ;
   };
End
```

5. PERFORMANCE STUDY

In this section, the performance of the whole miming process is evaluated, and this thesis also compares it with the Apriori algorithm in horizontal data, the adopted algorithm is also based on Apriori algorithm, our aim is to find out which data format is better for miming xml document, because nearly all the method is based on horizontal data.The algorithm is implemented in java, and the experiments are carried out on a 1.8GHZ with 256M RAM, running windows 2000. All the data is from database census which is commonly used (1.60M), with 14 attributes all together, first use "for xml" to change it to xml data, then miming from xml document. Fig 1 shows the result mining in two data format, Fd algorithm in vertical data format and Ariori algorithm in horizon data format.

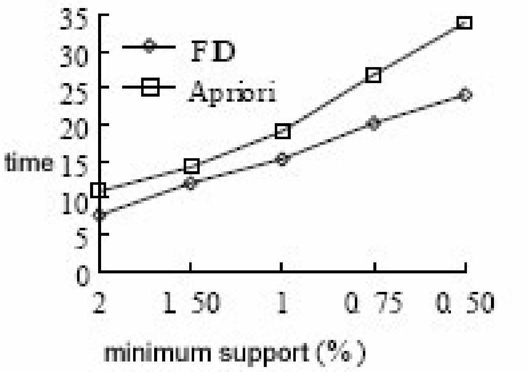


Fig.1. comparison between two data formats

6. CONCLUSIONS

From the result of fig 1, it is found out that FD algorithm is a little faster than the Apriori algorithm. All the work imitates the Apriori algorithm. It may be said that there is FP-growth algorithm, which is better than Apriori. But FP-growth algorithm is based on horizontal data, and is not suitable for vertical data .If better method is available for vertical data, the performance may be enhanced more greatly, so next work is to improve the algorithm for vertical data.

REFERENCES

[1] R.Agrawal. T.Imielinski. and A.Awami. “Minging association rules between sets of items in large databases” (*SIGMOD’93*. pages 207-216.ACM.May 1993

[2] D.Braga. A.Campi.S.Cen.M.Klemettinen *Discovering interesting information in xml data with association rules*.2002

[3] HAN Jiawei, KAMBERM. Data Mining: Concept and Techniques: Morgan Kaufmann Publishers Inc, 2006.

[4] H. Tan, T.S. Dillon, L. Feng, E. Chang, F. Hadzic. X3-Miner: mining patterns from XML database. International Conference on Data Mining. *Text Mining and their Business Applications*, 2005, pp. 287–298.

[5] A. Inokuchi, T. Washio, H. Motoda, “An apriori based algorithm for mining frequent substructures from graph data”, in: *Proceedingsof European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD,* 2000, pp. 13–23.

[6] M. Kuramochi, G. Karypis, “Frequent subgraph discovery”, *in:Proceedings of IEEE International Conference on Data Mining ICDM,*2001, pp. 313–320.

[7] M.J. Zaki, C.C. Aggarwal, XRULES: “an effective structural classifier for XML data”, in: *Proceedings of ACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining,* 2003.

[8] P. Bohannon, J. Freire, P. Roy, J. Simeon, “From XMLschema to relations”: *a cost-based approach to XML storage,ICDE,* 2002.

[9] M. Mani, D. Lee, “XML to relational conversion using theory of regular tree grammars”, *VLDB Workshop on EEXTT,* 2002.

Data Distribution Management based on Colored Petri Net and Lookahead

Xiangwei Liu, Xianwen Fang
Economy and Management Department, Anhui University of Science and Technology, Huainan 232001, China
Email:lxw7710@tom.com

ABSTRACT

For researching the high level architecture (HLA) in the data distribution management (DDM), the important object is enhancing the data filter efficiency and reducing computing quantity at the same time, providing well scalability for some kind of distribution simulation application under the HLA standard. Lookahead is an important time management concept in distributed simulation, in order to speed up the procedure, each simulation entity informs timestamp produced using Lookahead to other entities as far as possible early. Using the data distribution mechanism based on the colored Petri net and Lookahead, the data filters can be well carried out, the data filter efficiency can be enhanced, and the success rate of the data dispatching and receiving can be guaranteed. The simulation experiment result indicated that, the data distribution mechanism based on the color Petri net and Lookahead is better than the region match method.

Keywords: Colored Petri Net（CPN），Lookahead，Data Distribution Management (DDM)，High Level Architecture (HLA)

1. INTRODUCTION

HLA (High Level Architecture) provides an efficient way that solves the problem of simulation system scalability. HLA adopts customer server pattern, and data exchange among federation members by RTI (runtime infrastructure) server. According to the regulation of FOM (Federation Objection Management), the efficient information alternation and transfer mechanism among federation members is built. HLA provides data distribution management (DDM) service. DDM uses value-based data filter mechanism, and permits federation members to accept subscriber's attribute value and alternation information based on value or characteristic of publish members. DDM is a kind of important service provided by RTI, its essential target is to decrease the error data, which federation members received, and the net data flow quantity as little as possible. DDM is the key technique of realizing and is also the important means of realizing large-scale distribution simulation’s scalability. In the literature [3] presented a class-based DDM analysis method, this method only adapts to the small-scale federation members or small quantity entity federation. The literature [4] presented the region-based DDM analysis methods, whose main weakness is that every subscribe may receive a lot of unrelated data. The literature [6] gave out the grid method based on data filter mechanism, whose drawback is the waste of broadcast address. The essence of different data distribution management strategy is to choose different method which information is related. The essential difference also results in the discrepancy in the filter efficiency, the calculation and communication spending, the usability of broadcast address, and requirement for state information and so on.

The colored Petri net is a high-level Petri net, it has the characteristics including the basic Petri net and the high-level language. In the literature [7] , a study about DDM based on Petri net is presented, but matching time is not good sometimes.Lookahead is an important concept in the distributed simulation time management protocol; every simulation entity has time mark of the event oneself producing through lookahead to notify other entities earlier, in order to accelerate the operation of the procedure. In the HLA (the High Level Architecture) standard the explanation of lookahead function is:" guaranteeing all members of the union not to produce the event whose time mark is smaller than presently the union member of time plus the lookahead”. Using the lookahead, the matching time can be controlled, the characteristics of the colored Petri net and lookahead have large superiority to other methods to deal with data distribution management in the distributed simulation.

2. BASIC CONCEPTIONS

Definition 1[1,2]: A Colored Petri Net is a 9-tuple CPN＝（$\sum$,P,T,A,N,C,G,E,I）satisfying the following requirements:

(i) $\sum$ is a finite set of non-empty types, called color sets.

(ii) P is a finite set of places.

(iii) T is a finite set of transitions.

(iv) A is a finite set of arcs such that:

$$P \cap T = T \cap A = P \cap A = \Phi \, .$$

(v) N is a node function. It is defined from A into $P \times T \cup T \times P$.

(vi) C is a color function. It is defined from P into $\sum$.

(vii) G is a guard function. It is defined from T into expressions such that:

$\forall t \in T : [Type(G(t)) = Bool \wedge Type(Var(G(t))) \subseteq \Sigma]$

(viii) E is an arc expression function. It is defined from A into expressions such that:

$\forall a \in A : [Type(E(a)) = C(p(a))_{MS} \wedge Type(Var(E(a))) \subseteq \Sigma]$

where p(a) is the place of N(a).

(ix) I is an initialization function. It is defined from P into closed expressions such that:

$\forall p \in P : [Type(I(p)) = C(p)_{MS}]$

Definition 2[1]: A binding element is a pair (t,b) where t is a transition while b is a binding for the variables of t.

Definition 3[1]: A step is a multi-set of binding elements. A step Y is enabled in a marking M iff the following property is satisfied:

$$\forall p \in P : \sum_{(t,b)\in Y} E(p,t) < b > \le M(p) \, ^{\cdot}$$

When a step Y is enabled in a marking M_1 it may occur, changing the marking M_1 to another marking M_2, defined by: $\forall p \in P$:

$$M_2(p) = (M_1(p) - \sum_{(t,b)\in Y} E(p,t) < b > + \sum_{(t,b)\in Y} E(t,p) < b >$$

The first sum is called the removed tokens while the second is called the added tokens. Moreover we say that M_2 is directly reachable from M_1 by the occurrence of the step Y, which we also denote: $M_1[Y_1 > M_2$.

Definition 4[1]: An occurrence sequence is a sequence of markings and steps:

$M_1[Y_1 > M_2[Y_2 > M_3 \ldots M_n[Y_n > M_{n+1}$

such that $M_i[Y_i > M_{i+1}$ for all $i \in 1..n$. We then say that M_{n+1} is reachable from M_1. We use $[M >$ to denote the set of markings, which are reachable from M.

In different distributed simulation environments and different protocols, the concept meaning of lookahead is different subtly. In the conservative distributed simulation protocol, the event handling model of logical process adopts the way of the single queue single server, in this situation conveying the time stamp of event between logical process adopts three pieces of time amount (such as Fig.1): t_{cause} : the moment of sending events which entity knows at first, can be also regarded as source time mark of event; t_{commit} : the moment that the event can be committed and sent, the moment that the source event is finished; t_{effect} :as the effective moment of time mark of the sent event , or as time mark of the sent event.

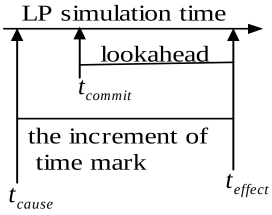


Fig.1. The *indication of lookahead*

So the definition of lookahead is adopted Min{$t_{effect} - t_{commit}$}, it means the minimum of the increment of time mark in the simulation course.

Definition 5[7]Provided t_{commit} means the moment that the event can be committed and sent, t_{effect} means the effective moment of time mark of the sent event, La= Min{$t_{effect} - t_{commit}$}, which means the minimum of the increment of time mark in the simulation course. Then call La as Lookahead.

3. THE DDM MECHANISM BASED ON COLORED PETRI NET AND LOOKAHEAD

In the HLA, the basic concept of supporting data filter is routing space. A routing space is a multi-dimension coordinates system, and is a data space, which is constituted by federation attribute -value. Members can use RS to express the range of data received and sent hopefully. The subclass of routing space called region, which can be divided into two regions:(1)update region. Update region is used to show that member's promises to send attribute-value into this region, and usually the attribute-value of the object in the region formed subset-space. When the attribute-value of the object changes dynamically along with the time, the update region forms a locus in the routing region at the same time. (2) subscribe region. Members use subscribe region to indicate that they want to receive some information in this region. Subscribe region is related with members, a member's subscribe region also dynamically changed along with time.

If an object's update region overlaps with a member's subscribe region, we call this member discovers the object. In the ideal state, RTI should guarantee to distribute the attribute-value to the members, which is discovered by the members.

In the data distribution management based on the colored Petri net and lookahead, once federation members' carry out subscribes behavior, RTI will send token that carries colored mark to publish where produced update data. Subscriber receives information that accorded to some conditions and satisfied the color requirement, at the same time computing the lookahead. As long as the publish has data update at any time, under the Petri net's effect of transition and arc, the token which has been loaded to sender will carry out data filtering, then select the better path having least lookahead and send to its subscriber. DDM process based on the colored Petri net and lookahead is shown in Fig.2.

In the DDM based on the colored Petri net and lookahead, if we want to build data publishing and subscribing relationship between publisher and subscriber. At first publisher and subscriber must describe the data that can be produced and consumed at the matching form. One publisher's matching forms as follows: (color information, message information); Subscriber's matching forms have (color information, message information). If and only if the color information between publisher and subscriber are the same, related transition can enable, and if this transition accorded with arc's banding condition, data's receiving-sending relationship can come true. The Receive Packet receive data from the Send place and affirm successful data, then really go into received region, at this time, data's receiving and sending can be come true. In the process of sending and receiving data, if data are not the same, DDM based on the colored Petri net and lookahead will resend the data again, and carry on data acknowledgement every time. If dada resend for many times (exceeded max-resend time ruled by system), then system will carry on next information sending automatically. Through the integrated process, DDM could be finished.

4. AN ANALYSIS ON SIMULATION RESULT

Compared with the region-matching method, DDM, based on the colored Petri net and lookahead(CPN & Lookahead), can efficiently realize data's matching and distributing. In the Lenovo M4600, making use of CPN/Tool simulation tool, adopting CPN & Lookahead method and region-matching method, we examine the relationship between different data-sent quantity and successful data-matching rate in the system, the result is showed in Fig.3. In the Fig.3, we can know that with the increase of data-sent quantity, DDM which adopting CPN & Lookahead method and region-match method, the successful data-matching rate will both decrease obviously. But concerning to the equal data-sent quantity at the same time, successful data-matching rate of CPN & Lookahead method is better than region-match method's. In the DDM mechanism based on the colored Petri net and lookahead, because of using data affirmation, data happening without order and resending time limitation mechanism, the successful rate of data-match is ensured, meanwhile, using the lookahead, then selecting the better path having least lookahead and sending message, the matching time can be reduced, and data resending is not unlimited, data-matching time is optimization. In the same flat based on the above, we have checked the matching time by adopting CPN &Lookahead method and region-matching method, and the relationship between data-sent quantity and data-matching time. In Fig.4, we can know that the increase of data-sent quantity, data matching time which adopting DDM of CPN &Lookahead method and region matching method increases obviously. But concerning to the equal data-sent quantity, data-match time of CPN &Lookahead method is fewer than region-matching method's, it is also better than CPN method presented in the literature[7].

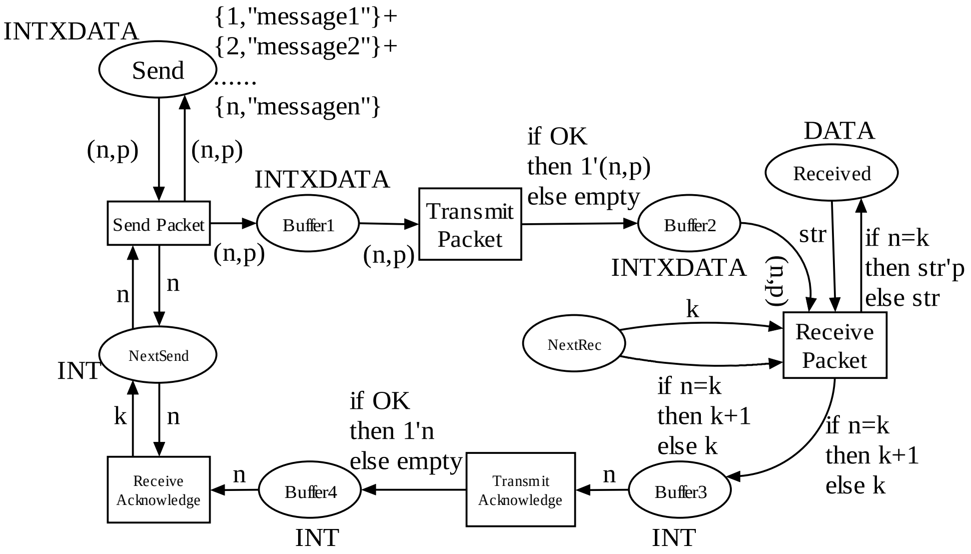


Fig.2. DDM process based on the colored Petri net and lookahead

About variable declaration in Fig.2:
Color INT= int with 1..10000;
Color DATA=string;
Color BOOL=Boolean;
*Color INTXDATA=product INT * DATA;*
Var n,k:INT; Var p, str: DATA; Var OK: BOOL

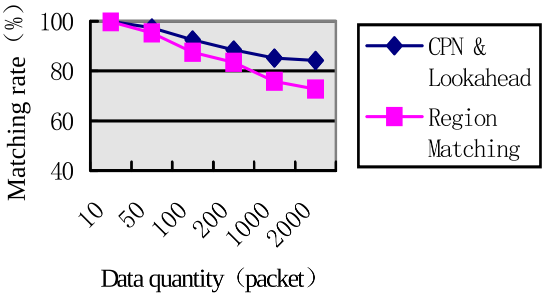


Fig.3. Two methods of the relationship between data-sent quantity and successful rate of data matching

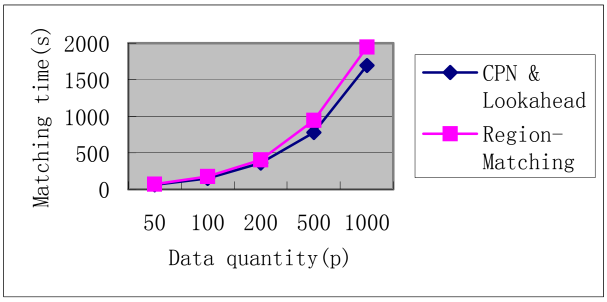


Fig.4. Two methods of the relationship between data-sent quantity and the time of data matching

5. CONCLUSIONS

This paper offered a data distribution management mechanism based on colored Petri net and lookahead, and using the colored Petri net to describe the process of data's sending and receiving. Making use of CPN/TOOL simulation tool, as to different data-sent quantity, adopting CPN & Lookahead method and region matching method, we check data-matching success rate and matching time in the system, then compare with simulation results to two kinds of methods. At last, the result shows that DDM based on colored Petri net and lookahead is superiority to region matching method.

REFERENCES

[1] Kurt Jensen. Colored Petri Nets. Springer -Verlag Berlin Heidelberg, 1996.
[2] Kurt Jensen. Colored Petri Nets: "A High-Level Language for System Design and Analysis." *Computer Science* 1991, 342-416.
[3] Mark Hyett, Roger Wuerfel. "Implementation of the Data Distribution Management Services" in the *RTI-NG Simulation Interoperability Workshop*, March 2002.
[4] Bachinsky S, Noseworthy R, Hodum F. "Implementation of the Next Generation RTI." *Simulation Interoperability Workshop*, March 1999.
[5] Petty M. "Geometric and Algorithmic Results Regarding HLA Data Distribution Management Matching." *Simulation Interoperability Workshop*, September 2000.
[6] Y.C.Zhang,G.J.Sun. "A new algorithm on DDM in distributed simulation," *Journal of system simulation,* 2005, 1:91-95.

[7] FANG Xianwen, ZHAO Yan, YIN Zhixiang. “The Study on Data Distribution Management Based on Petri Net,International Conference on Innovative Computing,” *Information and Control*, August 2006.

Xiangwei Liu was born in 1977 in Anhui province, China. She is a lecturer, she got Master degree in 2004.Her research interests are DDM and finance investment analysis.

Xianwen Fang was born in 1975 in Henan province, China. He is a Ph.D. student. His research interests are parallel computing and Petri nets.

Algorithms for Mining Association Rules in Image Databases

Li Gao, Shangping Dai, Changwu Zhu, Shijue Zheng
Department of Computer Science, Hua Zhong Normal University
Wuhan 430079, Hubei,China
Email: lgao@mail.ccnu.edu.cn

ABSTRACT

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as clustering, classification, and association rule mining. In this paper we take advantage of the genetic algorithm (GA) designed specifically for discovering association rules. We propose a novel spatial mining algorithm, called ARMNGA(Association Rules Mining in Novel Genetic Algorithm), Compared to the algorithm in Reference[2] , the ARMNGA algorithm avoids generating impossible candidates, and therefore is more efficient in terms of the execution time.

Keywords: Data Mining, Genetic Algorithm, Association Rules, Multimedia, Image Databases

1. INTRODUCTION

Because of the advances in information technology, vast numbers of images have accumulated on the Internet and in entertainment, education, and other multimedia applications. Therefore, how to mine interesting patterns from image databases has attracted more and more attention in recent years. Many data mining methods have been proposed such as association rule mining, sequential pattern mining, calling path pattern mining, text mining, temporal data mining, spatial data mining, etc [1] [2].

Association rule induction is a powerful method used to find regularities in data trends. By induction of the association rules, sets of data instances that frequently appear together must be founded. Such information is usually expressed in the form of rules. An association rule expresses an association between items or sets of items. However, only those association rules that are expressive and reliable are useful. The standard measures used to assess association rules are the support and the confidence of a rule. Both are computed from the support of certain item sets [3].

Genetic Algorithm (GA) is one self-adaptive optimization searching algorithm. GA obtains the best solution, or the most satisfactory solution through generations of chromosomes' constant evolution includes the reproduce, crossover and mutation etc. operation, until a certain termination condition is [4] [5].

Association rules mining Algorithm Based on a novel Genetic Algorithm (ARMNGA) is an optimal algorithm combing GA.
The contributions of this paper are:

- We take advantage of the genetic algorithm (GA) designed specifically for discovering association rules.
- We propose a novel spatial mining algorithm, called ARMNGA, Compared to the algorithm in [2], and the ARMNGA algorithm avoids generating impossible candidates, and therefore is more efficient in terms of the execution time.

2. ASSOCIATION RULES

Definition 1 confidence

Set up $I = \{i_1, i_2, i_m\}$ for items of collection, for item in $i_j(1 \le j \le m)$, $(1 \le j \le m)$ for lasting item, $D = \{T_1, T_N)$ it is a trade collection, $T_i \subseteq I(1 \le i \le N)$ here T is the trade.
Rule $X \rightarrow Y$ is probability that $X \bigcup Y$ concentrates on including in the trade.
The association rule here is an implication of the form $X \rightarrow Y$ where X is the conjunction of conditions, and Y is the type of classification. The rule $X \rightarrow Y$ has to satisfy specified minimum support and minimum confidence measure [6].
The support of Rule $X \rightarrow Y$ is the measure of frequency both X and Y in D

$$S(xy) = |xy| / |D| \qquad (1)$$

The confidence measure of Rule $X \rightarrow Y$ is for the premise that includes X in the bargain descend, in the meantime includes Y

$$C(x \rightarrow y) = S(xy) / S(x) \qquad (2)$$

Definition 2 Weighting support
Designated ones project to collect I = {*i*$_1$*, i2, i*$_m$*}*, each project i$_j$ is composed with the value w$_j$ of right *(0≤wj ≤ 1, 1≤j ≤m)*. If the rule is *X* →*Y*, the weighting support is

$$S_w(xy) = \frac{1}{k} \sum_{i \in xy} w_j S(xy) \qquad (3)$$

And, the *K* is the size of the Set *XY* of the project. When the right value *Wj* is the same as *i*$_j$, we calculating the weighting including rule to have the same support.

3 GENETIC ALGORITHMS (GA)

Genetic Algorithm (GA) is a self-adaptive optimization searching algorithm. GA obtains the best solution, or the most satisfactory solution through generations of chromosomes constant evolution includes reproduction, crossover and mutation etc.

Here is the general description of this problem:

$$F(x) = a \times S(x) + b \times C(x) \qquad (4)$$

a, b is constants, a *≥0, b≥0, S(x)* is the support, and *C(x)* is the confidence.

4. ASSOCIATION RULES MINING BASED ON A NOVEL GENETIC ALGORITHM

4.1 Encoding

This paper employs natural numbers to encode the variable A_{ij}. That is, the number of the lines of every range in the matrix A in which the element 1 exists is regarded as a gene. The genes are independent of each other. They are marked by $A_1, A_2 \ldots Aj, An$, in which $A_j \in [1,m], j \in [1,n]$ and A_n may be a repeatedly equal natural number.

When the distributive method at random is employed to produce the initial population comprised of certain individuals, the population must be in a certain scale in order to achieve the optimal solution on the whole. The best way is the generated *M* individuals randomly that the length is n，then the chromosome bunch encoded by the natural number is calculated as the initial population.

4.2 The Fitness

Formula (3) is properly transformed into:

$$F(xy) = W_s \times \frac{S(xy)}{S_{\min}} + W_c \times \frac{C(xy)}{C_{\min}} \qquad (5)$$

Here, $W_C+W_s=1$, $W_c \geq 0$, $Ws \geq 0$, S_{min}, is minimum support, and C_{min} is minimum confidence.

4.3 Reproduction Operator

Reproduction is the transmission of personal information from the father generation to the son generation. Each individual in each generation determines the probability that it can reproduce the next generation according to how big or small the fitness value is. Through reproducing, the number of excellent individuals in the population increases constantly, and the whole process of evolution head for the optimal direction. We are adopting roulette selection strategy; each individual reproduction probability is proportion to fitness value.

1) Compute the reproduction probability of all the individuals

$$P(i) = \frac{f(i)}{\sum_{i=1}^{M} f(i)} \qquad (6)$$

2) Generate a number r randomly, r=random [0, 1] ;

3) If $P(0)+ P(1)+\ldots+ P(i-1)<r< P(0)+ P(1)+\ldots+ P(i)$,the individual i is selected into the next generation.

4.4 Crossover Operator

Crossover is the substitution between two individuals of the father generation that is to generating new individual .The crossover probability Pc directly influences the convergence of the algorithm. The larger *Pc* is the most likely is the genetic mode of the optimal individual to be destroyed .However, the over-small of *Pc* can slow down the research process [7] .Here is the definition of the crossover operator:

Computing crossover probability Pc

$$Pc=\begin{cases} p_{c1}-\dfrac{(p_{c1}-p_{c2})(f(x)-\overline{f(x)})}{f_{\max}(x)-\overline{f(x)}} & f(x) \geq \overline{f(x)} \\ p_{c1} & f(x) \prec \overline{f(x)} \end{cases} \qquad (7)$$

Here, p_{c1}=0.9, p_{c2}=0.6, $f_{max}(X)$ is the maximum fitness value of the population, $\overline{f(X)}$ is the average fitness value of the population.

4.5 Mutation Operator

The role of the mutation operator lies in that it enables the whole population to maintain a certain variety through the abrupt change of the mutation operator when a local convergence occurs in the population. The selection of the mutation probability P_m is the vital point because it influences the action and performance of the ARMNGA. If P_m is over-small, the ARMNGA will become a pure random research .Here is the definition of the mutations operator, computing the mutation probability P_m

$$Pm=\begin{cases} p_{m1}-\dfrac{(p_{m1}-p_{m2})(f(x)-\overline{f(x)})}{f_{\max}(x)-\overline{f(x)}} & f(x) \geq \overline{f(x)} \\ p_{m1} & f(x) \prec \overline{f(x)} \end{cases} \qquad (8)$$

Here, p_{m1}=0.1, p_{m2}=0.001, $f_{max}(X)$ is the maximum fitness value of the population, $\overline{f(X)}$ is the average fitness value of the population.

4.6 Termination Condition

When the matching error ε≈0 or the condition is not coincident, the process will naturally stop.

5 EXPERIMENTS ON SYNTHETIC DATA

To check the research capability of the operator and its operational efficiency, such a simulation result is given compared with the GA in [2] ,The platform of the simulation experiment is a Dell power Edge2600 server (double Intel Xeon 1.8GHz CPU,1G memory , RedHat Linux 9.0).

We first compare the performance of our proposed method with the algorithm in [2].Fig.1 shows the runtime vs. the minimum support for both algorithms, where the minimum support varies from 0.25% to 2% for the synthetic dataset. Our proposed algorithm runs 2–5 times faster than the Apriori algorithm, because a large number of candidates can be pruned by using the ARMNGA pruning strategy. Therefore, as the minimum support threshold decreases, the runtime of the Apriori algorithm increases dramatically since it generates too many candidates when the minimum support is small.

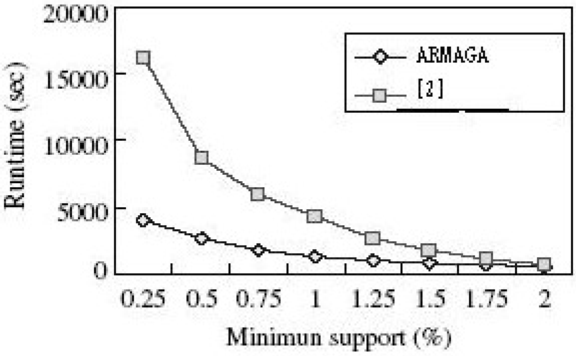


Fig.1. Runtime vs. minimum support.

Fig.2 shows the runtime vs. the average size of transactions for both algorithms, where the average size of transactions varies from 4 to 14 for the synthetic dataset. As the average size of transactions increases, the runtime of the algorithm in [2] increases dramatically, however, compared to the algorithm in [2], the runtime of our proposed algorithm increases slowly. The reason for increase in the runtimes of both algorithms is that the number of frequent patterns increases as the lengths of transactions are increased. Therefore, finding candidates present in a trans- action takes a longer time. Our proposed algorithm is more scalable than the algorithm in [2] because a large number of candidates can be pruned by using the ARMNGA pruning strategy.

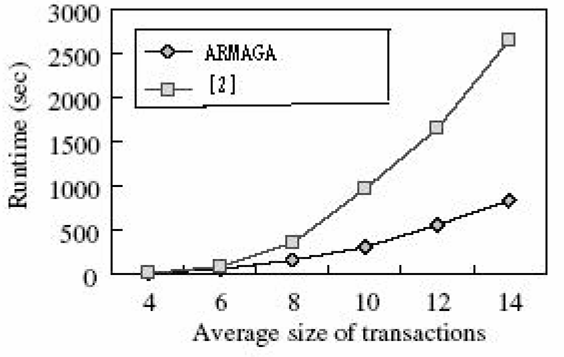


Fig.2. Runtime vs. average size of transactions.

Fig.3 shows the runtime vs. the size of an image for both algorithms, where the size of the image varies for the synthetic dataset. As the size of the image increases, the runtimes of both algorithms decrease; nevertheless, the runtime of the ARMAGA algorithm does not change very much.

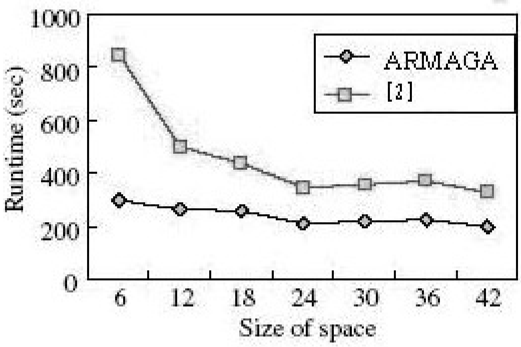


Fig.3. Runtime vs. image size

Fig.4 shows the runtime vs. number of objects for both algorithm, where the number of objects varies from 25 to 100 for the synthetic dataset .Since the average size of process and number of transaction are both fixed, the average support for the item sets decreases as the number of objects increases. Thus, the runtimes of both algorithms decrease slightly when the number of objects increases Nevertheless, our proposed algorithm is faster than the algorithm in [2].

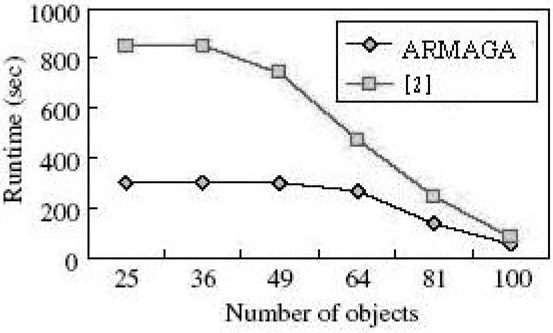


Fig.4. Runtime vs. number of objects.

From fig 1, 2, 3, 4, we can educe that ARMNGA has a higher convergence speed and more reasonable selective scheme which guarantees the non-reduction performance of the optimal solution. Therefore, it is better than GA and ARMA through the theoretic analysis and the experimental results.

6 CONCLUSIONS

The image data mining is a newly researching hot point in database area. But general data mining get knowledge from large quantities of data. We propose an Association rules mining based on a novel Genetic Algorithm, designed specifically for discovering association rules. We compare the results of the ARMNGA with the results of [2], and, it is better than GA and ARM through the theoretic analysis and the experimental results.

REFERENCES

[1] G. Bordogna, S. Chiesa, D. Geneletti,"Linguistic modelling of imperfect spatial information as a basis for simplifying spatial analysis,"in *Information Sciences 176* (2006) 366–389.

[2] Agrawal R, Imielinski T, Swami A,"Mining association rules between sets of items in large databases,"in *Proc. ACM SIGMOD Intl. Conf. Management Data*,1993.

[3] G. Chen, Q. Wei,"Fuzzy association rules and the extended mining algorithms,"in *Information Sciences 147* (2002) 201–228.

[4] Shijue Zheng,Wanneng Shu,Li Gao,"Task Scheduling Using Parallel Genetic Simulated Annealing Algorithm,"in *2006 IEEE International Conference on Service Operations and Logistics, and Informatics Proceedings* June 21-23, 2006, Shanghai, pp46-50

[5] P.Y. Hsu, Y.L. Chen, C.C. Ling,"Algorithms for mining association rules in bag databases,"in *Information Sciences* 166 (2004) 31–47.

[6]Li Gao, Dan Li, Dai Shangping,"A mining Algorithm of constraint based association rules,"in *journal of Henan University* Vol.33 (2003) pp.55-58

[7] Wu zhaohui,"Association rule mining based on simulated annealing genetic algorithm,"in *Compute Applications* Vol.25 (2005) pp.1009-1011

Li Gao is a associate professor, Department of Computer Science, Hua Zhong Normal University. She graduated from Hua Zhong Normal University in 1987; Her research interests are in distributed parallel processing, data mining, grid computing and e-commence.

Web Service Applications and Web-based Computing

Study of Hemodynamic Parameters Detecting Instrument Based on Embedded CPU Module*

Lin Yang, Da Li, Song Zhang, Yimin Yang
College of Life Science and Bioengineering, Beijing University of Technology
Beijing, 100022, China
Email: mickeymickey@emails.bjut.edu.cn

ABSTRACT

Objective: Cardiovascular disease is a serious threat to human life and health. Because of the frequent occurrences of cardiovascular disease, cardiovascular function has been noticed more and more. Methods: An Instrument Based on Embedded CPU Module is designed to detect hemodynamic parameters. It adopts simple rotatable coding switch for the input and control use, which is easy to handle. Results: The instrument completes the function of measuring blood pressure and printing the detection result rapidly. Otherwise, it can also save the result in order to track the detection and analyze the dynamic hemodynamic parameters. Conclusion: The instrument is easy to take and detect, non-invasively. It can be used not only in the clinic, but also in daily monitoring and detection of athletic cardiac function.

Keywords: Embedded CPU Module, Pulse Wave, Detecting Instrument

1. INTRODUCTION

Cardiovascular system is one of the most complex human systems. In the year of highly development in science and technology, many of the physiological problems of the cardiovascular system lack effective means of solution. Cardiovascular disease is a serious threat to (human) life and health. Because of the frequent occurrences of cardiovascular disease, cardiovascular function has been noticed more and more[1]. Cardiovascular flow parameters such as stroke volume, total peripheral resistance, blood flow semi-turnover rate reflect the cardiac function, vascular function and microcirculation function. Human cardiac function can be detected and analyzed objectively and accurately by these parameters through pulse wave contour[2][3][4]. In the prediction of cardiovascular disease, guidance in the treatment of cardiovascular and relevant diseases is extremely important.

Currently, complex algorithm and function design can be completed by PC-based detection system. However, because of PC's large cubage, high cost and inconvenient carrying, the use of the equipment is restricted. Further more, MC-based detection system has the advantage of small cubage, low cost and convenient carrying, however, has the disadvantage of limited resources, complex program, and difficulty of achieving complex functions. In this paper, a hemodynamic pulse wave parameters detecting instrument (MP-04) is designed using the embedded CPU module. The module system as a core module has such characteristics as small size, entire interface, faster operation speed, steady work, convenient update. At the same time, the module has the advantage of operating system support and simple software development. Therefore, more complex functions can be achieved.

2. SYSTEM DESIGN

Fig.1 shows the diagram of the overall system.

Embedded CPU module is the core module of the system. It is mainly responsible for receiving and showing data which is acquired by pulse wave data acquisition module, analyzing the hemodynamic blood flow parameters, and storing them and other parameters such as age, height, systolic pressure and diastolic pressure. It is easy to do the service of tracking detection and dynamic analysis.

The display device uses a resolution of 320 x 240 monochromes LCD. In addition, the system also uses a rotatable coding switch as the input and control interface, which makes the operation mode simple.

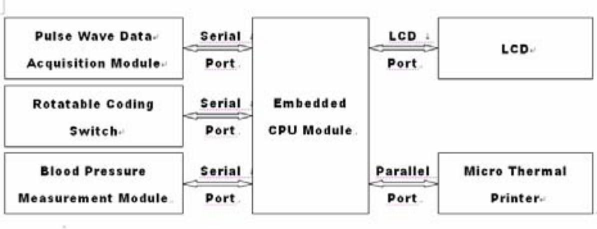


Fig.1. Structure frame of system

Pulse wave detection needs to input systolic pressure, diastolic pressure and other parameters. This system completes these two parameters detection automatically through blood pressure measurement module, which can avoid the inconvenient of manual measurement. CPU module sends control byte to initiate or terminate blood pressure module measuring blood pressure through serial port. Brachial artery blood pressure is tested by cuff module, using the oscillometric blood pressure measurement principles. After measuring, blood pressure module outputs heart rate, systolic pressure, diastolic pressure, mean pressure and other parameters.

Cardiovascular parameters can be computed after pulse wave detection. Hemodynamic parameters are printed by the embedded micro thermal printer through standard CENTRONICS parallel port. This small size face-shell printer can be easily embedded in equipment.

3. HARDWARE DESIGN

3.1 Embedded CPU Module

CPU module uses embedded processor 386, whose maximum operating frequency is 40 MHz. 4MB memory, four RS232 serial ports and one parallel port are integrated on the board. The board has CRT/LCD display and disk slot which can support multiple types of storage media (DiskOnChip /NVSRAM/Flash). We choose DiskOnChip 8M as the

* Technological project of Beijing "The development of portable detection and assessment system to athletic cardiovascular function" (No.Z0005190042011)

nonvolatile memory.

3.2 Pulse Wave Data Acquisition Module

Pulse wave data acquisition module consists of the front pulse wave analogical signal filter and amplification circuit and data acquisition circuit controlled by MCU circuit. As shown in Fig.2, the differential amplifier gain is adjusted by MCU through numerical control potentiometer. To obtain faster attenuation near the cutoff frequency, a band-pass filter with cascade connection of second-order lowpass and highpass is used. This filter can filter high frequency interference and low frequency baseline drift. Signal bias is adjusted by adjustable resistor through adjustable bias signal amplifier, thus it is in the input range of A/D converter.

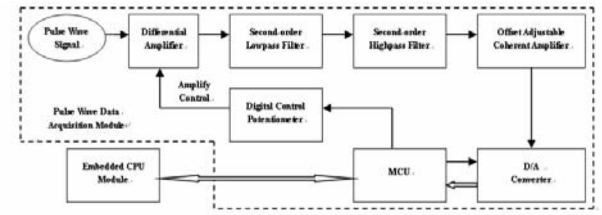

Pulse Wave Signal
Differential Amplifier
Second-order Lowpass Filter
Second-order Highpass Filter
Offset Adjustable Coherent Amplifier
Amplify Control
Digital Control Potentiometer
Pulse Wave Data Acquisition Module
Embedded CPU Module
MCU
D/A Converter

Fig.2. Structure frame of pulse wave sampling module

CPU module sends control byte to initiate and terminate the pulse wave data acquisition. A/D conversion is executed by MCU with a sampling rate. Main pulse wave signal frequency is 0～20Hz, therefore, 100 Hz sampling rate is used. MCU launches the A/D conversion as soon as receiving start orders. Sample rate is achieved by setting the timer overflow rate. When receiving suspension orders, MCU turns timers off immediately and terminates A/D conversion.

3.3 Rotatable Coding Switch

A rotatable coding switch is used as the input equipment of the whole system. When the switch turns clockwise, counterclockwise or pressed, MCU identifies the signal and sends the corresponding control byte to CPU module through the serial port, in order to complete the operation of rolling the menu upward or downward, adjusting input parameters and recognition of operations.

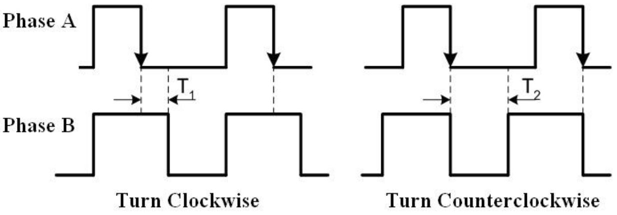


Fig.3. The output impulses of phase A, B from the rotatable coding switch

The switch has five pins, among which there are two for keystroke, two for the switch turning clockwise and counterclockwise, and the last for power supply. Two pulse signals are called phase A and phase B. The output impulses are shown in Fig.3.

As Fig.3 shows, both phase A and phase B increase when switch turns clockwise, but the falling edge of phase B is later than phase A. Both phase A and phase B decline when switch turns counterclockwise, but the rising edge of phase B is earlier than that of phase A.

The output of phase A is used to be the external interrupt source of MCU. MCU sets the falling edge as the interruption triggering method. When phase A has a negative jump and enters into an external interrupt service routine, as shown in Fig.4, the jitter removing time t1 and t2 should meet the condition of t1<T1 and t2<T2. T1 and T2 are determined by the rotation speed. Normally, T1 and T2 are both more than 5ms.

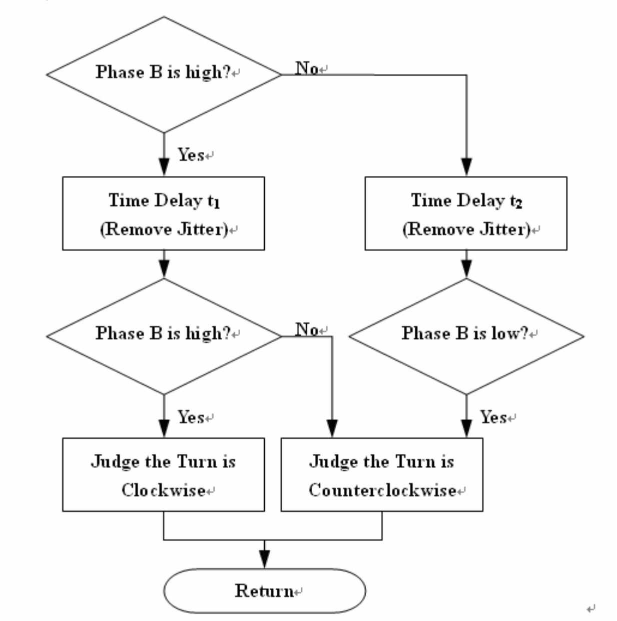

Phase B is high?
No
Yes
Time Delay t1 (Remove Jitter)
Time Delay t2 (Remove Jitter)
Phase B is high?
No
Phase B is low?
Yes
Yes
Judge the Turn is Clockwise
Judge the Turn is Counterclockwise
Return

Fig.4. The flow of external interrupt routine

It is similar when the switch button is pressed.

4. SOFTWARE DESIGN

Software is designed based on the development of CPU module. This module works similarly to Inter 386 and runs DOS system. Turbo C 2.0 integration environment is used. Program code can be edited, complied and debugged. A lot of library function provided by debug environment can be called to simplify the development process greatly.

4.1 Interface and Menu Design

The interface is developed in graphic mode through reading the 16 × 16 and 16 × 8 arrays of HZK16 and ASC16 to show the characters.

The menu interface is rolling brightened under the control of the rotary encoder switch. The software may enter the corresponding function of current menu by pressing the button. To achieve this function, each interface is described by the structure. The main attribute is shown in Fig.5. Each interface includes a certain number of menus which are followed by sequential numbers. A structure is used to show the display attribute of each menu. The display attribute mainly includes content, location, font and brightened. The pointer of menu structure points to the initial address of the menu structure.

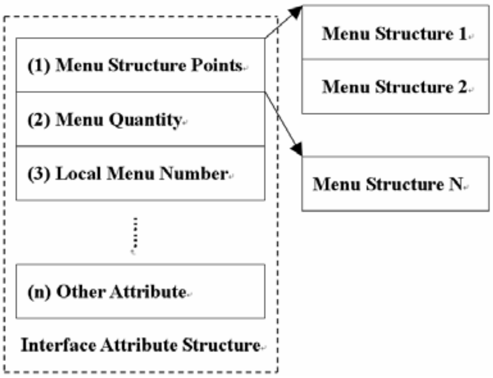


Fig.5. The attribute structure of interface

4.2 Serial and Printer Driver

The communication between CPU module and the other module is completed by the serial ports mainly. Serial initialization consists of baud rate setup, interrupt address setup and corresponding interrupt register setup. Printer is driven by standard sequence as printer we use is coincided with the CENTRONICS standard. The test results and pulse waveform are printed by sending printing order.

4.3 Pulse Wave Isolation and Hemodynamic Parameters Calculation

The data collected by pulse wave is a number of arrays which concludes many cycles of pulse wave. But the hemodynamic parameters we concerned are extracted from the isolation of a single cycle of the pulse wave. Therefore, a single pulse wave should be isolated.

600 data are selected to be the data source of single waveform after waveform leveling off. In order to separate single waveform from these 600 data, the key is to find the starting point and the end of a single pulse, that is, to find the minimum value. Then these data are computed as follows:

$$y(nT) = \frac{1}{10}[2x(nT) + x(nT-T) - x(nT-3T) - 2x(nT-4T)]$$

The signal computed has a quick ascending ramus, and the maximum value is close to the starting point of the pulse waveform. 450 intermediate points are selected to separate all the maximum values. Then a threshold value is computed by the maximum values. The threshold value will be closer to the minimum value. Then 300 intermediate points are selected to find all of the corresponding value with the threshold value (Suppose the number of value is N) .Go forward and find the value which is closest to zero. Again, 5 points near to the value are selected to find the minimum value. This value is the minimum point of pulse wave signal. Till now, N minimum points have been found in the 300 points. N points are corresponded to N-1 single wave. Then the final single wave is the wave of which the difference between the starting points and the end is the minimum.

Hemodynamic parameters calculation based on Wesseling's elastic tube model[5][6]. Cardio output, total peripheral resistance, blood viscosity and blood flow semi-turnover rate can be got according to the pulse waveform, systolic blood pressure, diastolic blood pressure, height, weight etc.

5. EXPERIMENTAL RESULTS

A detection instrument according to this paper is processed a corresponding experiment with TP-CBSII hemodynamic detection instrument, which is made by Beijing Redcom Co.,Ltd. A total of 45 subjects are examined. The radio pulse waveform correlation coefficient is 0.925. The correlation of cardiac output computed by pulse waveform is shown in Fig. 6.

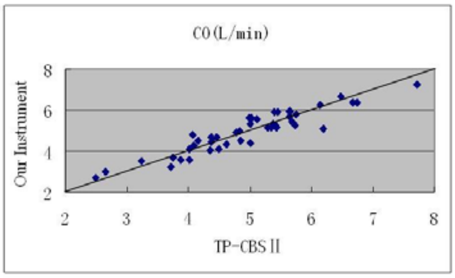


Fig.6. The contrast of the detected CO between this instrument and TP-CBSⅡ

6. CONCLUSIONS

The appearance of the detection instruments is shown in Fig.7. It is small in size, functional, simple operating. It is not only suitable for clinical care but also for daily monitoring on blood pressure, blood flow and other physiological parameters. It may improve the prevention and treatment of cardiovascular disease, which may help more people come to health from the status of sub-health.

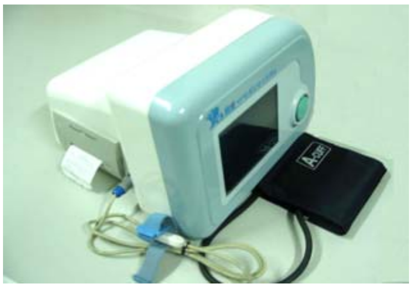


Fig.7. The appearance of this instrument

Meanwhile, the instrument is a non-invasive detection. With the combination of dynamic analysis of the test results, it can be used in the cardiac function detection of athletes. Then evaluation of the physical condition, training guidance and recovery suggestion can be shown to the players and their coaches to help them make improvement. The instrument will have a better future

REFERENCES

[1] KH Wesseling, *A century of noninvasive arterial pressure measurement from Marey to Penaz and Finapres*, Homeostasis, 1995,36:2-3

[2] Luo Zhichang,etal, *A new pulse contour method for noninvasive estimation of cardiac output.Automedica*, 1998,17:127-141.

[3] Mustafa Karamanoglu, "A System for Analysis of Arterial Blood Pressure," *Computers and Biomedical Reasearch*,1997,30:244-255

[4] KH Wesseling, Ben de Wit, Jan E.W.Beneken, "Arterial haemodynamic parameters derived from noninvasively recorded pulsewaves, using parameter estimation," *Medical and Biological Engineering*, 1973,11:724-731

[5] KH Wesseling, *A simple device for the continuous measurement of cardiac output.ADV Cardiovase Phys,*

1983, 5(Part II):16-52.

[6] Wormersly J R. “An elastic tube theory of pulse transmission and oscillatory flow in mammalian arteries,” *WADC Tech Rep*, 1967, 56-614.

Lin Yang is a doctoral student of College of Life Science and Bioengineering in Beijing University of Technology. He graduated from Qingdao Technological University in 2003 with specialty of automation. His research interests are in theory and application of hemodynamic parameters detection.

Distributed Campus Management System Based on SOA

Yang Yang, Qingping Guo
School of Computer Science and Technology, Wuhan University of Technology
WuHan, HuBei 430063, China
Email: honey0371_yy@163.com, qpguo@mail.whut.edu.cn

ABSTRACT

As the enterprise is trending to large-scale environment， originally information management system and existing system will partly not suited for the changing requirement . How to make most of existing information systems and set up low-cost, open and flexible integration system have already become the key factor to constructing information system in universities. Service oriented architecture (SOA) provides a solution to improve the reusability, scalability and efficiency in software development. Building a distributed management system based on Service oriented architecture (SOA) brings many attractive characteristics such as great efficiency, higher fault tolerant ability, good load balancing and reliability etc. in this paper I illustrate distributed campus management system based on the technique, comparing to the traditional system based on SOA, I have add my own strategy aiming to improve the search speed and efficiency when access some web services.

Keywords: SOA, WSDL, UDDI, Web Service, Web Service Chain, Mapping Table

1. INTRODUCTION

In recent years, developments in networking and telecommunications have opened up enormous opportunities for linking up disparate information sources and computational modules. This has led, on one hand, to the development of distributed information systems that integrate dispersed information sources. On the other hand, considerable interest has been generated in the area of software interoperability: the linking and integration of software modules to carry out complex computational tasks. An example of this second type of software integration is that of service oriented architecture which provides support for the automation of business or industrial processes involving human and machine-based activities. By using such architecture, organizations can accelerate throughput, reduce costs, and monitor performance of common, well-understood operational processes in their domain. Service Oriented Architecture (SOA) is becoming a favorite choice of software architects who struggle to provide solutions for distributed applications, while maintaining manageable system architectures. SOA building block is a set of loosely-coupled services. A service provides a unit of functionality by exposing its abstract interface on the network. The business functionality is implemented as coordinated interactions of services. SOA allows heterogeneous components to be easily integrated to satisfy business requirements. Furthermore, such systems are more flexible and adaptable than traditional. The paper will show you an architecture base on SOA which make you in detail understand the advantage of the technology.

Firstly, this paper introduces simple the concept of web service and SOA, and analyzes the core technologies to construct Web services architecture, XML, SOAP, WSDL and UDDI in detail. Secondly based on web services architecture technology, an open and distributed dynamic campus management system structure is brought out. Some suggestions to improve the access speed are put forward. Finally, the front design done on dynamic campus management system is summarized and further work is prospected.

2. THE ARCHITECTUREAND CHARACTER OF SOA

2.1 System Architecture of SOA

Service-oriented architecture (SOA) is not a new concept, back in 1996, Gartner Group had raised SOA. As yet, there is no one unified, widely accepted definition. In general view, SOA is a component model that links different functional modules of applications through the definition of a good interface and service contract (contract) .Interface is defined in neutral manner and is independent of specific hardware platforms, operating systems and programming languages, so that the system communications in a unified and standardized way. Such a neutral interface definition (no mandatory binds to the realization of specific) features is called loosely coupled [1].

At a conceptual level, SOA is composed of three core pieces:

(1) Directory: It also has another name “Registration Center” because it acts as an intermediary between providers and consumers. Most of these directory services are categorized by scientific taxonomies.
(2) Service provider:The Service Provider defines a service description and publishes it to the Registration Center.
(3) Service consumer: The service requester can use the directory services' search capabilities to find service descriptions and their respective providers.

The three activities the service consumer and provider in a SOA as depicted in fig. 1[2] can perform are:

(1) Publish: The service provider has to publish the service description in order to allow the requester to find it. Where it is published depends on the architecture.
(2) Discover: In the discovery the service requester retrieves a service description directly or queries the Registration Center for the type of service required.
(3) Invoke: In this step the service requester invokes or initiates an interaction with the service at runtime using the binding details in the service description to locate, contact and invoke the service.

2.2 Character of SOA

SOA has the following characteristics

(1) Services are loosely coupled. Services between requesters and providers are loosely coupled. Service requester doesn't know the technical details of providers, the requester requests and responds via messages instead of APIs.

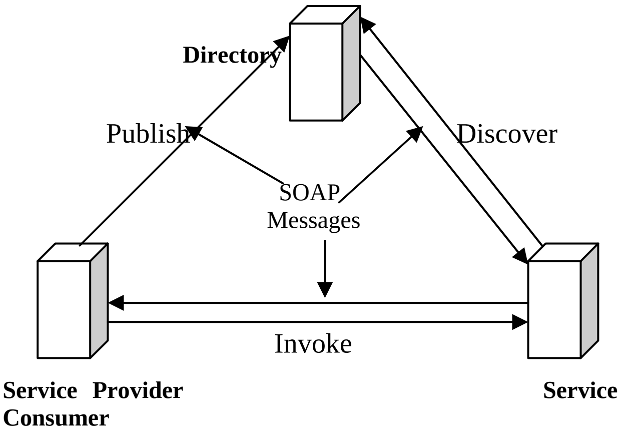


Fig.1. SOA architecture

(2) The thick granular service interface: the user and the service layer don’t need to reciprocating interact repeatedly
(3) Reused service: in order to implement the high levels of reusability, the service only worked in the particular processing context and was independent from the bottom implementing and the vary of consumer’s request. Designing the reused service is the most valuable job alike the database designing and the usual data modeling.
(4) The standard service interface: The abroad application and deep development of the XML and WEB service in the electron business affairs push the SOA toward a higher level and greatly increase the value of the SOA.
(5) Can be accessed from outside: the exterior consumer which is usually called business associate can access the same service like the corporation interior consumer.
(6) Can be used whenever: when the consumer calls the service, the SOA requires that there must existing the service provider to response it.
(7) Classification: using the different thick granularity grade to create service could resolve the problem of bad currency and difficulty of the ability to reuse.

3. TECHNOLOGIES FRAME OF WEB SERVICES

In order to realize Web services, there is self-definite protocol inter OP stack in Web services system. Protocol inter OP stack is new protocol inter OP technologies. They mainly have XML, SOAP, WSDL and UDDI. That is as follows in Fig.2 [3]:

XML (Extensible Markup Language) is a kind of markup language that may be created into self-definition markup. So the markups in paper files can be given some meaning by XML. XML files are made of markup, element and attribution. XML exchanges data simply and support intelligent search.

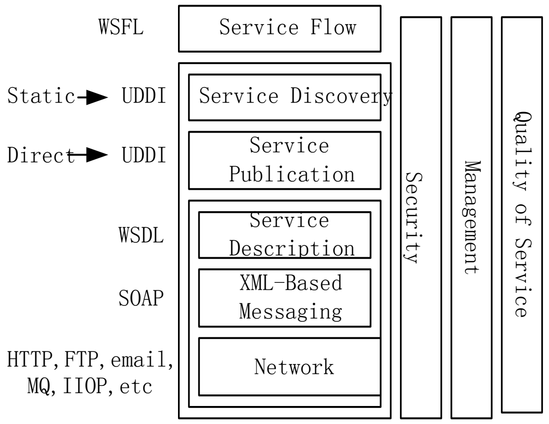


Fig.2. Web Services Protocol Stack

SOAP (Simple Object Access Protocol) is simple protocol that is used for communicating information in the distributed environment. It is based on XML protocol, and contains four parts: SOAP envelop defines that what is its content in message depiction; who sends the message; who should accepts and deals with it; how to deal with their frame.

WSDL (Web Services Description Language) provides a kind of grammar that can be described service into a group of information exchange port. WDSL files are a kind of depiction that is not relate to language and platform. WSDL describes services, visiting method, expecting response pattern. WSDL files can be exchanged through private or UDDI register center. And WSDL is also file format that is based on XML. It is used for pattern describe, message, operation, interface, position and protocol retained.

UDDI (Universal Description, Discovery, and Integration) provides a kind of middle system that is used for published and located service describe. UDDI supports different service definition patterns such as WSDL files, standard JAVA interface and XML files. UDDI describe all API of register center. The API finishes two basic tasks: register enterprise and service, locate and bind a service registered. Register and location is finished by means of UDDI command being put in body of SOAP message and sent into register center [5].

4. THE RELATION BETWEEN SOA AND WEB SERVICE

Web service is a technical specification, but the SOA is the design principle. Especially WSDL in the Web service, is an interface definition standard of SOA, this is the basic contact between the Web service and SOA. In essence, web service is a realization form of SOA, but not the only way to realize SOA (for instance CORBA). Beyond all doubt, web service is the most popular and successful form. The SOA is a very virtual concept, only put forward the concept of interface and protocol and without the realization and embodiment, but Web service embody them: the protocol which the web services use are all based on XML; the SOA should only be considered that it has three kinds of roles, but the web service in these three kinds of roles all have their respective specific way of realization. Fig.3 [4] reflects the mapping relation of SOA and the web services.

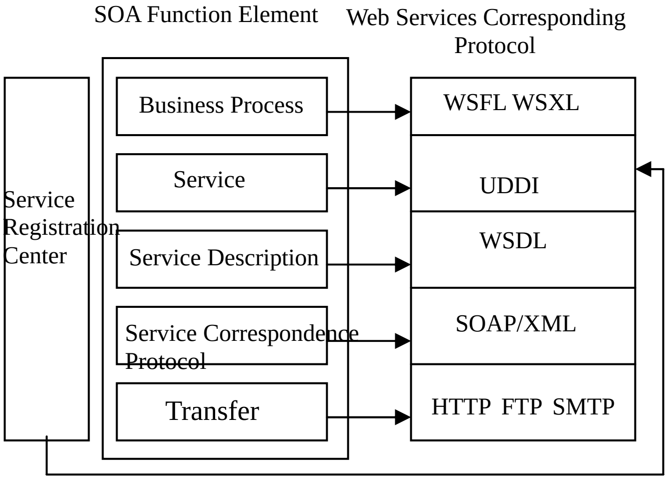


Fig.3. Web Services Protocol Stack

Seeing from the above mapping relation, the Web Services is the most suitable technique aggregation to implementing the SOA recently.

(1) Through looking up the catalog, we can change the service providers dynamically rather than influence the applications configuration.
(2) Through using the WSDL and the request of the SOAP based on text, we can implement the interface which could receive a great deal of data one time.
(3) All the communication of the Web Service is carried out through the SOAP based on XML, and the different versions could be simply discerned by the different DTD or XML Schema.

5. ANALYSIS OF DISTRIBUTED CAMPUS MANAGEMENT SYSTEM BASED ON SOA

The model I will introduce can be used in many field which has the similar characteristics, such as some large-scale enterprise who own plenty of subsidiary companies and the university which ever been combined by two or three small-scale college, due to some reasons, the database system of each dispersed locations can not be congregated together. My university is a representative example, which was integrated by three colleges. So this paper I will expound the system through depicting a supposed distributed architecture of certain integrated university, of course based on SOA.

As I had referred above, the dispersed campus or college has their own database system, and it is difficult to congregate the database, we hope the management architecture we construct can realize the systemic unification and interaction at logic although distribute in the different physical structure. So we will choice proper method to make the information and database interact each other and realize the unification of data and manipulation. SOA is the best choice, whose advantage and benefit have been detailed introduced above. Adopting this scheme can solve many rough problems as follows[6]:

(1) The data of each system can be shared at any moment.
(2) The function of each system and system itself all can be reused under some condition.
(3) The manger owning certain privilege can carry out the overall management to the whole university system, each system is loosely coupled
(4) The system has the ability of enlargement and expansion.
(5) The system has the strong compatibility which can make compatible with existing system and satisfy the requirement at the lowest cost.
(6) The service can provide the secure service access which also can be controlled.

6. THE OVERALL ARCHITECTURE OF DISTRIBUTED CAMPUS MANAGEMENT SYSTEM BASED ON SOA

We all know that the most advantage of utilizing the service oriented architecture is the fact that all existing and useful information system will reuse in the new system model, which can reconstruct the new management architecture at the lowest cost. As to university, there are plenty of existing system can be reused after combination. For example, OA system used by administrator、students management system and so on.

However, existing system can not be used directly, they must be packaged and constructed into web service, and then by special tools these web services will be published to the Registration Center, where services can be categorized and managed in uniform mode. When the client or the user wants to access the web service or employ certain function of some services, they only call these web services using HTTP through SOAP protocol.

The detailed procedure of one web service formed from packaged to be employed was as follow:

(1) The service provider registered the web service in the UDDI Registration Center after one web service packaged and encapsulated.
(2) When the users need some web services, they will search the web services from UDDI Registration Center and read the WSDL document about the web services.
(3) If the users found the web services they needed, they will call the services through SOAP request message encapsulated by HTTP request.
(4) If some problems occur during the whole process, the service provider will redeploy the web service and meanwhile renew the related technique description information at UDDI registration center

From the steps illustrated above, we can see that after the reconstruction and encapsulation, the existing systems are all presented in the form of web services and waiting for being visited and called. Then I will show you the overall architecture and the design thoughts.

As shown in Fig 4, the system I design contains the following layers: the presentation layer, the interface service layer, the internal web service layer, and the database layer. The presentation layer is set of web-based user interfaces for customers to manually manage the configuration of the system and perform transactions. The interface service layer exposes a set of web service interfaces for clients to interact with the system automatically.

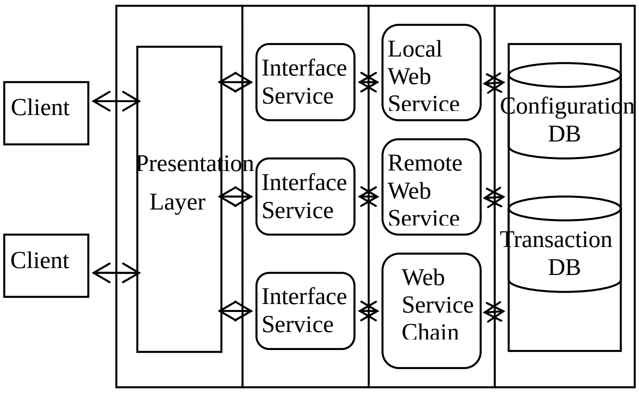


Fig.4. multilayer architecture of distributed campus management system based on SOA

Web services in the system are grouped into three different types service groups: local web service、remote web service and web service chain. Each type has special domain of being visited. Local service means the user only access the local database system and local web service; remote service refers to the user only visit the remote service; web service chain refers to a series of continuous services performed in one time automatically under the control of the workflow process which have been defined previously, it can also been comprehended to the an array web services containing several local web services and remote web services, as shown in Fig 5.

Each group of services is deployed in the same server. Resiliency and scalability are achieved by deploying the same group of services on a cluster of server boxes. Then we will focus on the problem of how distinguish between the three types web service. To solve the problem, we will induce the concept of configuration parameters.

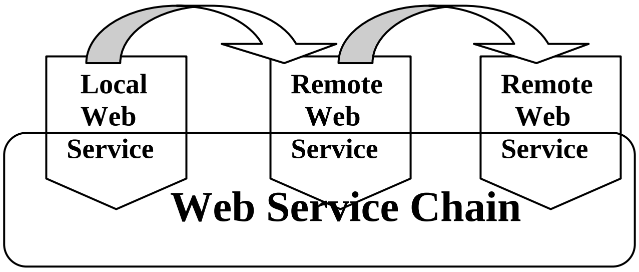


Fig.5. web service chain

In order to distinguish diverse type web services, each service may require a different set of configuration parameters. To simplify the issue, we centralize the management of system configuration by using a single configuration database for the entire system. All servers bootstrap there configuration by loading the information from the configuration database on startup and the configuration database can list the special configuration parameter of each type web service, we also can dynamically add the new parameter when the new web service bring out. To support the dynamic update of configuration information, when the configuration is updated, a message is sent to a notification service. The notification service then notifies each server regarding the configuration update.

Another of the challenges in our implementation is efficient message passing and transformation. To avoid unnecessary overheads of message transformation, messages are passed in their native format through message queues. However, when messages go through an interface service, it is transformed to a compatible format, in most cases, an XML document. When an error occurs during a transaction, details of the error are sent to an exception message queue. Messages in the queue are then reviewed by the administrator for appropriate handling. The system requires tractability of the history of all transactions, in case disputes or legal issues arise. The system provides an audit log service, which logs the history of all messages. The audit log can be reviewed, monitored, and queried.

7. THE DISTRIBUTED IMPLEMENT OF CAMPUS MANAGEMENT SYSTEM BASED ON SOA

I do hope the distributed system I design display the distributed characteristics distinctly, so firstly I will illustrated some drawback I found from the traditional system based on SOA: the most serious disadvantage is the speed problem, and the countless search work is processing repeatedly. So I just put forward my opinion aiming at making some improvement on this problem.

In order to improve the performance and speed, I design some tactics or some engine to quicken the search process.

I just conceive of designing some mapping tables (mapping to UDDI registration center) at each local database to solve the problem. And there must be three tables at least, respectively record the mapping information of different type service: local web service、remote web service and web service chain. The table used for recording local web services can be design into the form of the UDDI registration center, it acts the role as the cache memory in the computer, because the record sequence of the table I design is arranged according to the frequency of certain local web service being accessed and used, we just can consider the table to be the local UDDI registration, the user doesn’t need search web service information from the UDDI registration center.

The second table I introduce is the record table of remote web service, comparing to the first type table, the most obvious distinction is the service domain is not local, but the remote, so the design construction of the table doesn’t need to be complex and detailed as the first type table, because it can’t utilize the predominance of local resource including local database resource and local web service resource. So we just can design the second type table to be the mapping table which records the position of the remote web service at the UDDI registration center. Of course, the record sequence of the table is arranged according to the frequency of remote web service being accessed.

The third type table I illustrate aims at the web service chain, we all know, it is inevitable that some function or operation will access several services in one times, we can consider the situation to be a web service chain. According to its feature, we can design the third table at the foundation of the second type, but we should also distinguish the distinct character of the two type. The third type table should focus on the web services sequence and record the each service sequence number at UDDI orderly, other features designed are similar with the second type table.

There are some common feathers I must refer here, each record of the three table is added when the user implement some operation, the operated web service’s related information and sequence number will be recorded at corresponding table according to the configuration parameter, the process has some common points with log files. Meanwhile the record of each table will be arranged according to the frequency of web services

8. CONCLUSIONS

With the development of distributed management architecture of enterprise and university, many researches have proposed the distributed technique scheme to deal with the management of different systems. In this paper, we propose the distributed management system based on service oriented architecture. This architecture brings many attractive characteristics such as great efficiency, higher fault tolerant ability, good load balancing and reliability etc. Compared with traditional distributed management system, a distributed management system based on service oriented architecture has the following advantages. It supports large-scale applications, has better load balancing and fault tolerant ability, and make execution more efficiently. Compared with other distributed management systems, it can be used in a larger scale and supports service distributed on the internet. And it is more convenient for managing the changes in systems, and more adaptable to the dynamic environment. We argue that, as the benefit showing above, building a distributed workflow management system based on service oriented architecture can help enterprises to manage their processes flexibly and efficiently with minimum deployment cost.

REFERENCES

[1] Bai Xiaoming, Song Ruliang, Hou Zonghan. “The study on secure distributed workflow architecture based SOA.” 2006 International Conference on Power System Technology.

[2] Chen Dan, Yuan Jie. “Appling SOA to distributed scientific research information system.” *Computer Engineering and Design*.2006;27(24):4759-4761

[3] Doug Tidwell, *Web services—the next revolution on Web*, in 2001. http://www-900.ibm.com/developerWorks/

[4] Sun Hualin, Zhao Zhengwen. “Study on service-oriented architecture based on Web Service.” *Information Technology*. 2007, 1:50-53

[5] Shao Zhenye, Hua Junhan. “A Design and Implementation of Dynamic E-business System Based on WEB Services.” *DCABES 2004 PROCEEDINGS*. 2004:516-520

[6] Wang Jinling, Zhu Shisheng, Fu Qunwei. “Study of SOA Department Based on Web Service.” *Modern Electron Technology*. 2007, 4:155-158

Yang Yang is a master degree candidate in the School of Computer Science and Technology, Wuhan University of Technology. She majors in computer application and her research interests are web service and network security.

Research on Internet Voting Schemes and Protocols *

Bo Meng [1], Huanguo Zhang [2]
[1] School of Computer, South-Center University for Nationalities
Wuhan, Hubei 430074, China
[2] School of Computer, Wuhan University
Wuhan, Hubei 430072, China
Email: [1] mengbo@263.net.cn.[2] liss@whu.edu.cn

ABSTRACT

With the popularization of Internet and advance of process of democracy of nation, the desire of Internet voting is more and more intense. Internet voting protocol is the key and base of Internet voting scheme. In this paper, firstly, the Internet voting model and homomorphic encryption scheme, blind signature scheme and Mix net scheme are analyzed. Secondly, the status and implementation of properties that Internet voting protocol should have are introduced. Thirdly, the property of invariableness is proposed by us. Finally we analyze the typical protocols such as FOO, CGS, JCJ, and ACQ according to the properties that the Internet voting protocol should have.

Keywords: Internet Voting Scheme, Internet Voting Protocol, Protocol Security, Electronic Government

1. INTRODUCTION

Voting is that a formal expression of preference for a candidate for office or for a proposed resolution of an issue. Down the ages all kinds of technologies and tools were used to vote in the society activities, such as stones, paper ballots, datavote, punchcard, electronic voting and so on.

With the popularization of Internet and advance of process of democracy of nation, a new voting system called Internet voting is introduced. Internet voting is that voting done by using a computer to cast a ballot over the Internet. Internet voting is classified four types by R. M. Alvarez and T.E. Hall in [1]. They are remote Internet voting, Kiosk Internet voting, Polling place Internet voting, and Precinct Internet voting.

Remote Internet voting is voting by using a computer that is not under the physical control of election officials; the ballot is cast over an Internet connection. Kiosk Internet voting is that voting is done at certain locations by using a computer under the physical control of election officials to cast a ballot over the Internet. Polling place Internet voting is that voting done at any valid polling place by using a computer under the physical control of election officials to cast a ballot over the Internet. Precinct Internet voting is that voting that is identical to polling place Internet voting except that the voter can vote only at his or her own precinct polling place. Unless otherwise indicated, when we say Internet voting we mean remote Internet voting.

In this paper, firstly, the Internet voting model and the Internet voting scheme composed of homomorphic encryption scheme, blind signature scheme and Mix net scheme are researched. Secondly, the status and implementation of properties that the Internet voting protocol should have are introduced. Thirdly, we propose the property of invariableness. Finally we analyze the typical protocols such as FOO, CGS, JCJ, and ACQ according to the properties that the Internet voting protocol should have.

2. INTERNET VOTING SCHEMES

The participants in Internet voting mainly consist of voter, registration authority, tallying authority, and authority generating the ballot, bulletin board. Bulletin board is publicly readable. Any participant can write in his own section, but nobody can delete or change anything in the bulletin board. Internet voting is composed of four main phrases: preparation phrase, registration phrase, voting phrase and tallying phrase.

Internet voting has been researched for about twenty years. Internet voting protocol is the key and base of Internet voting scheme.

Internet voting scheme can be classified into two types based on if they need authority. One type needs not authority, such as [2]. This kind of protocols is fewer. The other type needs authority. Many Internet voting protocols [3~33] belong to this type. These protocols can be categorized by different technologies into three schemes: homomorphic encryption scheme, blind signature scheme and mix net scheme.

2.1 Homomorphic Encryption Scheme

Internet voting protocols [3~18] belong to homomorphic encryption scheme. Homomorphic encryption is used in this kind of scheme. Homomorphic encryption method is that encryption of the sum of ballot is obtained by multiplying the encrypted votes of all ballots The purposes of homomorphic encryption method are protection of the voter's privacy and advancement of the efficacy of tally ballots. Generally the homomorphic encryption scheme is not receipt-free. The first voting protocol of this scheme was introduced by Benalooh[16]. The voting protocol proposed by Cramer, Gennaro and Schoenmakers [5] is representative. Homomorphic encryption scheme is described as in Fig.1.

(1) The authority generates ballot and sends them to the bulletin board.
(2) The voter makes his choice and gets the ballot from the bulletin board correspondingly and encrypts it with the homomorphic cryptosystem.
(3) The voter sends encryption of ballot with the homomorphic cryptosystem to the bulletin board.
(4) Due to the homomorphic property, an encryption of the sum of ballot is obtained by multiplying the encrypted votes of all ballots. Finally, the result of the election is computed from the sum of the ballots, which is jointly decrypted by the authorities.

*Supported by the National Natural Science Foundation of China (60373087,60473023) and by the Foundation of South-Center University for Nationalities (YZZ06026)

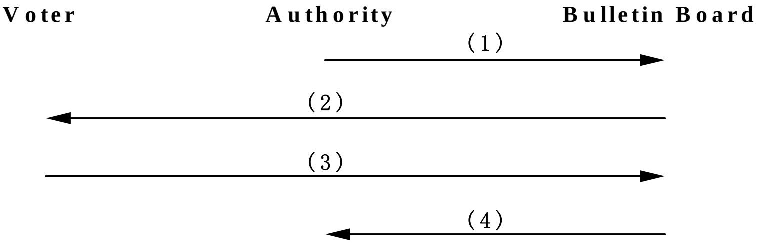


Fig.1. Homomorphic encryption scheme

2.2 Blind Signature Scheme

Internet voting protocols [19~26] are blind signature scheme. The scheme described as in Fig.2 uses blind signatures technology. The voting protocol proposed by Fujioka, Okamoto, and Ohta in [21] is representative. Generally the blind signatures scheme is not receipt-free because the blinding factor is the receipt. The voter firstly obtains a token, a blindly signed message unknown to anyone except himself. Next, the voter sends his token together with his vote. These protocols require voter's participation in more rounds. Generally the protocols need two authorities. One is administrator which responsible for issuing the ballots and generating the blind signature of ballots. The other is collector which responsible for tallying the ballot and publish the result.

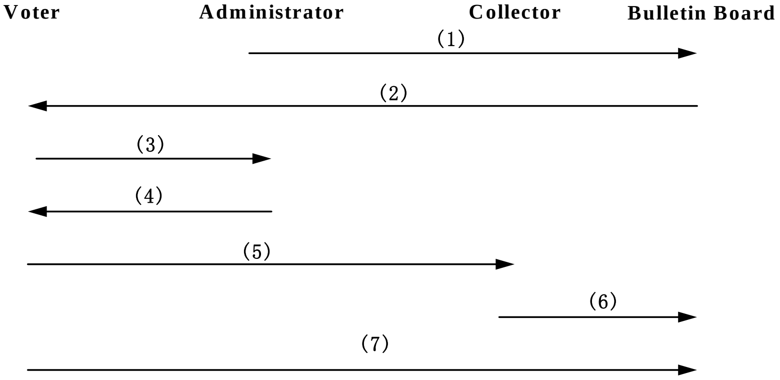


Fig.2. Blind signature scheme

(1) The authority generates ballot and sends them to the bulletin board.
(2) The voter makes his choice and gets the ballot from the bulletin board correspondingly.
(3) The voter creates the signature of his blinded ballot and sends it to administrator.
(4) The administrator verifies the voter's signature of his blinded ballot and generates the blind signature of his blinded ballot and sends it to the voter.
(5) The voter removes the blinding factor and gets the signature of the ballot. Then the voter sends it to the collector through anonymous channel.
(6) The collector checks the administrator's signature of the ballot and sends them to the bulletin board.
(7) The voter checks if his ballot is on bulletin board.

The protocol ends.

When all of the voters had voted, the collector begins tallying and publishes the result of voting on bulletin board

2.3 Mix Net Scheme

The third scheme is Mix net scheme, such as [9,14,27,28,29,30,31,32,33]. The key idea of Mix nets is to permute and modify the sequence of objects in order to hide the correspondence between elements of original and final sequence. David Chaum introduced this idea in 1981 as a realization of anonymous channel. The protocols of mix net scheme use the mix net to mix the possible ballot and sending done permutations secretly to the voter. The scheme describe as in Fig. 3.

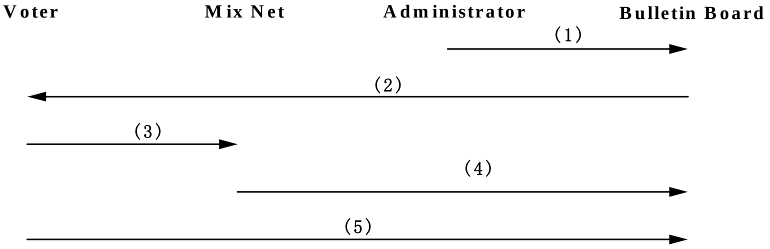


Fig.3. Mix net scheme

(1) The authority generates ballot voted t and sends them to the bulletin board.
(2) The voter makes his choice and gets the ballot from the bulletin board correspondingly.
(3) The voter sends his ballot to Mix net
(4) The Mix net processes the ballot and sends them to bulletin board.
(5) The voter checks if his ballot is on bulletin board. The protocol ends.

When the voting phase ends, the authority begins tallying and publishes the result of voting on bulletin board.

3. ANALYSIS OF THE INTERNET VOTING PROTOCOLS

The secure and practical Internet voting protocols should have the following properties:

Basic properties:
Privacy: all votes should be kept secret in the voting. No coalition of participants not containing voter himself can gain any information about the voter's vote. Privacy means that the link between the voter and his vote is disposed or inaccessible to everyone (including authority), even if all of the public communication is monitored.
Completeness: All valid votes should be counted correctly.
Soundness: Any invalid vote should not be counted
Unreusability: No voter can vote twice
Fairness: No participant can gain any knowledge about the partial result before the tallyinging stages because the knowledge of the partial result could affect the intentions of voters who have not yet voted
Eligibility: only eligible voters can cast the voters. Every voter can cast only one vote.
Except above these properties, we think that the Internet voting protocol should have property of invariableness. The invariableness is that whatever tally the ballot the result of vote is invariableness.

Expanded properties:
Universal verifiability: Any one can verify the fact that the election is fair and the published tally is correctly computed from the ballots that were correctly cast.

Receipt-free: The voter cannot produce a receipt to prove that he votes a special ballot. Its purpose is to protect against vote buying. Receipt-free property was introduced by Benaloh and Tuinstra [3]. They proposed a receipt-free scheme based on the voting-booth. Hirt and Sako in [9] point out that their scheme is not receipt-free.

Coercion-resistant: A coercion-resistant [13] voting protocol should offers not only receipt-free, but also defense against randomization, forced-abstention, and simulation attacks [13].
Randomization attack: The idea is for an attacker to coerce a voter by requiring that she submit randomly composed balloting material. The effect of the attack is to nullify the choice of the voter.

Forced-abstention attack: This is an attack related to the previous one based on randomization. In this case, the attacker coerces a voter by demanding that she refrain from voting.
Simulation attack: an attacker coerce voters into divulging private keys or buying private keys from voters and then simulating these voters at will, i.e., voting on their behalf.

Research on Internet voting protocol focuses on how to implement these properties. At present the hot point is how to realize receipt-free and coercion-resistan with few assumpition and constraints.

A lot of Internet voting protocols have implement Receipt-free and Coercion-resistant through ad hoc physical assumptions and trusted third parties, such as, one- or two-way untappable channels and/or anonymous or private channels [9,19,27]; third-party (trusted) honest verifiers [34]; smart cards [30]; tamper-resistant machines [10] ; third party randomizers [6,14,39]; voting booths [3,35] ; with special visual encryption tools [36]). Also schemes based on deniable encryption [37,38]. Reliance on ad hoc physical assumptions or trusted third parties is problematic, because it undermines the security, flexibility, robustness, trustworthiness, and ease of use of an election scheme.

The protocols in [13,11] are better in the implementation of Receipt-free and Coercion-resistant prosperities. They don't use strong physical assumption.

Research on the coercion-resistant is at the beginning. It is researched first in [13,11]. [13,11] mainly applied the credential of voter and designated verifier proof to accomplish it. Voter can cheat the coercer by producing a false credential. Owning to designate verifier proof the coercer cannot verify the proof.

In the protocol [11] the election authorities provide shares of credentials to each voter, along with designated verifier proofs of each share's validity. Using homomorphic encryption, the voter assembles the shares and combines them with her own vote that is cast on a public bulletin board. All messages in the bulletin board can be decrypted by a coalition of the election authorities after the voting phase of the election is completed.
The key idea in the protocol [13] is for the identity of a voter to remain hidden during the election process, and for the validity of ballots instead to be checked blindly against a voter roll. When casting a ballot, a voter incorporates a concealed credential. This takes the form of a ciphertext on a secret value that is unique to the voter. The secret value is a kind of anonymous credential. To ensure that legitimate voters cast ballots, the tallying authority performs a blind comparison between hidden credentials and a list of encrypted credentials published by an election registrar alongside of the plaintext names of registered voters. The idea of protocol [13] is good. But we find that the protocol is not the receipt-free and coercion-resistant they claimed it is receipt-free and coercion-resistant in [13].

According to these properties we analyze these protocols of FOO [21], CGS [5], JCJ [15], and ACQ [11]. These protocols are typical. Owning to the space limitation we only give the analyzed result described as in Table 1.

Table 1. Result of analysis

Properties	FOO [21]	CGS [5]	JCJ [15]	ACQ [11]
Privacy	√	√	√	√
Completeness	√	√	√	√
Soundness	√	√	√	√
Unreusability	×	×	×	×
Fairness	√	√	√	√
Eligibility	√	×	×	×

Invariableness	√	√	√	×
Universal verifiability	√	√	√	√
Receipt-free	×	×	√	×
Coercion-resistant	×	×	×	×

4. CONCLUSIONS

With the popularization of Internet and advance of process of democracy of nation, the needs of Internet voting are more and more intense. Internet voting protocol is the key of Internet voting scheme. In this paper we research the Internet voting model and the Internet voting scheme composed of homomorphic encryption scheme, blind signature scheme and mix net scheme. At the same time the status and implementation of properties that the Internet voting protocol should have are introduced. Thirdly we proposed the property of invariableness. Finally we analyze the typical protocols such as FOO, CGS, JCJ, and ACQ according to the properties that the Internet voting protocol should have. From the analysis result we can know that until now the practical Internet voting protocol with these properties have not introduced.

REFERENCES

[1] R. Michael Alvarez and Thad E. Hall, Point, Click, and Vote: *The Future of Internet Voting, Brookings Institution* Press, 2004. http://www.brookings.edu/press/books/pointclickandvote.htm.

[2] R. A. DeMillo, N. A. Lynch, M. Merritt, "Cryptographic Protocols", In the *proceeding of Of 14th Annual ACM Symposium on Theory of Computing*, 1982,pp. 383-400.

[3] Josh Benaloh,Dwight Tuinstra,"Receipt-free secret-ballot elections," in *the proceeding of STOC '94*, 1994,pp.544–553.

[4] Ronald Cramer, Matthew Franklin, Berry Schoenmakers, "Moti Yung. Multi-authority secret ballot elections with linear work," In *the proceeding of EUROCRYPT '96*, Springer-Verlag, LNCS1070, 1996,pp.72–83.

[5] Ronald Cramer, Rosario Gennaro, Berry Schoenmakers. "A secure and optimally efficient multi-authority election scheme," in *the proceeding of EUROCRYPT '97*, Springer-Verlag, LNCS 1233,1997,pp.103–118.

[6] Olivier Baudron,Pierre-Alain Fouque,David Pointcheval,Guillaume Poupard,Jacques Stern. "Practical multi-candidate election system," In *the proceeding of PODC* '01, ACM, 2001: 274–283.

[7] Ivan Damg°ard , Mads Jurik,"A generalisation, a simplification and some applications of paillier'sprobabilistic public-key system," In *the proceeding of Public Key Cryptography* '01, Springer-Verlag,LNCS 1992,2001,pp.119–136.

[8] Ivan Damg°ard,Mads Jurik,Jesper Buus Nielsen,"A generalization of paillier's public-key system with applications to electronic voting,"2003, http://citeseer.ist.psu.edu/cache/papers/cs/27234/http:zSzzSzwww.daimi.au.dkzSz~ivanzSzGenPaillier_finaljour.pdf/damgard03generalization.pdf.

[9] Martin Hirt and Kazue Sako,"Efficient receipt-free voting based on homomorphic encryption," In *the proceeding of EUROCRYPT* '00, Springer-Verlag, LNCS 1807,2000,pp.539–556.

[10] Byoungcheon Lee , Kwangjo Kim,"Receipt-free electronic voting scheme with a tamper resistant randomizer," In *the proceeding of ICISC2002*,2002,pp. 405–422.

[11] Alessandro Acquisti,"Receipt-Free Homomorphic Elections and Write-in Voter Verified Ballots,"*Technical Report* 2004/105,International Association for Cryptologic Research,May 2,2004, and Carnegie Mellon Institute for Software Research International, CMU-ISRI-04-116,2004. http://www.heinz.cmu.edu/~acquisti/papers/acquisti-electronic_voting.pdf.

[12] Jonathan Goulet,Jeffrey Zitelli. "Surveying and Improving Electronic Voting Schemes". http://www.seas.upenn.edu/~cse400/CSE400_2004_2005/senior_design_projects_04_05.htm

[13] Ari Juels , Markus Jakobsson. "Coercion-resistant electronic elections," 2002. http://www.vote-auction.net/VOTEAUCTION/165.pdf

[14] Martin Hirt. Multi-party computation: "Efficient protocols, general adversaries, and voting," *PhD Thesis, ETH Zurich*,2001.

[15] Ari Juels, Dario Catalano, Markus Jakobsson: Coercion- resistant electronic elections,"One older version was at Cryptology ePrint Archive,"Report 2002/165. http://eprint.iacr.org/; latest version on Juels web page: http://www.rsasecurity.com/rsalabs/node.asp?id=2030 as of June 2005.

[16] Josh C. Benaloh. "Verifiable secret-ballot elections," *PhD Thesis, Yale University, Department of Computer Science*, 1987. Number 561.

[17] Kazue Sako and Joe Kilian. "Secure voting using partial compatible homomorphisms," In *the proceedings of CRYPTO'94*, Springer-Verlag, LNCS 839, 1994: 248–259.

[18] David Chaum. "Secret-ballot receipts and transparent Integrity," Draft, 2002. www.vreceipt.com/article.pdf.

[19] Tatsuaki Okamoto. "Receipt-free electronic voting schemes for large scale elections," In *the proceeding of Security Protocols Workshop*, Springer-Verlag, LNCS 1361, 1997: 25–35.

[20] David Chaum. "Elections with unconditionally- secret ballots and disruption equivalent to breaking rsa," *In the proceeding of EUROCRYPT* '98, Springer-Verlag, LNCS 330, 1988: 177–182.

[21] Atshushi Fujioka, Tatsuaki Okamoto, and Kazuo Ohta. "a practical secret voting scheme for large scale elections," In *the proceeding of Auscrypt* '92, Springer-Verlag, LNCS 718, 1992. 244–251.

[22] Michael J. Radwin, "An untraceable, universally verifiable voting scheme," http://www.radwin.org/michael/projects/voting.html

[23] Wen-Sheng Juang, Chin-Laung Lei, Pei-Ling Yu. "A verifiable multi-authorities secret elections allowing abstaining from voting," *International Computer Symposium*, Tainan, Taiwan, 1998.

[24] Patrick Horster, Markus Michels, Holger Petersen. "Blind multisignature schemes and their relevance to electronic voting,".In *the proceeding of 11th Annual Computer Security Applications Conference*. IEEE Press, 1995: 149–156.

[25] Lorrie Cranor and Ron Cytron. Sensus: "A security-conscious electronic polling system for the

Internet," In *the proceedings of the Hawaii Conference on System Sciences*, 1997.

[26] Miyako Ohkubo, Fumiaki Miura, Masayuki Abe, Atsushi Fujioka, and Tatsuaki Okamoto. "An improvement on a practical secret voting scheme," In *the proceedings of ISW* '99,1999,pp.225–234.

[27] Kazue Sako , Joe Kilian. "Receipt-free mix-type voting scheme," In *the proceeding of EUROCRYPT* '95, Springer-Verlag, LNCS 921,1995,pp.393–403.

[28] Choonsik Park, Kazutomo Itoh, Kaoru Kurosawa. "Efficient anonymous channel and all/nothing election scheme," In *the proceeding of Advances Cryptology - EUROCRYPT*'93, Springer-Verlag, 1993: 248–259.

[29] Markus Jakobsson, Ari Juels, and Ronald L. Rivest. "Making mix nets robust for electronic voting by randomized partial checking," In *the proceeding of USENIX* '02,2002,pp.339–353.

[30] Emmanouil Magkos, Mike Burmester, and Vassilios Chrissikopoulos. "Receipt-freeness in large-scale elections without untappable channels," In *the proceeding of I3E*,2001,pp.683–694.

[31] Birgit Pfitzmann. "Breaking an efficient anonymous Channel," In *the proceedings of EUROCRYPT* '94. Springer-Verlag,LNCS 950,1995,pp.332–340.

[32] Markus Michels , Patrick Horster. "Some remarks on a receipt-free and universally verifiable mix-type voting scheme," In *the proceedings of ASIACRYPT* '94, Springer-Verlag, LNCS 1163,1996,pp.125–132.

[33] Masayuki Abe. "Universally verifiable mix-net with verification work independent of the number of mix-servers," In *the proceedings of EUROCRYPT* '98,. Springer-Verlag, LNCS 1403,1998,pp.437–447.

[34] Youngcheon Lee , Kwangjo Kim. "Receipt-free electronic voting through collaboration of voter and honest verifier," 2000. hhtp://citeseer.nj.nec.com/lee00receiptfree.html.

[35] Andrew Neff. "Detecting malicious poll site voting Clients," 2003. http://votehere.com/vhti/documentation/psclients.pdf.

[36] David chaum, Secret-Ballot Receipts:"True Voter-Verifiable Elections," *IEEE security and privacy*, January-February 2004 (Vol. 2, No. 1): 38-47.

[37] Ran Canetti, Cynthia Dwork, Moni Naor, Rafail Ostrovsky,"Deniable encryption," In *the proceeding of CRYPTO* '97,Springer-Verlag, LNCS 1294, 1997,pp.90–104.

[38] Zuzana Rja¡¦skov´a, Electronic Voting Schemes, master thesies,"Department of Computer Science Faculty of Mathematics," *Physics and Informatics Comenius University*,Bratislava,Apr 2002.

[39] Aggelos Kiayias, Moti Yung. "The vector-ballot e-voting approach," http://theory.lcs.mit.edu/~rivest/voting/papers/KiayiasYung-TheVectorBallotEVotingApproach.pdf

Bo Meng is a Full Associate Professor of School of Computer, South-Center University for Nationalities. He got his doctor degree from Wuhan University of Technology in 2003; from 2004 to 2006, he work in Wuhan University as a postdoc. He Researches on Information Security,electronic Commerce. Electronic Voting.

Technical Criterion and Model of Electronic Data Exchange and Share Based on XML Technologies *

Yefu Wu, Wei Zhong, Dingfang Chen
Wuhan University of Technology, Wuhan, 430063, P. R. China
Email: wuyefu@whut.edu.cn

ABSTRACT

Currently, many enterprises or government departments have owned some independent application systems, which results in the phenomena of “information isolated islands”. So constructing a platform of data exchange and share for all the application systems is a critical task. It is well known that it is necessary to constitute some technical standards or criterions to guide the design of the platform. In the paper, we analyze the technical needs of data exchange and share platform, and then brings forward an XML-based and message-based model of data exchange and share which is based on XML and message. The model provides an open, flexible technology solution for data exchange and share. In the end, we introduce the mechanism of data exchange and share, the type and structure of message, and how to implement security of message.

Keywords: Data Exchange, Data Share, XML, Message, Adapter, Heterogeneous Data Integration

1. THE PROBLEMS ON ELECTRONIC DATA EXCHANGE AND SHARE

With the development of information technology, more and more information systems are applied in enterprises or government departments. But many systems are running independently, so the information isolated islands appear. Furthermore, because more and more new systems will be put into use, the data integration between new systems and existing ones needs to be taken into account in advance. Presently, the urgent need for them is constructing a platform of data exchange and share for all application systems to exchange data each other.

From the exchanged message format, there are three ways for electronic data exchange conventionally. Firstly, the exchanged message is text file. Secondly, it is database file. Finally, it is EDI message [1].

The first way requires that the text format accords with the agreement between data’s sender and receiver strictly to understand the data, which results in increasing the length of programming codes and restricting the system’s expandability. The second way requires that the data’s receiver can understand the structure of the data table. But the data’s receiver almost could not do it. The third way is exchanging EDI message which needs to develop a series of special EDI messages and has a huge investment.

The advent of the XML provides a new and open approach to electronic data exchange. It will be proved that XML and related technologies are the most excellent technologies for electronic data exchange and data [2].

The advantages of applying XML and related technologies in the process of electronic data exchange are given here [3]:

(1) XML can describe the semantics of exchanged data;
(2) The low-cost software design is available, because XML format is open and standardized;
(3) XML technologies can encrypt a part of information in XML document, which increases the speed of data encryption and decryption greatly;
(4) XML technologies support the SOAP protocol, Web Services and Grid Services.

According to the basic needs of data exchange and share, we analyze the technical requirements of the data exchange and share platform, and then put forward a new technical criterion of data exchange and share. The criterion is based on XML standard, and supports both synchronous and asynchronous operations. The data transmission adopts message-based mechanism. The format of exchanged data follows XML standards. The protocols of network transportation are SOAP and HTTP/HTTPS/FTP /SMTP. All of the technologies above ensure that the criterion is open and standardized [3].

2. THE REQUIREMENTS OF ELECTRONIC DATA EXCHANGE AND SHARE

The technical requirements of data exchange and share are given here:

(1) Using advanced and scientific technologies. The technologies should have a long life at present and even in the future
(2) Adopting open standards. The methods and protocols of data exchange and share should accord with a series of international standard. By adopting the open standards, the complexity of the data exchange and share is decreased.
(3) Developing software with low-cost and using software in flexible way. Clients should use the platform with the method of asynchronous or synchronous to send data. The platform should provide the mode of "push" or "pull” data and support visual modeling tools to define the relationship of the exchanged and shared data.
(4) Security and reliability. The platform should provide the mechanisms of digital signs, digital certificate, access audit and so forth.
(5) Across-platform running. The platform should support multi-platform, multi-OS, the heterogeneous database and different application system [4].

The functions of the platform are as follows:

(1) Standardizing the exchanged data. The data is described as open and normative XML-based document.
(2) Exchanging data among different systems. The function

* The paper has been supported by the following projects:
(1)The Key Technologies R & D Program of Hubei Province: Universal data exchange technology research in electronic commerce (2001AA105B05), P. R. China.
(2)The Key Projects of Wuhan Communication Committee: The standards and criterions for Wuhan transportation informatization, P. R. China.

accomplishes the data exchange between data exchange center and different clients.

(3) Integrating data. The basic data of clients can be integrated into the data center.

(4) Sharing data. Integrated data can be stored in the data center and be accessed by clients.

(5) User interface. The function provides some services for the clients' application system.

(6) Charging clients for using the center. These fees include the fee of registering clients, exchanging data and accessing shared data.

(7) Security and audit. The platform can accomplish include the data encryption and decryption certificate authority and operation audit.

(8) Platform Management. The platform needs to carry out some administration day-to-day [5].

3. A CRITERION OF XML-BASED ELECTRONIC DATA EXCHANGE AND SHARE

This criterion is abbreviated to EDISML. Referring the EDI and ebXML standards, we present the contents of EDISML. EDISML includes the following five parts [6]:

(1) Technical Architecture of Data Exchange and Share. It specifies the basic framework of EDISML. And it is the basis of other technical specifications. It provides the semantic framework for data exchange and share. It also provides the mechanism of discovering clients, consulting with each other and carrying out the activities of exchanging and sharing data among the clients.

(2) Message Architecture. It specifies the format of the exchanged data and the rules of packaging data. It provides a series of specifications about data format, and rules of sending and receiving messages. It also specifies the security regulation for the data center and clients.

(3) Registry Service Center. It specifies the interactive model between clients and the data exchange center. The registry service center provides the registration index for the information of client. The indexed information is stored in the exchange center database. The center also saves database pattern and database mapping rules for every client.

(4) Encoding or Decoding Rules. It specifies the rules of encoding or decoding exchanged data which is related to the center and all clients' systems, such as time encoding rules, ship encoding rules, and so on.

(5) Platform Actualization Rules. It specifies some basic guide lines of selecting hardware and software platforms, and the requirements of the toolkits of designing application systems.

4. THE TECHNICAL ARCHITECTURE OF XML-BASED DATA EXCHANGE AND SHARE

Message-based communication is the core of the technical architecture. Firstly data is presented in XML format, and then it is encapsulated into message with SOAP. At last, it is transferred to receiver with the help of HTTP, SMTP or FTP. The platform of data exchange and share is a 4-layer structure, as is shown in the Fig.1.

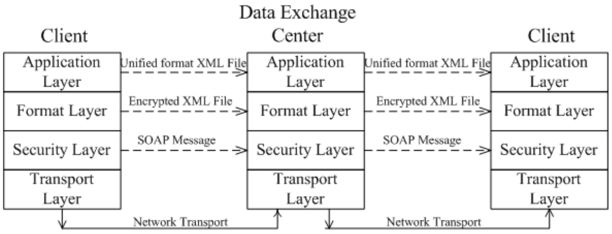


Fig.1. The 4-layer structure of data exchange platform

The architecture consists of three divisions. They are Data Exchange Center, Data Share Center, and Client-Adapter. The blue print of the architecture is shown in Fig. 2.

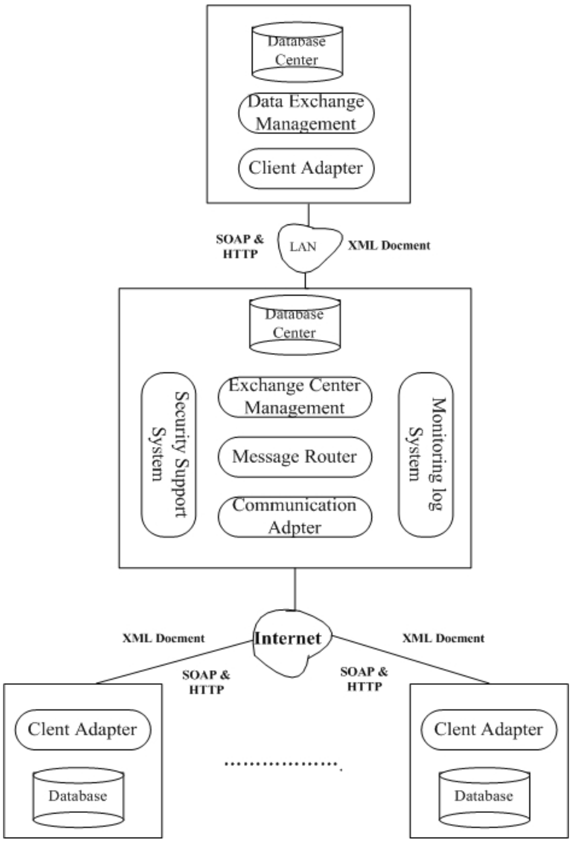


Fig.2. The architecture of data exchange and share platform

4.1 Data Exchange Center

Data exchange center is the core of the data exchange and share platform. It carries out sending, receiving analyzing and routing messages. It also supports the security of exchanged messages. The working mechanism of the data exchange center is expressed in Fig.3.

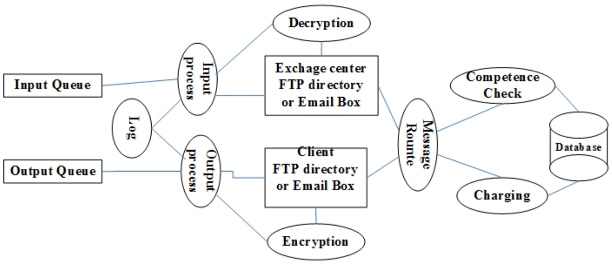


Fig. 3. The working mechanism of data exchange center

The center provides a FTP directory or Email box for itself and. every client. The client's FTP directory or Email box temporarily stores the XML-based document which will be sent the client. The center's FTP directory or Email box stores all received XML documents temporarily which are from all of clients. The center consists of five modules which are given as follows.

Communication Adapter: The communication adapter mainly connects the center with client-adapters, and achieves the exchange of messages between the center and clients. It includes several function modules, such as receiving message, sending message, encrypting message and decrypting message.

Message Router: The message router is a key part of the data exchange center. It analyzes the electronic data or messages which are from the client, and then transmits it to the exchange center module to carry out auditing privilege of accessing to data and charging clients for using the data. Message can de divided into four categories: Data Message, Requesting Data Message, system Message and Requesting Downloading Message. According to the type of the message, some jobs are done by the message router.

(1) If the message is either of Data Message, Requesting Data Message and system Message, the message router should be distributed it to the corresponding client's FTP directory or Email box.

(2) If the message is Requesting Downloading Message, the message router will analyze it, and then all messages in the corresponding client's FTP directory or Email box should be sent to the output queue by the router, and then to be sent to the client-adapter by communication adapter.

Exchange Center Management: The module manages the shared data in the center, and the clients' privileges of access to the shared data. It also includes the function of charging and registering clients. A lot of tables are created to save the shared information, the roles' privileges, the charge information, and the clients' registration information.

Monitoring Log System: The module monitors and controls each process of the platform, such as lifecycle of process, state of process, exception of process and so forth.

Security Support System: In data exchange center, a department-level CA authentication center will be set up. The security support system is called by message-receiver, message-sender, and message encryption and decryption sub module.

4.2 Clients-Side System

The application system of every department is a clients-side system. The client-adapter mainly handles the communication between the data exchange center and clients. The client's database saves the messages received from the center. The working mechanism of the client-adapter is shown in Fig.4.

The working processes of the client-adapter are given here:

(1) The client-adapter gains a data request message from own FTP directory or Email box, then puts the message into the client-side input buffer queue.

(2) The input process disassembles the message which is encapsulated with SOAP and encrypts the message by calling the corresponding module of the security support system to get an XML document. The data request in XML document is executed by the message process module. So the requested data is generated.

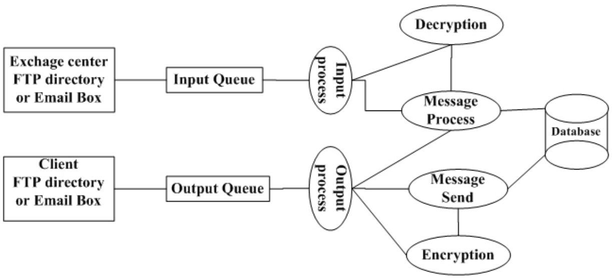


Fig.4. The working mechanism of the client-adapter

(3) The data is formulated by the application of XML standards, and then exported into the output buffer queue.

The module of sending message can send Data Request Messages automatically, and send the Data Messages to other clients or the center directly.

4.3 Data Share Center

The data share center is the data storage center of the data exchange and share platform. All shared data must be stored in the database in the data share center. The shared data includes the business data which comes from the client and the basic code data which are the basic data published by the center. The client's adapter in data share center carries out sending messages to the data exchange center and receiving messages from the data exchange center. The adapter in the center has the same function as the client-adapter of each client.

5. THE PROCESS OF EXCHANGING DATA

Message is the core of the data exchange and share platform. Sending or receiving data is a basic task for the platform. The data (i.e. message) can be transmitted from departments to the data exchange center or from the data exchange center to departments or from one department to another department. The general courses of data exchange between department A and B are presented as follows:

(1) Department A requests the database mode of department B from the data exchange center firstly. The client-adapter of department A generates a query sentence by means of the database mode, and then gets related data automatically, and then encapsulates the data into a message with SOAP. The message will be sent to the data exchange center after the message is encrypted.

(2) After the message has been received in the data exchange center, it is decrypted. Then the client's privileges of accessing to the inquired data are audited. Meanwhile, the exchanged center management charges the client for the request. There are two ways of charging clients. One is according to the flow of bytes, and another is according to the number of messages. After the message is passed, it will be sent to department B if department B is online.

(3) When the client-adapter of department B has received the message, the client-adapter query the department B's database or files according to the inquiry sentence in the message to get the requested data. Then after the data is encrypted and encapsulated, the data will be sent to the data exchange center.

(4) After the data exchange center has received the data message, the message is placed into FTP directory or Email box of department A temporarily, waiting for department A to download.

(5) While department A is online, the client-adapter of department A downloads the message from own FTP

directory or Email box, then processes it further, including saving it in the local diskette or writing it into the related to database's tables.

6. THE STRUCTURE AND SECURITY OF MESSAGE

Message is composed of a sequence of XML elements. We have design the structures of the four kinds of messages.

6.1 System Message

The root element of system message is <systemMessage> element which includes two sub elements: <msgHeader> and <msgBody>. The DTD diagram of system message is shown as Fig.5.

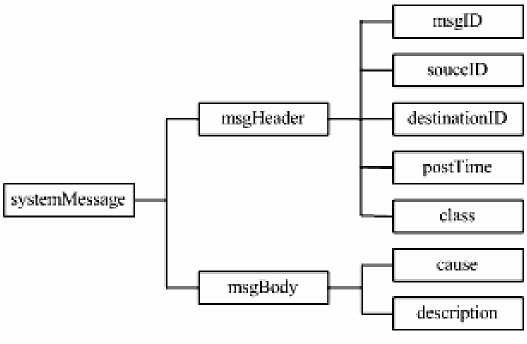


Fig.5. The DTD diagram of the System Message

In the DTD, the <msgHeader> element is the header of the system message, which describes the basic information of the message: msgID (the number of message), sourceID (the node of sending message), destinationID (the node of receiving message), pasTime (the date and time of sending message), and class (the type of message: 1- sending message success, 0-sending message error). The <msgBody> element is the body of the system message, which includes the two elements: <cause> and <description>. The <cause> element describes the season of sending error, and the <description> element describes the content of sending error.

6.2 Requesting Data Message

The root element of this message is <rquestMessage> element which includes two sub elements: <msgHeader> and <msgBody>. The DTD diagram of the message is shown as Fig. 6.

In the DTD, the <msgHeader> element is the same structure as system message. But, the <class> element describes the types of requesting data: 1-data query, 2-files acquirement 3-both. The <msgBody> element is the body of the message, which includes two elements: <datas> and <files>. If <class> element is 1, the <datas> element describes the SQL statements of data query. If <class> element is 2, the <files> element describes the files' name. If <class> element is 3, the <datas> element and <files> element describe the SQL statements and the files' name.

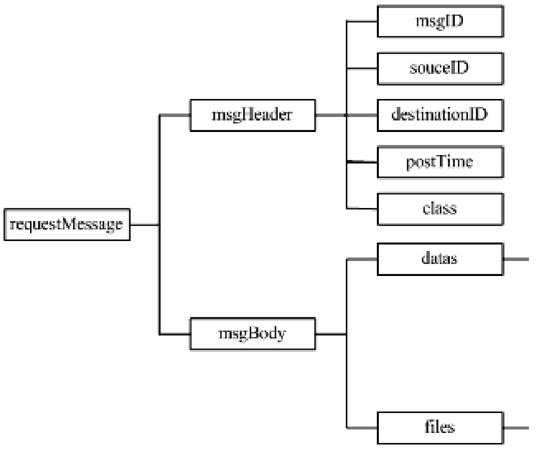


Fig.6. The DTD diagrams of the Requesting Data Message

6.3 Data Message

The root element of this message is <dataMessage> element which includes two sub elements: <msgHeader> and <msgBody>. The DTD of the message is shown as Fig.7.

In the DTD, the <msgHeader> element is the same structure as system message. But, the <class> element describes the type of message: 1-the results of data query, 2- the shared files, and 3- modified data automatically. The <msgBody> element is the body of the message, which includes a sequence of <item> element. The <item> element describes the results of data query. The ID attribute of <item> describes the method of sending data. The Operator attribute of <item> describes the action of processing data: inserting data or deleting data or updating data. The <item> includes two sub elements: <dataAttribute> and <dataValue>。 The <dataAttribute> element describes the table field, and the <dataValue> element describes the field value, which consists of <row> element and <column> element. and <item> element

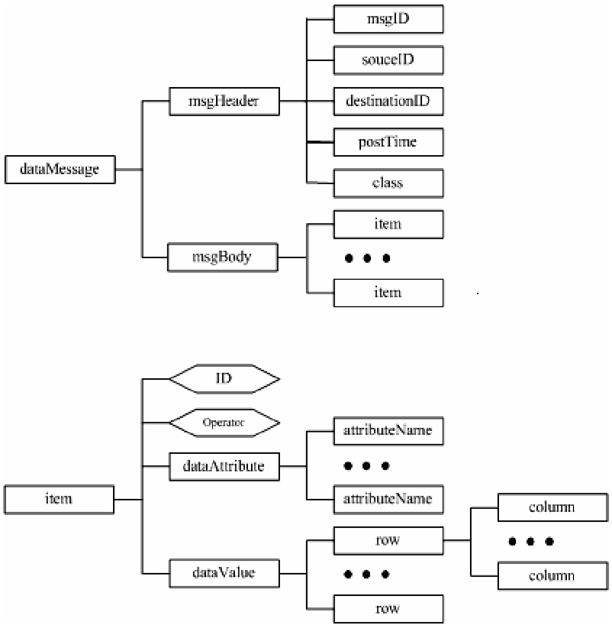


Fig.7. The DTD diagrams of the Requesting Data Message

6.3 Requesting Downloading Message

The root element of system message is <loadMessage> element which includes two sub elements: <msgHeader> and <msgBody>. The <msgHeader> element is the same structure as system message. But, the <class> element is empty. The <msgBody> element is empty.

6.4 Encrypting Messages

The security of data should be ensured because the exchanged data are transmitted on the Internet. In the model, there are two ways of encrypting messages. One is encrypting a part of elements of message. Another is encrypting all of elements of message. The ways are supported by XML standards [7].

The principles of encrypting message are as follows:

(1) The shared key is generated by application of Triple-DES or other arithmetic for the data which is the parts or all of elements in the message. The elements are some sensitive data and are asked to be encrypted.

(2) The parts or all of elements in <msgbody> element is encrypted by application of the shared key.

(3) The shared key is encrypted by application the public key to generate an element, i.e. <EncryptedData> element, which replaces the encrypted elements.

The clients and the center will encrypt messages by application of digital certificate of the enterprise's CA center. It is a good approach to saving the private digital certificate into an USB-KEY device for each of client.

7. CONCLUSIONS

XML is one of the hottest network technologies. XML provides a new and open approach to electronic data exchange. It has proved that XML and Related technologies are the most effectual technologies for electronic data exchange and data integration. Through studying the key technologies of data exchange and XML standard, the paper brings forward a universal model of data exchange and share, which is based on the XML standard and SOAP protocol. The model takes advantage of XML and related technologies. The model is a contributing thought for developing electronic data exchange and share platform.

Of course, some factors may not be covered in the model, such as the process of sending the same message again, the efficiency in the encrypting or decrypting mass data and the security model of transferring data/message. The solutions of all the issues will be studied further in the near future. For example, we are studying how to apply SAML and XACML standards into the model to strengthen security.

ACKNOWLEDGEMENT

The following individuals have contributed to this work:
Chaozong Wu, Kun Xu, Hongtao Liu, Ting Li.

REFERENCES

[1] Ye-fu Wu, Xiao-ling Yuan, "Researching of the Message Security and Identity Authentication of the EDI System", *Mini-Micro Systems*, vol.23, Feb. 2002.

[2] Tianhuang Chen,Qing-mei.Zou,"Information sharing technology between heterogeneous databases based on XML,"*Journal of Wuhan University of Technology*,Feb 2005.

[3] Xiao-ling Yuan,Ye-fu Wu,"Researching of the Unusual e-Commerce Model Based XML Technology about B2B," *Journal of Wuhan University of Technology*, vol.27,Apr 2005.

[4] Zun-chao Li, Ying-qiang Xu, et al, "XML-based secure data exchange between heterogeneous databases,"*The Computer Engineering and Applications*, vol.13, 2005.

[5] OASIS, http://www.oasis-open.org/committees/download.php/17817/ebxmljc-WhitePaper-wd-r02-en.pdf, 24 Apr 2006.

[6] Fan-bo Ye, Ren-zhong Tang, et al, "Data integration platform for small-medium enterprise process integration," *Journal of Zhejiang University*, vol.40, Nov 2006.

[7] Manish verma, "XML Seurity，Part II", http://www-128.ibm.com/developerworks/cn/xml/x-seclay2/index.html,1. Dec 2003.

Yefu Wu is a associate professor within the College of Computer at Wuhan University of Technology with much experiences in teaching and research. He has taught many curriculums, such as XML and Electronic commerce, Object-oriented Programming and Visual Programming, and so on. He graduated from South East University (Nanjing, China) in 1988. Presently he is a Ph.D. candidate. His research interests include electronic commerce and computer network, in which he has done some work deeply, developed some application systems and published many papers. He was awarded a 3rd prize of science and technology advancement in Hubei Province, China in December 2002.

Research on Flexible Workflow Management Based on Web Services

Guangchao Wu, Shu Yu
School of Mathematical Sciences, South China University of Technology
Guangzhou 510640, China
Email: [1]Brisk_su@hotmail.com, [2] Yushu_scut@126.com

ABSTRACT

Aiming at some existing problems in traditional workflow management systems, such as bad flexibility, poor reusability and weak interoperation, this paper proposed a flexible workflow management system based on Web services. By choosing Web services as the management objects of activities or business processes in workflow and using ontology technology to increase the QoWS, the system can allow the dynamic binding and invocation of workflow and further solve the integration and sharing problems of workflow information.

Keywords: Flexible Workflow Management System,Web Service,BPEL4WS, Quality of Web Service,Ontology

1. INTRODUCTION

With the rapid development of network and computer technology, workflow systems become the cross-enterprise and cross-platform applications from the inner of the enterprises. Traditional WfMS (Workflow Management Systems) can't meet the requirements of enterprises' development anymore. Many enterprises have their own application systems, such as SCM (Supply Chain Management), CRM (Client Relationship Management) and ERP (Enterprise Resource Plan), and these application systems are often changed, which as a result causes the error or interruption when running them as business processes in a workflow. In order to enable the enterprises in different places can finish their cooperation tasks more fluently, some workflow management mechanism with good flexibility and high self-adaptability is urgent. It is needed to solve the poor interoperation problems and integrate the workflow processes in different enterprises or application systems in each associated enterprise.

The characteristics of high openness and cross-platform make the Web services-based WfMS become an inevitable trend. So, according to some existing problems in current workflow management researches, this paper proposes a flexible WfMS based Web services. By incorporating Web services and ontology technology, the system can allow the dynamic service binding and invocation between workflow and Web services during the execution time of activity instances, and shield the changes occurred inside the enterprises. In this way, the business processes in each enterprise can run normally and the flexibility and self-adaptability of workflow can be greatly improved.

2. RELATED WORK

2.1 Tradition Workflow Management

Traditional WfMS pay more attention to modeling simple and static process, and required all information in the processes should be static. They can't model those dynamic processes and can't support the heterogeneous distributed environment. But many current business processes in most of the enterprises include a lot of geographic discrete execution units. In addition, traditional workflow models mainly focus on those processes within enterprises and rarely consider how to support the cross- enterprise workflow. Consequently, users in traditional WfMS can only call the applications inside the enterprise and is unable to invoke other cross-enterprise or cross-platform applications. All workflow information can't be shared and integrated in time.

2.2 Current Related Researches

Currently, some scholars become to propose some Web-based WfMS. Authors in [1] use Browser/Server architecture and middleware technology to support the distributed workflow execution, and promote the cooperation among many Web-based workflow servers. But it is unpractical because it requires all workflow servers should be based on Web and lacks of reusability. It can't completely allow the dynamic binding and invocation of processes and is unable to permit the seamless information exchange and cooperation. Authors in [2] use XML-PDL (XML Process Definition Language) as the process definition language and employ J2EE architecture to build the system, but it also lack of dynamic process binding and invocation ability.

As a new application integration technology, Web service uses XML document to describe and query Web services and return the results. It offers a dynamic integration solution to enable all services use UDDI (Universal Description, Discovery and Integration) standard to find, bind and call the services dynamically. Its characteristics of loose coupling, reusability and cross-platform make it one of the best choices to solve the problems in traditional or Web-based workflow systems. So, this paper tries to use Web services to implement activities or business processes in workflow and also employ BPEL4WS (Business Process Execution Language for Web Services) to integrate the businesses processes with different process definition languages, which can finally allow the cooperation and integration among the WfMS that cross enterprises or platforms in Internet.

3. FLEXIBLE WORKFLOW MANAGEMENT SYSTEM BASED ON WEB SERVICES

Web services-based flexible WfMS can improve the performance of workflow system in the following four aspects:

(1) Using Web services to implement the activities in workflow. In traditional WfMS, workflow can only invoke the applications inside enterprises and can't interact with other cross-platform or cross-enterprise applications. But in Web services-based WfMS, many activities can be directly carried out by the Web services providing similar or the same functions in Internet, in stead of being developed repeatedly.

(2) Improving the flexibility and self-adaptability. Using Web services to implement business processes can offer good dynamic ability and louse-coupling ability. When a

user defines a process, he isn't necessary to bind some specific application immediately, but only need to declare the correspondent interface types of the Web services. So long as the interface types of other Web services in Internet meet the requirements, they can be dynamically chosen to complete the process at any moment. In this way, the enterprises can select the most suitable cooperative companies at any time according to their own needs.

(3) Improving the reusability and interoperation ability. Deploying processes in workflow as Web services and registering them in UDDI not only can meet their own needs, but also can allow the applications in other enterprises to access them. As a result, the integration and recursive definition of process can be achieved and the functions of WfMS become more powerful. Furthermore, the interaction means between cooperative parts are more flexible and the trade costs can be decreased greatly.

(4) Improving the sharing and integration ability of information. In traditional WfMS, information can only be used within enterprise and can't be shared with other workflow systems, not to mention integration. Consequently, many information islands are produced. However, because the nature of Web service is an application integration technology, which uses XML document to exchange information, it can provide very good interoperation ability. So, using Web service to implement cross-organization and cross-platform workflow can take full advantages of Web service and also can solve the information sharing and integration problems.

3.1 Architecture

Because Web service is based on SOA (Service Oriented Architecture), so the WfMS based on Web services should fully take this feature into account. Besides making full use of the advantages of Web service, we should also consider the present situations of enterprises and try to cut down their costs and spending. The flexible WfMS based on Web services is shown in Fig. 1.

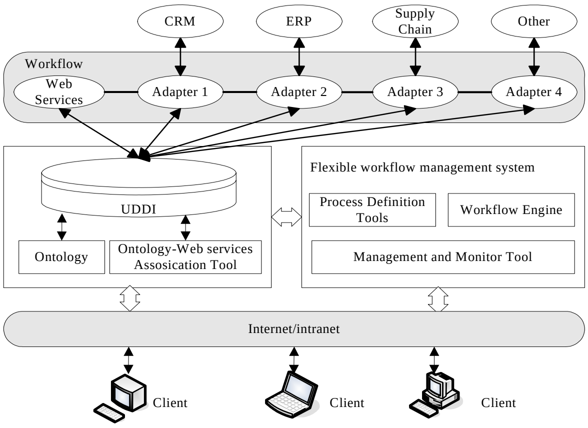


Fig.1. Architecture of flexible WfMS based on web service

In Fig. 1, Workflow server is composed of process definition tool, workflow engine and management and monitor tool. The process definition tool mainly takes charge of process modeling on enterprise workflows which integrate various kinds of Web services, and use BPEL4WS[4] as the formal specification of business processes and business interaction protocols to enable the interoperation of WFMS in different enterprises. Workflow engine is responsible for how to run the business processes according to the pre-defined process model and how to interact with UDDI. Workflow management and monitor tool watches and supervises the execution instances and status of workflow, such as user and role management. While the workflow design tool cooperates with ontology-Web services association tool and creates process ontology and relevant Web services. The outputs are depicted in the form BPEL4WS and become the input of transaction processes and workflow engine.

For the Web service adapter, it is mainly adopted to make full use of those original business application systems in enterprises, which are left due to the historical reasons. For many enterprises, it is unpractical and impossible to totally discard all current business application systems. So, they can use Web service adapters to package and encapsulate these systems in the form of Web services and allow the invocation bye using SOAP (Simple Object Access Protocol) through Internet.

3.2 Workflow Model

The process meta-model defined by Workflow Management Coalition includes six basic entities[5]: workflow type definition, activity, transition conditions, workflow relevant data, role and invoked application. The basic process definition meta-model is shown in Fig.2.

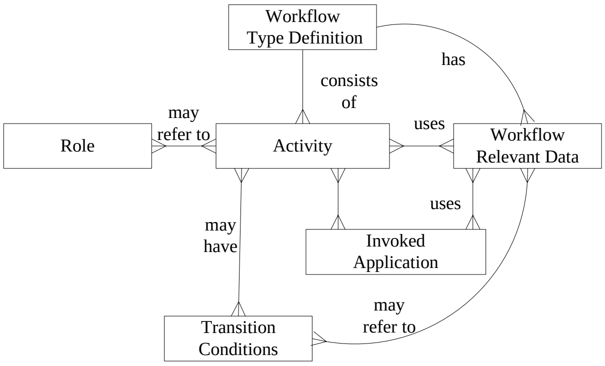


Fig.2. Basic process definition Meta-Model

However, many enterprises often use different modeling tools to express workflow processes because of their own requirements. As a result, many WFMS can't communicate with each other due to the different description about workflow processes. To solve this problem, it is necessary to use united process definition language.

At the present time, the most frequent discussed script language used to describe processes includes Microsoft's XLANG and IBM' WSFL. XLANG supports the graph-oriented process and is able to describe many kinds of process models. WSFL allows the structured construction of processes and uses a directed graph to define and execute these business processes. BPEL4WS combines these two standards and becomes a Web service integration language that supports various kinds of business processes through some very natural means. In BPEL4WS, WSDL (Web Service Description Language) is not only used to describe process and its service, but also used to integrate a group of existing services and define them as a new integrated service. The interfaces of the whole service are also described as WSDL.

In order to allow the communication and interoperation among business processes, it is important to describe them using a common modeling language. To accomplish that, it is necessary to convert various process definition languages in different enterprises into BPEL4WS format and make it as the input of work engine. The conversion process of different process definition languages in our system is shown as follows:

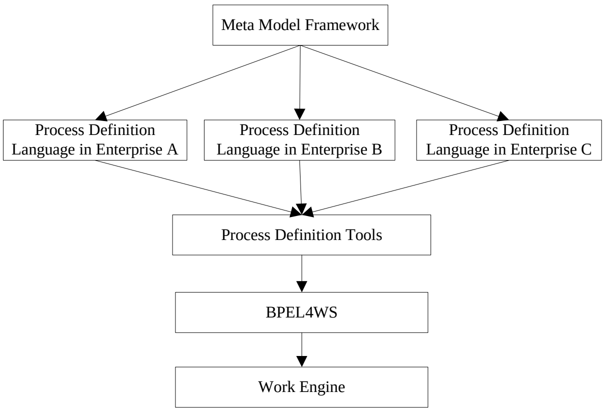

Meta Model Framework
Process Definition Language in Enterprise A
Process Definition Language in Enterprise B
Process Definition Language in Enterprise C
Process Definition Tools
BPEL4WS
Work Engine

Fig.3. The Conversion of Different Process Definition languages

3.3 Workflow Engine Design

The major functions of workflow engine includes: explaining process definition, debugging the process execution, creating and managing the process instance execution, debugging the activity execution and creating the work item, maintaining workflow control data, relevant data and users' work list[3]. The structure of workflow engine is shown as follows:

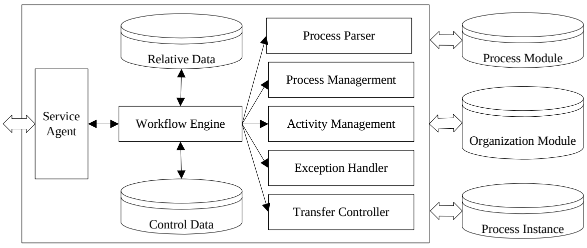


Fig.4. Structure of Workflow Engine

In Fig.4, service agent mainly takes charge of all interaction events with UDDI. It will search UDDI registration center to decide which service will be most suitable one according to the restriction conditions on activity in workflow.

3.4 Web Service Invocation

The process of invoking outer Web services by workflow can be divided into two phases. At the beginning of modeling phase, the user first declares the port type of the Web service. During the execution phase, if some activity needs to invoke a outer Web service, workflow engine will choose a specific implement of the port type, which points to concrete ports and operations that have only one access point.

In tradition WfMS, activities are always pre-bound to a concrete module and workflow can't replace this module dynamically during the execution time. While in the Web services-based WfMS, the user doesn't need to point out the executor of an activity when defining a workflow. He just needs to declare the public interfaces of the operations and his requirements. During the workflow's execution time, the workflow engine will search UDDI registration center and find out a series of candidate Web services that meet the requirements, and decide which one will be used to fulfill the business process. When choosing the Web service, some QoWS (quality of Web service), such as execution time, dependability or reputation, can be imposed.

The interaction process among workflow, UDDI and Web services in WfMS is shown in Fig. 5.

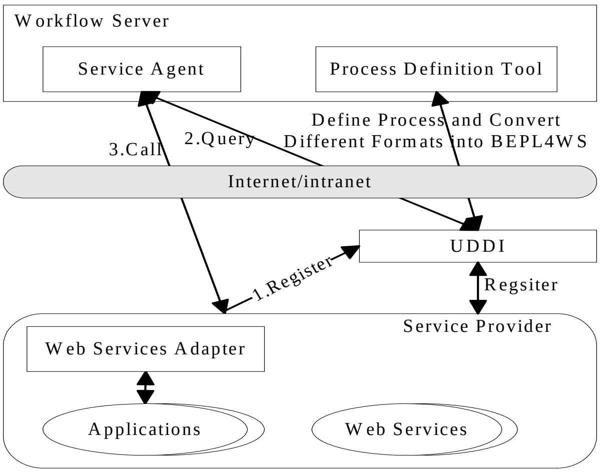


Fig.5. Interaction among workflow server, UDDI and Web services

3.5 Ontology Technology

With the explosive growth of the number of Web services, users in WfMS not only care whether the service can provide the needed functions or not, but also pay more attention to the QoWS. The QoWS is critical for the execution results and can improve the stability of workflow. However, most of the current Web service description languages are mainly based on phraseological description. For example, WSDL describes the basic properties, such as message format, data type, operation, protocol binding and service address, as well as how to interact with Web services. But it doesn't have the language description ability and can't show the semantic meanings of Web service. The tModel in UDDI also only provides a markup mechanism and service lookup can only base on string matching technology, such as service name or other pre-defined fields.

So, how to add QoWS description information about Web service and apply it in WfMS becomes a very important problem. To address this problem, we introduce ontology technology into the WfMS and UDDI to add the semantic information about the services. The ontology of Web service can definitely show the meanings of the service itself and make it possible to understand the service at the semantic level. In this way, when the business logic is modified, workflow can change accordingly and won't be affected by the change of the system.

4. CONCLUSIONS

In this paper, we first propose a Web services-based flexible WfMS and present the system architecture. Then, we analyze several key technologies, including workflow model, workflow engine, Web service invocation and ontology and demonstrate how to use Web services to implement the activities and business processes in traditional workflow management. We also discuss how to use ontology technology to add the semantic meanings to the services in order to improve the flexibility and self-adaptability of workflow system.

REFERENCES

[1] Zh.Piefa,H.Kaihu,Zh.Jinfeng,Q.Liangchun,"Research on enterprise workflow manager system based on Web [J]," *Development & Innovation of Machinery & Electrical Products*,2006,19(5),pp.31-33.

[2] Zh.Hongshan,"A design of workflow management system based on Web [J],"*Journal of Capital Normal University (Natural Science Edition)*, 2006, 27(2):12-14.

[3] L.Hongxin,F.Yushun,"Web service-based integration and interoperation of heterogeneous workflow management system [J],"*Information and Control*,2003, 32(3),pp.193-197.

[4] IBM,Web services:BPEL4WS Theme. http://www.ibm.com/developerworks/cn/webservices/ws-theme/ws-bpel.html,Jun 2003.

[5] WfMC. The Workflow Reference Model [R]. (WfMC-TC00-1003), technical report,Workflow Management Coalition, Hamnshire,1995.

[6] Ivan Mecar,Alisa Devlic,"Agent-oriented semantic discovery and matchmaking of Web services [A]", *In:8th International conference on telecommunications-ConTEL*,2005,pp.603-607.

The Design and Implementation of VoiceXML-based Voice-Driven Email Client

Qing Yang, Gen Feng
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, 430063, China
Email: qingyang@whut.edu.cn

ABSTRACT

Voice-driven email client integrates voice technology and voiceXML-based voice browsing technology, extends traditional email service to PSTN (Public Switched Telephone Network), telephone is thus becoming a new kind of terminal and “visible” email is being turned into “audible” email. Voice-driven email client facilitates users to receive and send emails by telephone, and is able to carry out the effective and efficient authentication for user’s identity. Voice gateway is the core for voice email service, the architecture and working flow of voiceXML-based voice gateway is introduced; the design and implementation of three key modules in voice-driven email client are in detailed discussed.

Keywords: Voice browsing technology, Voice technology, VoiceXML, Voice-driven email service

1. INTRODUCTION

VoiceXML[6](Voice Extensible Markup Language) was proposed by four leading companies in 1999 for developing voice browsing technology. W3C accepted and approved it as a recommendation in 2000 in order to promote the development of voice browsing on the Web. VoiceXML is defined on the basis of XML and provides the facility for controlling the voice interaction process. VoiceXML has got widely support from computer industry and communication industry, and is thus becoming the key technology for voice browsing[2] on the Web.

Recently, with the development of voice technology, telephone is becoming another kind of information terminal on the Web. Email is the most popular communication way on the Web, and some people hope to manage their mailboxes and send/receive email by telephone. The combination of voice technology and traditional email technology and voiceXML based voice browsing technology makes the requirement possible.

This paper firstly discusses the model for voice-driven email server, and then proposes the architecture of voiceXML-based voice-driven email client, at last clarifies the implementation of the key modules.

2. THE MODEL FOR VOICE-DRIVEN EMAIL SERVICE

Voice-Driven email is compatible with traditional way, and extends email service from Internet to PSTN.

PSTN[7](Public switched telephone network) connects to Internet through voice gateway[4], and voice gateway becomes information processing middleware between PSTN and Internet. Fig.1 shows a model for voice-driven email service.

Voice gateway mainly has two components: voice server and document server. Document server (i.e. web server) is mainly used to store VoiceXML script, and carry out business logic such as database operation, and receive or send email as Pops/SMTP proxy. Voice server provides the bidirectional .connections between PSTN and Internet, and mainly includes VoiceXML parser[2] and voice processor. Voice processor performs two kinds of operation, they are: Speech synthesis or Text To Speech[1] (TTS), and Automatic Speech Recognition (ASR). VoiceXML parser performs the business logic by following the working flow specified in a specific VoiceXML script.

According to VoiceXML, <record> mark could be used to store the speech in memory in the forms of “audio/wav”, “audio/basic”, and so on. So it is not difficult to record the speech in audio file and attach audio file to the email. Through such a way, “visible” email is being turned into “audible” email, and it is feasible for users to get information service by using voice-driven email client.

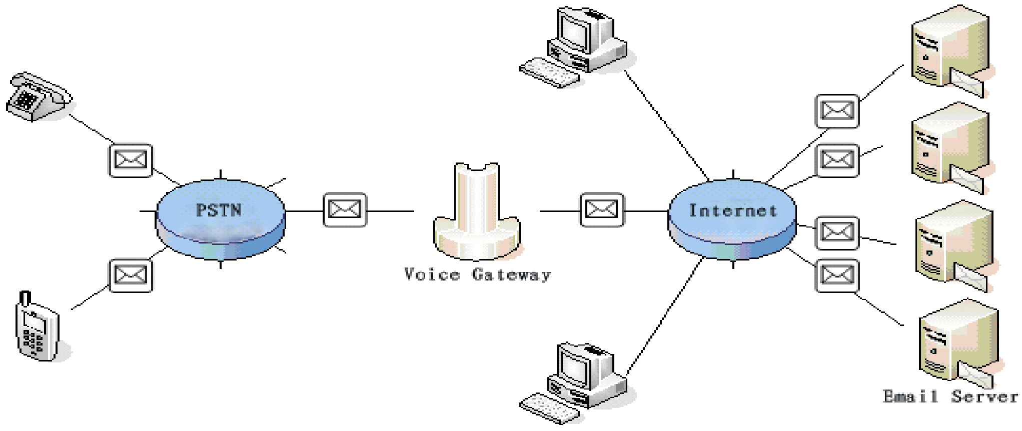


Fig.1. voice-driven email service model

3. THE ARCHITECTURE OF VOICE GATEWAY

Fig.2 shows the architecture of voice gateway. Voice-Driven email client is usually deployed in document server and consists of identification module, receiving/sending module and operating module. Identification module collects all required authentication information and submits it to SMTP[5]/POP3[3] remote proxy; Receiving and sending module are the local proxy and used for receiving or sending voice emails attached with audio file. Operating module manages deleting, moving, copying and marking junk and so on.

E-mail server is actually the SMTP/POP3 remote proxy, receives and sends emails by following SMTP/POP3 protocol after authentication. As the SMTP local proxy, sending module is used for dynamically synthesizing voice emails and sending the voice emails to SMTP remote proxy; As the POP3 local proxy, receiving module is mainly used to receive voice emails from POP3 remote proxy, parse them to audio and then play audio to phone terminal.

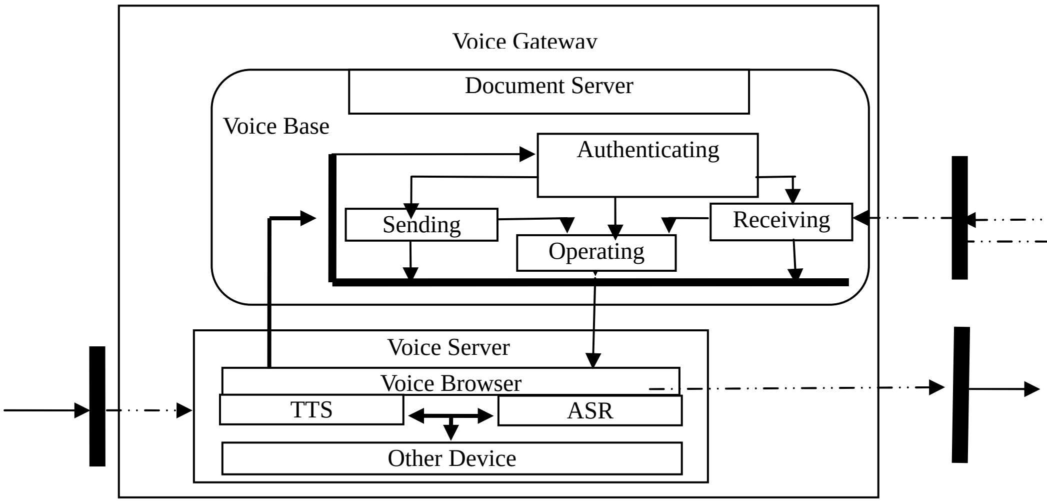


Fig.2. Structure of voice gateway voice-driven email gateway

Voice base mainly stores voiceXML script and audio files which is used to decorate the VUI (Voice User Interface). VoiceXML script defines VUI and controls the voice interaction process when receiving or sending voice email. Audio files is the complement of voiceXML script and helps making the voice interaction process more friendly and efficiently.

All business logic is processed in document server, and processing result is dynamically synthesized and then submitted to voice browser[2] in the form of voiceXML document. Voice browser parses voiceXML document and then play some required voice information to users.

The purpose of dynamic synthesis is to embed the processing result into voiceXML document. Fig.3 shows the working flow of voice-driven email client.

4. THE IMPLEMENTATION OF VOICE-DRIVEN EMAIL CLIENT

Three key modules in voice-driven email client will be in detailed discussed below.

4.1 Identification Module

Authentication information is required by most SMTP/POP3 remote proxy before sending and receiving emails. The function of identification module is to gather the authentication information through the telephone, and to submit it to SMTP/POP3 remote proxy.

The voice-driven email client provides two ways for identification, one is the traditional account/password identification, and another is the voice characteristic identification. The identification result (including account and password) will be submitted to SMTP/POP3 remote proxy. Fig.4 shows the working flow and deployment structure of identification module.

The primitive authentication information is different from the digit authentication information. Primitive account and password is a string with irregular structure, voice recognizer is hard to carry out the speech analysis on it. However, digit account and password (usually in form of a string of digit letter) is much easer to be recognized, so identification module takes digit authentication information as the input.

Primitive authentication information is required by SMTP/POP3 remote proxy (e.g. fenggen2003@sohu.com as primitive account). Digit authentication information will be mapped to primitive authentication information. For example, “58974569” (Digit authentication information) is mapped to fenggen2003@sohu.com (primitive authentication information).

Another identification way (i.e. voice characteristic identification), firstly analyzes the user’s voice, and extract voice characteristic code , then save them in the characteristic storehouse which is deployed in voice gateway. When user registers by voice, identification module will confirm this user’s identity by matching voice characteristic with the voice input, and then get primitive authentication information.

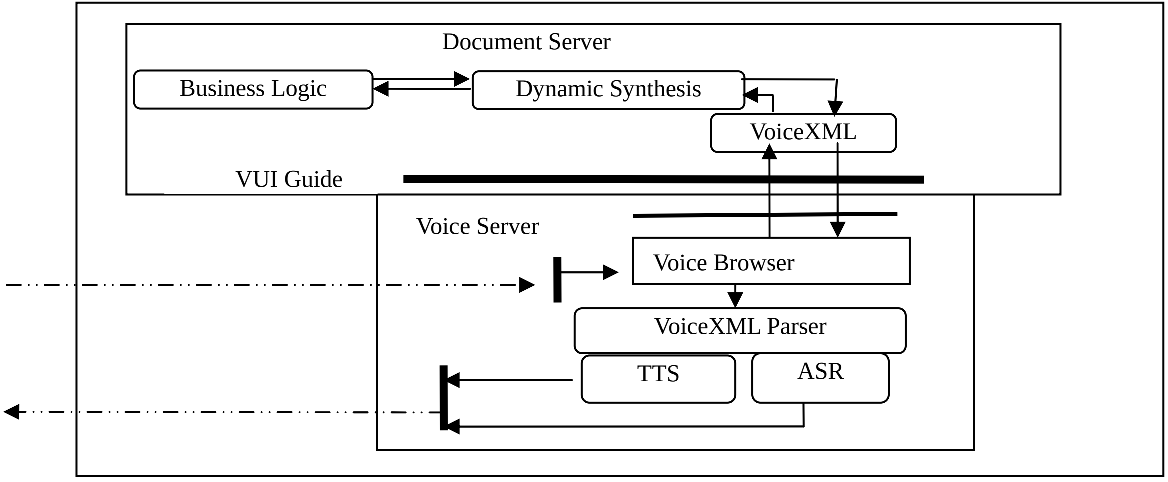


Fig.3. working flow of voice-driven email client.

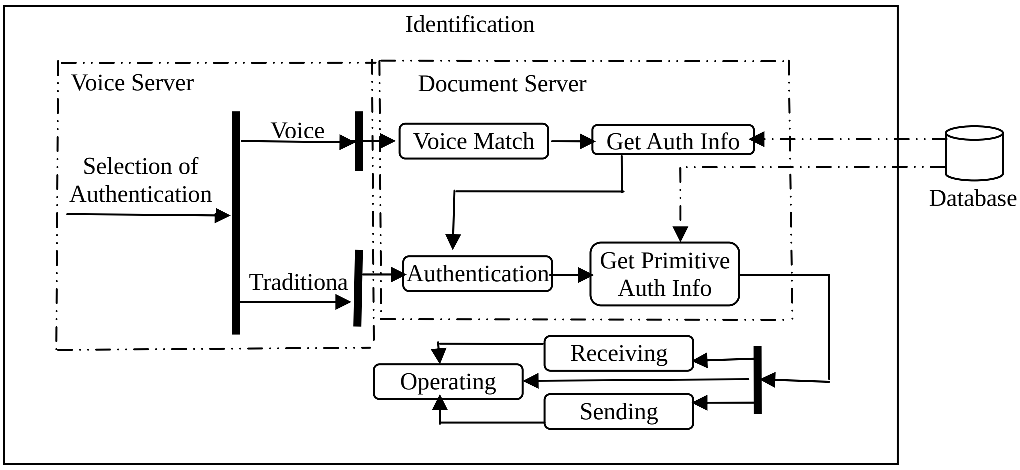


Fig.4. The working flow of identification module

4.2 Voice Email Sending Module

Sending module consists of three main parts: synthesis of voice email, sending out email and saving email to local host. Fig.5 shows the working flow of sending module.

There are three steps during the synthesis of voice email: asking user to choose a receiver from a list which comes from POP3 remote proxy; recording user’s voice and transforming it to an audio file, attaching audio file to email, and filling in the fields such as subject; at last sending out the voice email.

Natural language without any grammar restraint can not be translated into plain text through voice recognizer, so sending module firstly records user’s speech into audio file and then attaches the audio file to email. Email’s subject would mark this email being a voice email, and the emails with such a mark might be received by the voice-driven email client in the same way.

The last step of sending module is sending voice email to receiver’s mailbox through SMTP proxy. Saving voice email into local host is also convenient by carrying out operating module.

4.3 Voice Email Receiving Module

Receiving module is mainly composed of two parts: downloading voice email and speaking voice email to user. Fig.6 shows the structure and working flow of receiving module.

Downloading part carries out two operations: one is to distinguish which email is voice email; another is to download voice email from POP3 remote proxy.

After downloading voice email, next step is to save the downloaded emails to database. These voice emails could be divided into different types and be saved into different folders.

According to user’s choice, voice-driven email client will inquire the database and speak voice email out after separating attached audio file from voice email.

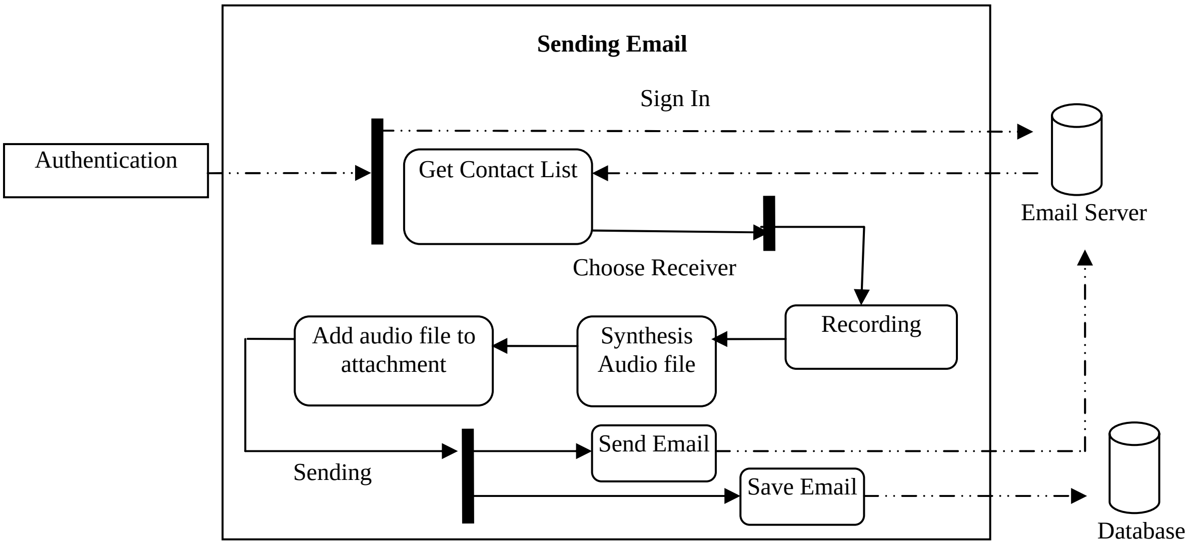


Fig.5. The working flow of voice email sending module

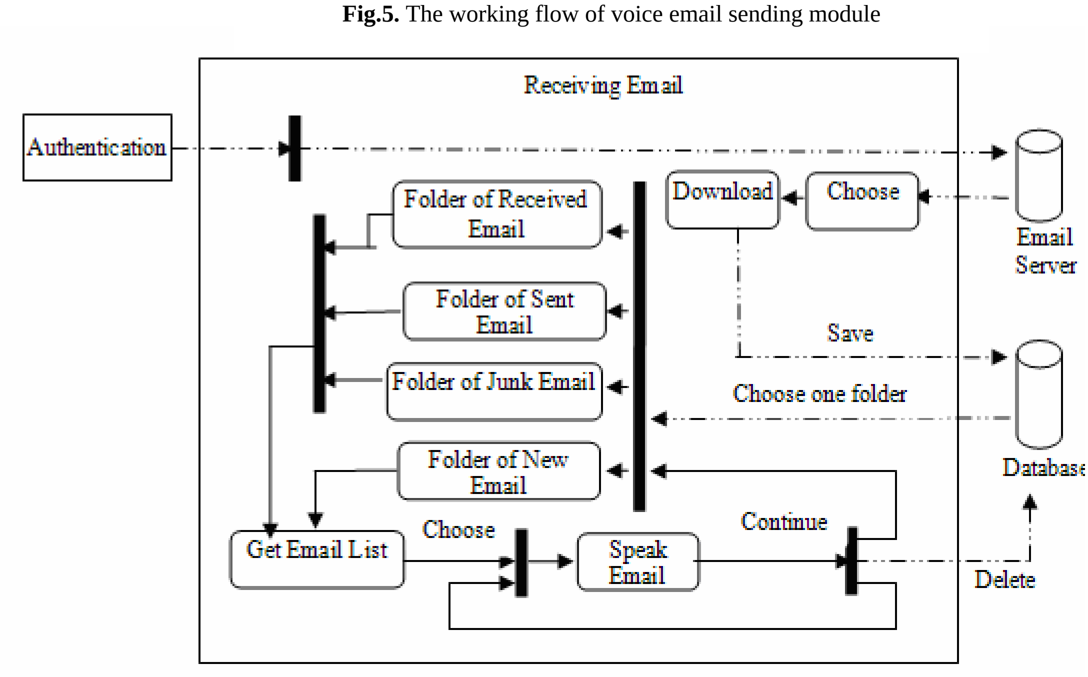


Fig.6. The working flow of email receiving module

5. CONCLUSIONS

VoiceXML is capable of meeting the needs of voice web service. VoiceXML based voice-driven email client is a feasible way to extend the traditional email service from Internet to PSTN. Email content is thus transformed from "visible" to "audible".

The programming and deployment of voice-driven email client prototype has been finished. The test shows that voice-driven email client provides effective support to voice email service, although the voice interaction process still needs to be improved.

REFERENCES

[1] Chen Lu, "The Application of TTS Speech Engine," *Journal of Taizhou Polytechnical Institute*, 2007, (01).

[2] Honglan Li, Baozong Yuan, "Research on Voice Browsing," *2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*.

[3] J. Myers, Carnegie Mellon, M. Rose, "Post Office Protocol - Version 3," *RFC* 1939, May 1996

[4] Jonathan Eisenzopf, "Selecting a VoiceXML Gateway," http://itmanagement.earthweb.com/netsys/article.php/975321 2002/2/14

[5] Jonathan B. Postel, "Simple mail transfer protocol," *RFC* 821, August 1982

[6] VoiceXML specification, "Voice browser working group of world wide web consortium," http://www.w3c.org/voice/

[7] Yin Jianqi, Wen Bin, "Internet Telephony and Interface to PSTN," *TELECOMMUNICATIONS SCIENCE*, 1998 (04).

Research on the Software of Multiple Service Access System Public Platform

Chuanqing Cheng [1], Li Wang[2]
[1]Compute Science Department, Wuhan University of Science and Techonology
Wuhan,China
[2]School of Telecommunication,Wuhan University
Email: ccqcjl2005@126.com, wl3833@126.com

ABSTRACT

This paper firstly introduces a general multiple service access platform, which can support LAN, xDSL or PON access technology. The system design is introduced and a ATDA is discussed in detail. Finally the paper gives the software structure to show the public platform can adapt to different access technology smoothly

Keywords: EPON IGMP Proxy IGMP snooping multicast

1. INTORDUCTION

With development of IP broadband router technology and the discuss of IP network, the MAN has rapidly been the load-carrying network of future telecommunication services. The router with large capacity, High bandwidth, High process large capacity will be the load-carrying network core. The IP MAN and Layer2 broadband access network structure with clear of matter and logic is essential. The structure need distributed structure hardware Component Architecture and line-rate hardware switch to meet the requirement of MAN.

According to the difference of access technology and media, current broadband access network can be divided into Ethernet access network, xDSL access network, fiber access network based on PON. On the service supplier view, a long-term coexistence of different access technology is integrant. On one hand, the current user should be maintained .On the other hand, the preliminary operation of FTTx must be developed.

Although the Ethernet, xDSL and PON technology have their different characters, the demand of large capacity, Wide bandwidth and High process ability need a distributed structure must be used in broadband access network system. From the hardware view, all type of services enter the system with a inserted card and be connected with core card to be a master-slave distributed system. From the software view, core card and service card have their own related self-governed software. But the software have logic connection with each other to build up a system, maintaining the running of device and manage or configure it.

The section 2 of the paper introduces the system design of multiple services access system. The auto topology discovery scheme is expatiated in section 3.The Section 4 discusses the software system structure. The section 5 is conclusion and future works.

2. SYSTEM DESIGN

Core card can be divided to two units: switch and management. The main chip of switch module is Broadcom5690, which can support 12 ethernet ports with 1G speed and a 10G concatenation port with 10G speed. The switch capacity can be up to 32Mbps and support line rate switch of all the ports. Ten ports of BCM5690 can be as slot ports to connect service cards. The other two ports can be as uplink ports. The core of management module is Motorola micro processor with 64M SDRAM and 16M flash. The I/O of management is composed of console and 100M FE ports. The core card used vxWorks embedded operating systems, which has been reduced and compressed to image and be saved in flash. The software undertakes the task to control hardware and to carry out core switch as well as system management.

The EPON,xDSL,LAN access network system are most popular access technology .The device can be divided to user access system, core switch system and uplink system. The structure figure is just like this:

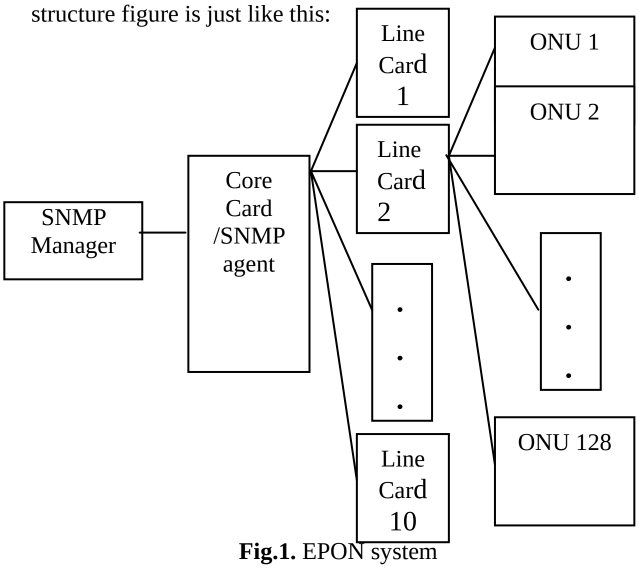


Fig.1. EPON system

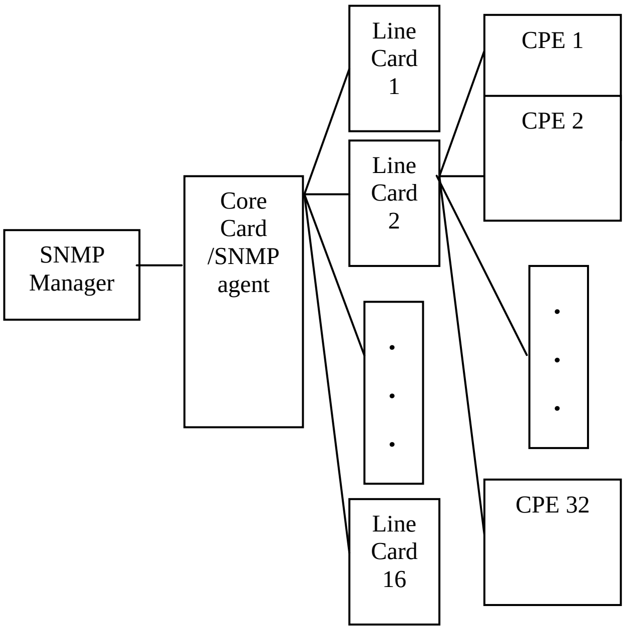


Fig.2. XDSL system

3. RECOMMENDED LINE AND PARAGRAPH SPACES AUTO topology Discovery algorithm

The Service card connects to core card by one of its Ethernet port as a downstream device. There should be a auto-discovery function. When the service card is inserted to the system, core card should discover the card automatically and admit it to the administration.

A auto topology discovery algorithm(ATDA) is adoped. The algorithm uses a auto discovery protocol (ADP) to discovery the neighbor device and maintain the connection. The protocol is on the base of vxWorks protocol stack. The switch chip usually can trap some specific multicast packets such as BPDU to CPU ,hence the software can process them conveniently. ATDA borrow the idea, modify the ordinary BPDU a bit, keep destination MAC no change as 0180c2000000.But we change the payload of the packet to fill in the private information.

On the same time, ATDA allot a default IP address to every service card inserted into the system, such as 192.168.110.1/24, 192.168.110.2/24,The core card has its own ip address of course ,which is in the same subnet with the service card .These IP address are used only in the internal but not will be diffused to public network. The system is administrated by a public IP address.

When the core card received the ADP packet, it means that the service card has entered the system. The core card can setup TCP connection to the service card. When the connection is set up, the data path is build ,by which service card can be administrated.
The flow of service card is showed in Fig 3

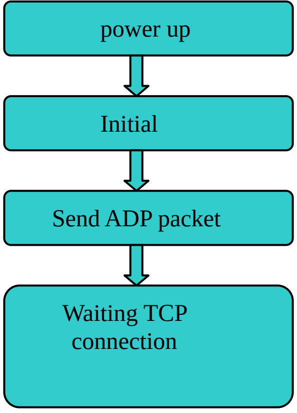


Fig.3. flow of service card

The flow of core card is showed in Fig 4

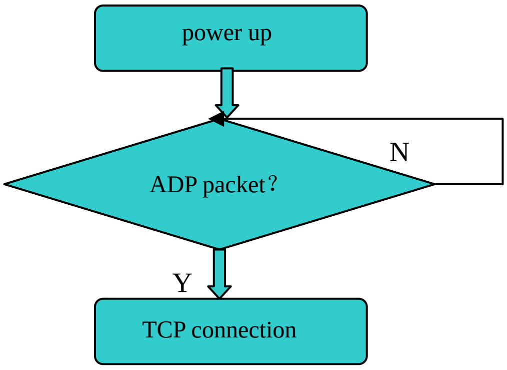


Fig.4. Flow of core card.

- Service card powered up, finish the initialization .Then send the ADP packet to the uplink port of the card self,The card information ,like slot number,slot type ,can be packaged in the packet.
- The core card is polling until the CPU process the ADP packet, parse it and get the service card information, record and maintain the information.
- The core card set up TCP connection to service card by private IP .After finishing setup connection, the data path is ready and the core card can start polling service data.

4. SOFTWARE SYSTEM STRUCTURE

The software system can be divided into two parts: kernel and user process. Kernel function is the key of the system and user process is the function extend on the base of kernel.Layer4 TCP/UDP protocol,Layer3 IP protocol and the driver is implemented in kernel. Router protocol and upper layer protocol is implemented by user process. The key task of design is driver, protocol and system management module.

4.1 Driver

Considering the hardware and software will be the public platform of multiple service access, three-layer structure is adopted in the driver to decrease the change of itself.

The bottom layer is hardware layer, which abstracts the most fundamental interface to control the switch module. The middle layer is function layer, which abstracts the interface to control switch module according to function. The upper layer is interface layer, include network device and character device, which supply a hardware independence interface and shield concrete hardware. The application can not need to change with the change of hardware.

4.2 Protocol Module

Protocol module include :layer2 and layer3 protocol ,QoS, multicast protocol and etc. As a public platform of multiple service access, The system is a master-slave concatenation system in fact. To be a integrate system ,VLAN,multicast, QoS are all need the cooperation of service card. To configure these functions, on one hand, software system must configure the core card, on the other hand, send the configuration information to service card by private protocol. Each service card parses the protocol and configure their own hardware or software. So as a integrate system, the interface must be independence with service card and not changed with the service change.

Some protocol of this module, such as spanning tree protocol(stp), does not come down to the service card ,so it will be migrated to other device expediently

4.3 SYSTEM MANAGEMENT

Besides the protocol module of core card self but also the service card, the platform undertake the task of management of all the system. The system management include :service card on/of management, service card type management, service card parameter management, polling of service card performance.

System management is on the base of data path discussed in section 2.Using some public interface and different private protocol , the system finished the master-slave configuration mode and maintain the data and saved data.

5. CONCULUSIONS

In this paper, we introduce software of multiple service access public platforms. A general multiple service access platforms are introduced firstly, which can support LAN, xDSL or PON access technology. The system design is expatiated and an ATDA is discussed in detail. Finally the paper gives the software structure to show the public platform can adapt to different access technology smoothly.

REFERENCES

[1] Jim Metzler，*Lynn DeNoia. Layer 3 switching: A guide for IT professionals[M],* Prentice Hall, 1999.

[2] *StrataXGS BCM5690 theory of operations[Z],* Broadcom Corporation,2002.

[3] IEEE Std 802.1Q-1998. *IEEE standards for local and metropolitan area networks: Virtual bridged local area networks[S].*

[4] IEEE Std 802.1D-1998. *IEEE standards formedia access control* (mac) bridges[S].

[5] Andrew S. Tanenbaum. *computer networks*(third edition)[M]. Prentice Hall,1996.

[6] Alessandro Rubini , Jonathan Corbet. *Linux device drivers*(second edition)[M].O'Reilly and Associates Inc, 2001.

[7] Christensen M,Kimball K, Solensky F,*Consideration for IGMPandMLDsnoopingswitches*<draft-ietf-magma-snoop-09.txt>[S]. Internet Draft,2003.

[8] ZHANG Xiao-zhe, WANG Xian-lei, XU Ye, FANG Gui-ming, “Design of software system in Layer3 switch”, *Computer Engineering and Design,*Vol26,2005

Learning Resources Discovery Based on Semantic Web Services

Qizhi Qiu
School of Computer Science and Technology, Wuhan University of Technology
Email: qqz@whut.edu.cn

ABSTRACT

With the development of network technology, learning resources have played an important role in e-Learning. It is necessary to find the more suitable learning resources for both educators and educatees. In order to find and compose the learning resources that are on different software and vendor platforms, this paper proposes a discovery model based on SOA, where all the learning resources are regarded as Web Services. With reference to IEEE LOM and CELTS-24, a construction of learning resource ontology is also proposed, which is the foundation of discovery. Meanwhile similarity-matching algorithm is employed to enhance the finding accuracy.

Keywords: Learning Resources, Web Services, Semantic Web, Ontology, Similarity.

1. INTRODUCTION

With the modernization of instructional management, there have been so many learning resources on the Internet, which are on different software and vendor platforms. This phenomenon is called as *Information Isolated Island*. How to make the existing learning resources accessible is one of the research focuses. The popular solutions to this sort of problem are EAI (Enterprise Application Integration) and Web Services. Researchers have done more on them. Niu proposes a feasible solution to the problem of the modernization of legacy systems, where the educational resources are dispersed and isolated [1]. Lin uses QoS to optimize the discovery of learning objects [2]. There are some drawbacks in those existing solutions, such as semantics and so forth.

This paper proposes a different discovery mechanism for learning resources. The mechanism aims at two problems. The first is description lack of semantics, including the loss of mapping OWL-S (Web Ontology Language for Services) into WSDL. And the second is the disagreement between the descriptive standards for learning resources. The following are helpful to implement the mechanism:

1.1 Semantic Web Services

Web services provide a standard means of interoperating between different software applications, running on a variety of platforms and/or frameworks. In order to provide a more comprehensive, more expressive framework for describing all aspects of services, Semantic Web technologies have been built inside the web services, for example, OWL.

1.2 Standards for Learning Resources

Although designers and developers of online learning materials have an enormous variety of software tools at their disposal for creating learning resources, descriptive labels can be used to index learning resources easily. There are many descriptive standards for the learning resources:

- The IMS Learning Resource Meta-data Specification aims to make the process of finding and using a learning resource more efficient by providing a structure of defined elements that describe or catalog the learning resource and requirements about how the elements are to be used and represented [3].
- The IEEE LTSC (Learning Technology Standards Committee) has been developing internationally accredited technical standards, recommended practices, and guides for learning technology. The IEEE LTSC coordinates with other organizations, such as ADL (Advanced Distributed Learning Initiative, a program of the US Department of Defense), Dublin core metadata initiate, IMS Global learning consortium, that produce specifications and standards for learning technologies.
- CELTS (Chinese E-learning Technology Standardization) is developed by Chinese government to cover many aspects of learning technology and to address most conceptual issues encountered in the process of E-Learning, with reference to mainstream E-Learning standards/specifications, especially IEEE LTSC. The sub-standard for QoS characterizes CELTS [4].

1.3 Fuzzy Theory

Fuzzy theory is a challenge to tradition. Most natural phenomenon can be described more accurately by fuzzy set theory. It has changed the way of human cognition; its logic is based on a closed interval [0,1] instead of *either 0 or 1*. Up to now, fuzzy theory has showed its power in handling the uncertain, non-linear problems; and because of its flexibility, it can cooperate with many other technologies, such as automatic control, data-categorizing, cognitive science and so on.

2. PROPOSED MODEL

Fig.1 shows the model, which is proposed also in SOA. The difference between the proposed one and the popular one is how to describe the Web Services. The ontology-based descriptive method will benefit the finding operation of SOA because of both syntactic and semantic description.

2.1 Compatibility

In order to avoid the semantic loss in the mapping process between OWL-S and WSDL, the model abandons the popular WSDL. In addition to, there is an enhanced UDDI in the model. Compatibility characterizes the model.

- Compatible with SOA [5]: as shown in Fig.1, the model also supports three main operations (such as publishing, finding and binding) between three roles (that is, service provider, registry and service requester). That makes it possible for developers to develop the Web applications independent of environment, programming languages and so on; and it also leads to the reuse of existing resources, such as software, hardware and some other resources.
- Compatible with the UDDI's extendability [5,6]: tModel is a useful element of UDDI, which can support the defined semantics. tModel is a kind of metadata about invocation information including service name, publishing organization, URL , etc. In the proposed model, tModel is

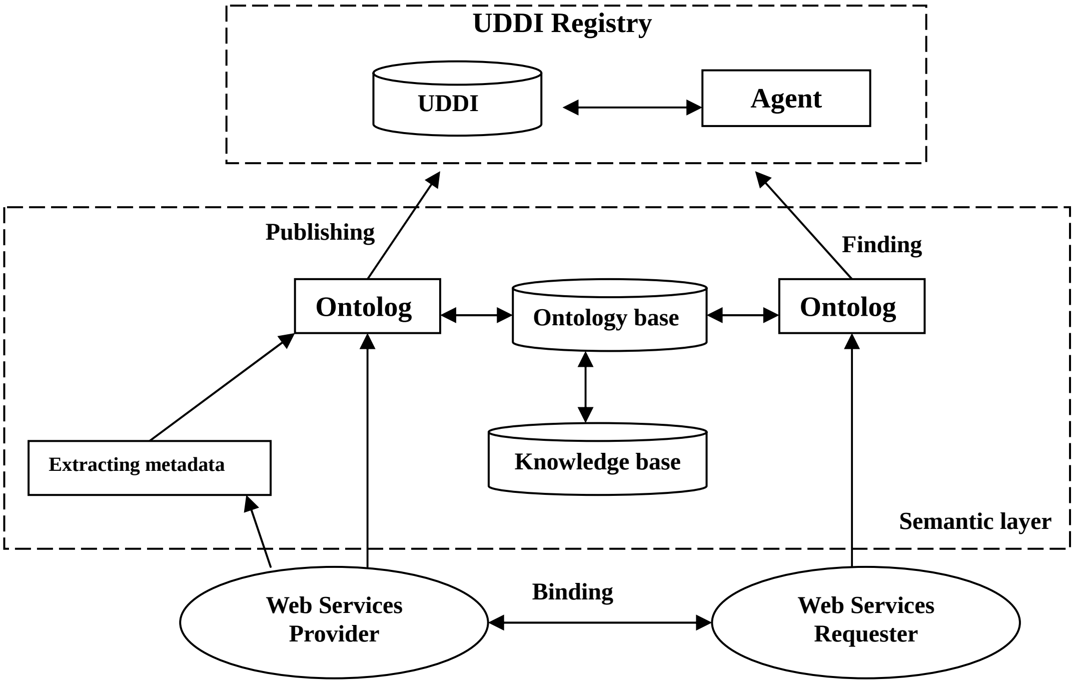


Fig.1. The Proposed Model

used to describe a categorization for learning resources, meanwhile there is an ontology management Agent. With the help of tModel, Web services are characterized and categorized according to the keywords, so the requestors can locate the suitable service.

- Compatible with XML [7,8]: although OWL-S is abandoned in Fig.1, OWL is employed to describe the Web Services ontology. As we know, all data type supported by RDF and XML Schema can be used in the OWL document, so we can make the XML-based message mechanism work well with SOAP in the proposed model.

2.2 Three Major Roles

2.2.1 Provider and Requester

As we know, the provider entity is the person or organization that provides an appropriate agent to implement a particular service, while a requester entity is a person or organization that wishes to make use of a provider entity's Web service [9].

There are four different ways to develop a web service with the help of WSTK (IBM Web Services ToolKit), and only two of them is shown in Fig.1. The fisrt one is for a new web services. Besides implementing the functions of the web service, the provider should decsribe it to follow the standard in section 3. The other one is for a exsiting one. In order to provide the metadata for those existing services, the model in Fig.1 employs an agent that can extract the metadata of the Web Services to follow the defined framework for the ontology.

2.2.2 Registry

In SOA, a registry is an authoritative, centrally controlled store of information. UDDI is often seen as an example of the registry approach. In order to find the services which aren't described in WSDL+OWL-S, the model makes use of UDDI tModel, and an agent is used to transform the ontology.

3. ONTOLOGY

The proposed model in Fig.1 is for general purpose to enhance finding efficiency. The following sections will discuss learning resources discovery.

Ontology is an important part in the semantic web architecture proposed by Tim Burners-Lee in 2000 [10], and it has been chosen to describe the metadata about learning resources.

3.1 Requirements Analysis

3.1.1 IEEE LOM

IEEE standard for learning object metadata (LOM) specifies a conceptual data schema that defines the structure of a metadata instance for a learning object. All elements in the schema are divided into nine top categories: General, Life Cycle, Meta-Metadata, Technical, Educational, Rights, Relation, Annotation, and Classification, and each of these branches comprises several elements. Although LOM has been an authoritative standard, it is too cover-all to work out. Meanwhile some of sub-elements are conflicted or overlapped. Table 1 shows the elements that are chosen to describe the learning resources in this paper [3].

Table 1. Metadata about Learning Resources

category	element	date type
general	identifier	character
	title	character
	language	character
	description	character
	keywords	character
life cycle	version	character
	contribute	character
technical	format	character
	size	numeric
	duration	numeric
educational	intended end user role	character
rights	cost	numeric

3.1.2 CELTS

Most researches on standardization of quality of learning resources are concerned about ：Institutional Support、Course Development、Teaching/Learning、Course Structure、Student Support、Faculty Support、Evaluation and Assessment. As a sub-standard of CELTS, Specification for Service Quality Management System of e-Learning (CELTS-24) is the first one for quality of services of the learning resources [4]. Its framework is composed by five dimensions: reliability, responsiveness, assurance, validation of learning resources and empathy.

Table 2. QoS of the learning resources

name	comments
scientificity	evaluation about the reliability, validation in terms of knowledge
accessibility	evaluation about the organization of knowledge (e.g. wizard) and its accessible difficulties
real-time	interval of the update

CELTS-24 is intended to make the users choose the e-Learning institutions according to their preference, while the purpose of the model shown in Fig.1 is to assist with learning resources discovery. In the model, the resources are regarded as Web Services.

QoS of the Web Services is concerned by requesters. With references to validation of learning resources in CELTS-24, Table 2 shows the QoS attributes of the learning resources.

3.2 Ontology of Learning Resources

Definition 1: A Web Service is defined as:

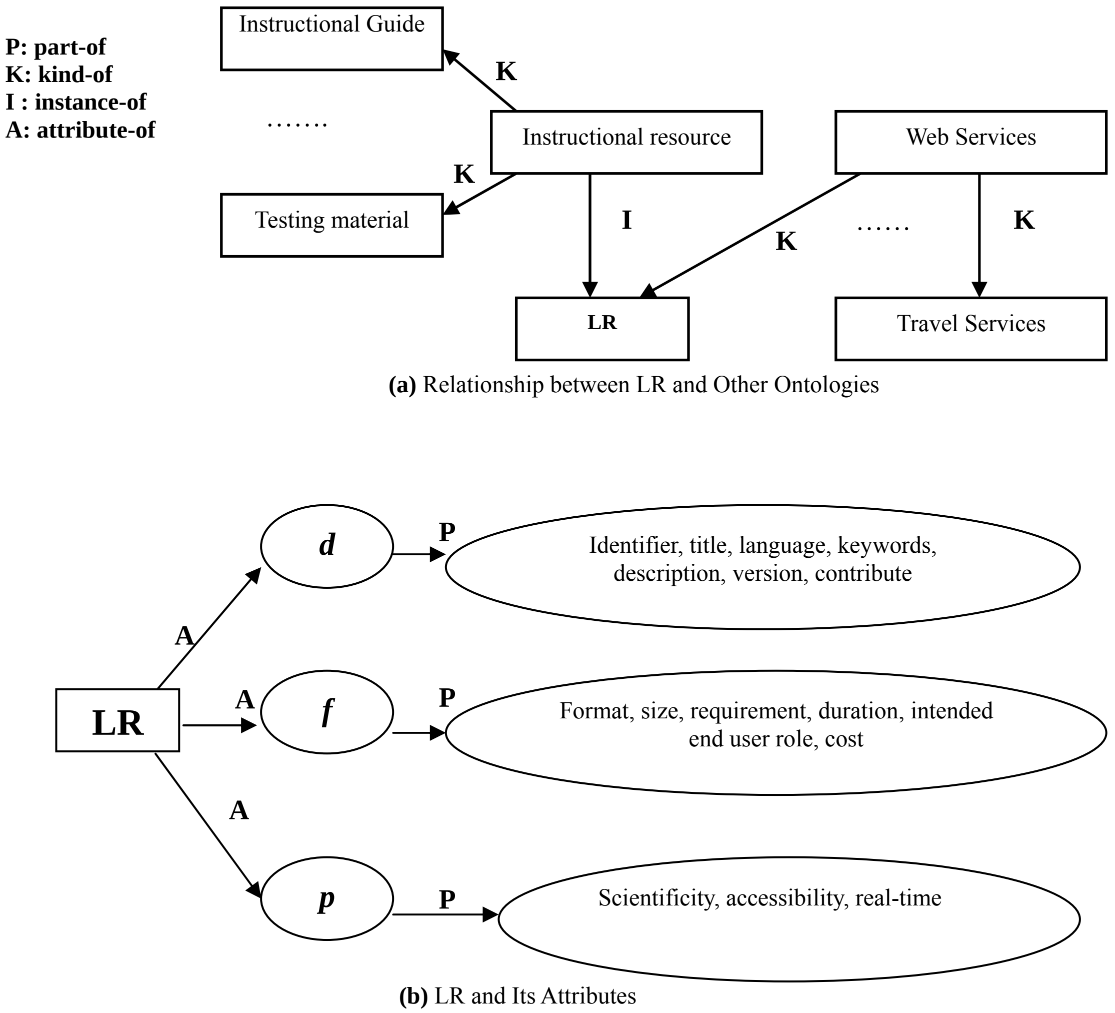
Service（*d*，*f*，*p*）

Where:

- *Service* is a Web Service that supports semantics;
- *d* is the basic information about *Service*.
- *f* is the functional information about *Service*.
- *p* is the performance information about *Service.* [11]

Ontology of learning resources (for short: LR) is a kind of domain ontology, which is used to define the domain knowledge. According to the analysis in 3.1 and the above definition of web services, the ontology is composed by three elements as shown in Fig.2.

Fig.2 shows the framework of ontology LR and the hierarchy of the ontology. In fact, the ontology LR is also a triple: <concept, attribute, axiom>. Making use of the three sets, we can describe the ontology not only the thing's nature, but also the relationship between the things. Furthermore more reasoning can be made by the set of axioms.


P: part-of
K: kind-of
I : instance-of
A: attribute-of
Instructional Guide
K
…….
Instructional resource
Web Services
K
Testing material
I
K
……
K
LR
Travel Services
LR
d
f
p
A
A
A
P
P
P
Identifier, title, language, keywords, description, version, contribute
Format, size, requirement, duration, intended end user role, cost
Scientificity, accessibility, real-time

(a) Relationship between LR and Other Ontologies

(b) LR and Its Attributes

Fig.2. Construction of LR

4. MATCHING ALGORITHM

4.1 Tversky Function

Psychologist Tversky proposed his feature-based model, which regards similarity as a function as follows:

$$s(a,b)=\theta \cdot f(A\cap B)-\alpha \cdot f(A-B)-\beta \cdot f(B-A) \quad \text{Eq. (1)}$$

where:

- *s* is the similarity between concept *a* and *b;*
- *A*, *B* is the feature set of concept *a* and *b*, respectively;
- *A*∩*B* represents the public features of *a* and *b*, while *A-B*, *B-A* represents the distinct features of *a* and *b*, respectively;
- Parameter θ、α、β are weight coefficients.

Tversky function shows that similarity increases with the increase of public features or decrease of the distinct features.

In the term of data retrieval, the information on Web is characterized by the sparsity. So it is not practical to describe it in *either-or* way. Owing to the fuzzy theory, similarity matching-algorithm is proposed, which can quantitatively depict the similar degree between two services, so that the requesters can choose the services according to their own preference.

4.2 Matching Process

Finding is one of the major operations of SOA; it receives the consumers' requests and finds the relevant one to match the request in UDDI. The matching process employed in the proposed model matches the requester and provider in the terms of function and performance; therefore the finding problem is transformed into computing function optimal solution with performance constraints. There are several different matching policies in the matching process.

Nomenclature:

- s(*a,b*) is similarity of a and b, where *a* is the sample (also called as a provider), *b* is the object. *a* and *b* are either Web Services or attributes.
- |*a*|、|*b*| is the number of the attributes of *a*, *b* respectively.
- |*a*∩ *b*| is the number of public attributes of *a* and *b*.
- |*a-b*|, |*b-a*| is the number of distinct attributes of *a* and *b*, respectively.

4.2.1 String-matching

Because of mature specifications [3,4] for LR, most attributes of vector *d* and *f* (Definition 1) in LR, such as *language, format*, can be indexed by the traditional matching based on string. It is an effective method that indexes information by thesaurus with the result *either 0 or 1*. By this way the searching range shrinks greatly.

4.2.2 Numeric Computing

In order to get a more accurate value about the numeric attributes matching, the proposed model computers a ratio.

Suppose *Req* is a numeric attribute of the requester, while *Pro* is that of the provider,

If $Pro > Req$, then $s(Pro, Req)=0$;

If $Pro \leqslant Req$, then $s(Pro, Req)=Pro/Req$. Eq. (2)

Eq.(2) is suitable for attributes *cost, size, duration real-time* and so forth.

4.2.3 Similarity-matching

As for those are not exactly matching during the string-matching and vector *p* (Definition 1) in LR, similarity-matching will be employed (more details in 4.3). The totality of similarity s (*a*,*b*) is computed by Eq. (3)

$$s(a,b)=\Sigma\ (\mu_i \quad s_i(a,b)) \quad \text{Eq. (3)}$$

where

$\sum\mu_i=1$, μ_i is the weigh coefficient

Therefore, computing and sorting will find the most suitable LR.

4.3 Similarity Algorithm

Similar to *object* in Object-Oriented technology, there is a hierarchy structure in ontology, therefore the feature-based model about Tversky Function, which is feature-based, is suitable for the ontology matching.

A simplified form of Tversky function is applied:

$$s(a,b)=\frac{|a\cap \mathrm{b}|}{|b|}-\frac{|a-b|}{|b|}-\frac{|b-a|}{|b|} \quad \text{Eq.(4)}$$

By making use of class hierarchy and relationships between concepts, it is easy to compute each vector in Eq.(4). Eq.(4) is for the different classes, and the following conclusions are drawn for particulars by analysis:

If *a* and *b* are the same class, then s(*a*,*b*)=1.

If *a* is a subclass of *b*, then s(*a*,*b*)=1.

If *b* is a subclass of *a*, then s(*a*,*b*)= $\frac{|a|}{|b|}$.

Eq.(4) is suitable for those attributes which have defined hierarchy, such as *keywords*.

4.4 Others

Distinguished to other Web Services, LR is a kind of Web Services whose purpose is to help the requester get knowledge according to their interests. So its correctness and convenience are important. Attributes *scientificity* and *accessibility* are responsible for this.

Each attribute is assigned to a statistic because it is hard to give a fair evaluation. This paper computes the average of three values to assign the attributes. The three values are evaluations from provider, learner and certificated authority. In order to keep the consistency with the above computation, all the evaluations range from 0 to 1. Therefore during the matching process, they can be computed by Eq.(2).

5 CONCLUSIONS

The purpose of this paper is not only to apply Semantic Web Services technologies to manage the learning resources, but also to enhance the accuracy of discovery. As for the instructional management systems (for short: IMS), standardization and scientificity are the challenges. With reference to the existing standards, ontology LR is built to support the discovery based on semantic and QoS.

The related tool kit is Apache Geronimo Web application server, IBM UDDI registry, Java programming language, Protégé ontology editor, etc.

The further work includes:

- Study on more matching algorithms for Web Searching, such as GOT (Global Optimization Technology), cluster analysis. Simplicity and swiftness are crucial for the algorithms.
- Perfect the construction of both ontology Web Services and Learning Resources. The challenge is how to design an open architecture.
- Apply the methodology to other domains, such as e-Commerce.
- Study on the communication between the international standards (such as IEEE LOM) and CELTS. What this paper does is to select the elements from the related standards to build ontology LR. Up to now, those different

standards for learning resources are not compatible. That means the provider of learning resources must make a compromise: either international or local.

REFERENCES

[1] Niu Zhicheng et al, “Education Resources Integration Research Based on Web Service Technology,” *Journal of Shenyang Jianzhu University* (Natural Science) , No.6, 2006

[2] Lin Hao, “Discovery of Learning Objects Based on Web Services and Its Optimization,” *Journal of Qingdao Technological University*, No. 2, 2006

[3] “Standard for Learning Object Metadata" http://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html, 31 August 2006

[4] Chinese E-Learning Technology Standardization Committee, “Specification for Service Quality Management System of e-Learning,” http://www.celtsc.edu.cn/680751c665875e93, Mar. 2003

[5] Gu Ning et al, “Principle and Development of Web Services,” China Machine Press, 2006

[6] Organization for the Advancement of Structure Information Standards, “UDDI Executive Overview: Enabling Service-Oriented Architecture,” http://uddi.org/pubs/uddi-exec-wp.pdf, October 2004

[7] David Martin, Mark Burstein, “OWL-S: Semantic Markup for Web Services,” http://www.daml.org/services/owl-s/, March 2006

[8] Oloivier Damerron, Natalya F. Noy, et al, “Accessing and manipulating ontologies using web services,” *Proceeding of Semantic Web Services Workshops at ISWC 2004*

[9] David Booth. et al, “Web Services Architecture,” http://www.w3.org/TR/we-arch/#reqpro, 11 Feb. 2004

[10] Tim Berners-Lee, “Semantic Web–XML2000,” http://www.w3.org/2000/talks/1206-xml2k-th1/silde10-0.html, 2000

[11] Qiu Qizhi, et al, “Study on Ontology-based Web Services Discovery”, *Proceeding of CSCWD 2007*, April 2007

Qizhi Qiu is an Associate Professor in School of Computer Science and Technology, Wuhan University of Technology. She got her master degree from Wuhan Transportation University in 1999, and now she is a Ph.D candidate of Wuhan University of Technology. Her research interests are in e-Commence, Web Services, Semantic Web.

The Application of Service-Oriented Architecture in an OA System*

Wang Chao, Wan Yan
School of Economics and Management, Beijing University of Posts and Telecommunications
Beijing 100876, P. R. China
Email: dynasty0628@sina.com, wanyan@bupt.edu.cn

ABSTRACT

SOA is a newly founded method to guide the integration of enterprise's information system recent years. This article shows the basic concept and a real application of it. Through analyzing the structure of the case system we can educe advantages and some disadvantages of SOA. The practice is proved to have advantages in the aspect of reusability, extensibility and data synchronization. At the same time using middle-ware can realize rapid change but bring forward higher hardware requirement.

Keywords: SOA, Integration of Information System, Middle-ware

1. EDUCTION OF THE PROBLEM

The work of development and test for a newly founded commercial bank's office automation system has just finished. At the beginning of the designing, through analyzing the character of the enterprise and the requirement of them, we considered the system should own character of maintainability, extendibility, can be integrated and rapid change. Based on these characters we decided to adopt SOA architecture.

2. BRIEF INTRODUCTION OF SOA

SOA is short for Service-Oriented Architecture. It came into being on the basis of the development of Web technology to solve the problems of Web technology's information system integration. It is used to guide the designing of enterprises' next generation information system. The essence of SOA is to make every part of an information system working cooperatively on the basis of communication. By doing so, it is able to reduce or eliminate information fragment and realize loose couple among systems. Generally speaking, SOA is not a kind of technology but a way of linking systems together. It connects different function units (called services) of application programs by defining compatible interfaces and protocols among them. The interfaces are defined neutrally, so it is independent of hardware platforms, operating system and programming language that are used for the realization of service. Consequently, the services built in various systems are able to communicate with each other in a uniform and universal way such as through middleware. Middleware is the core of SOA, located above operating system, network, and database while below application software. It was put forward to solve distribution isomer problem. Thus, it is also independent of hardware platforms, operating systems and programming languages.

3. A CASE OF SOA — SOME BANK'S OA SYSTEM

This refers to the newly built National Joint-venture Commercial bank, whose office automation system is based on SOA. Its system is built on the software of IBM Lotus Domino series and adopts B/S architecture. According to the theory of SOA, the system is divided in several modules and every module has its function and offers standard interfaces. The whole system can be divided into several parts. They are independent from each other and stored in different Lotus database. The first part is organization structure module. It contains all the organization structure information of the bank, including organization's hierarchy, department leaders and personnel. The personnel information synchronizes the bank's AD (Active Directory) system through Websphere. The module can offer different functions through the different interfaces offered, including supplying all departments' information of the bank, querying the department by some personnel's name and returning the department leader and personnel according to the department name, etc. The second part is personal information module, which defines the interfaces of obtaining personal messages and pending tasks. When the users log on the website, they can get their own messages from the individual front page. The third part is the flow management module. It contains all flow information of the whole OA system's application modules. The system offers standard interface to define flow and expresses it with the type of visibility. With these interfaces, operator can define author, reader, authority and selection condition for every flow node.. And then it connects with every application flow form through setup management. The fourth section is the aggregation of application process modules which are based on the three modules discussed above and can be designed with standard development template. When a form is build it can connect its corresponding flow in the flow management module through simple visible operations. Among these application modules, a few of them need to be integrated with some modules outside the OA system. For example, the module for checking work attendance should read the data from SQL Server through ODBC to determine if the application is legal and stationary application module from ERP to check the res and its amount (see Fig.1).

4. SYSTEM'S SOA FEATURES

4.1 Standard format for information storage

In every module of the system, if the information needs to interact with that outside, it is saved in standard format of XML. At the same time, all kinds of standard interfaces are defined to get the information. In this way, the information format is ensured to be universal. No matter the OA system or others, if only have enough authority, they can obtain information using these interfaces.

In each module, it may be necessary to store different kinds

* This paper is supported by the open fund of Key Laboratory of Information Management and Information Economics, Ministry of Education P.R.C. F0607-10.

of information, so several XML archives will be generated according to the associations between them. This ensures users to obtain the exact information they want rather than redundant information or information containing redundancy.

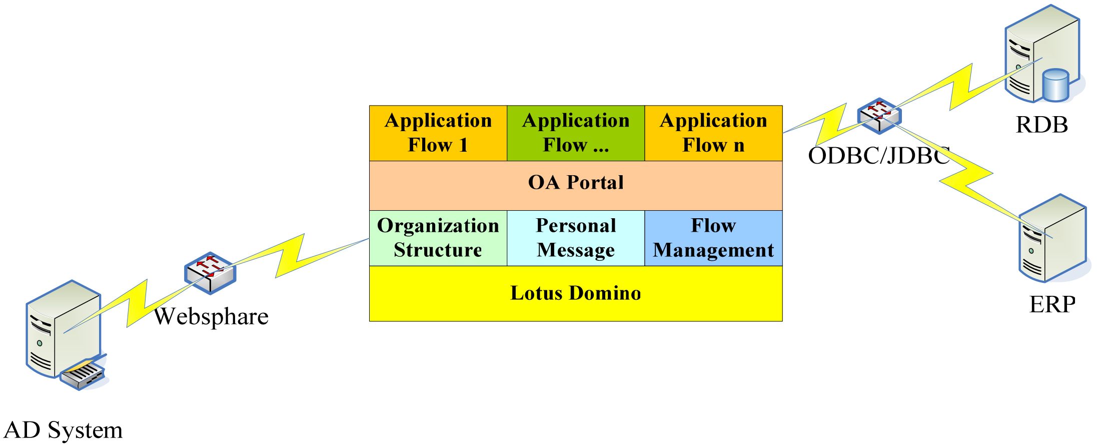


Fig.1. OA system's structure of some bank

4.2 Platforms Independence

First of all, with respect to the fundamentals of the system, Lotus Domino is independent of platforms. Thus the systems built above have the same character. Secondly, information stored in other modules of the system is expressed in standard XML, so the systems of any other platforms can use the obtained information directly. Thirdly, though some middleware based on Windows platforms such as ODBC is adopted to integrate the system with outside ones, that's because the server in this case is installed on Windows operating system. When changing platforms, it can easily use other middleware instead, such as replacing ODBC with JDBC.

4.3 Integrating with other systems

Although OA is relatively independent from the whole bank's systems, it is a not information fragment since it also has intercommunion with outside systems through middleware. That's because all the information about data is stored in related database or ERP, when needed it is read or updated with the connection utility of middleware.

The character above fits that of SOA completely. If its modules are considered as components of SOA, then the system is a typical SOA system.

5. THE ADVANTAGES OF ADOPTING SOA

5.1 Simple Structure of the System and the Strengthened Reusability

The function of the modules is decided according to various entities in the system. Every module has its own relatively independent function and serves for the whole system. Other modules, if only have enough authority and parameters, are entitled to call the service. In this way, the reusability of the module is strengthened greatly. For example, a series of function buttons are defined in a sub-module and called by all the application flow modules. In the calling progress, the module will transfer parameters according to the flow's setup and identity of the users logging in and define the authority of users. Thus, there is no need to design function buttons for every application module. Then the reusability of the module is strengthened and the system's designation is simplified, since the only job to do is dividing the system into several modules by functions.

5.2 Good Extendibility

Since the function modules of the system are relatively independent and offer all kind of service interfaces according to requirement, when there is a new business requirement, there is no need to overthrow original structure. The only thing to do is to add new modules or new functions on the original module, and in this way new module can serve the whole system. For example, to add a new function of sending short messages, the task can be achieved by adding a new module of sending short messages. Then by simply adding a button to the function button definition module, the function can be realized in all the application modules.

5.3 Middleware, Foundation of Flexibility

Traditional system integration has experienced three stage, data integration, function integration and agent integration. But all of them are based on the structure of point-to-point connection, which causes the difficulty in implementation and maintenance and at the same time makes the cost rises. Because the relation between sub systems is tightly coupled, it is difficult to change and adjust itself to the development of enterprise business. Meanwhile, integration method of point-to-point asks to know the detail of every system, and this also increase the difficulty of integration. If adopting middle-ware, sub system links with middle-ware directly and all the service the systems can offer is registered in middle-ware and then if one changed, there is no need to change all other ones but middle-ware only, so rapid changes are realized. See fig.2, when using middle-ware and one system changed, the work to do is changing middle-ware but not all the other ones. The broken lines stand for adopting traditional point-to-point, when one changed, all the other related should be changed.

5.4 Data Sharing and Synchronization

Since SOA system is independent from hardware platform, operating systems and programming languages, it is convenient to share data maintained on other platforms. Data sharing has two advantages, the first is that it can reduce the work of maintaining data and there is no need to maintain data aimed at every system; the second is data synchronization, it can effectively diminish or eliminate the phenomenon of data disagreement. Let's take a simple example, there are all kind of systems in the bank, including OA, ERP, Financial, HR etc, even every department has its own system. Traditionally, the users have to remember their user names and password for every system. When need to change them, they have no choice but changing all of them one by one. After data sharing, the user's information is maintained in AD and all the changes operated in one system will take effect in all the other system.

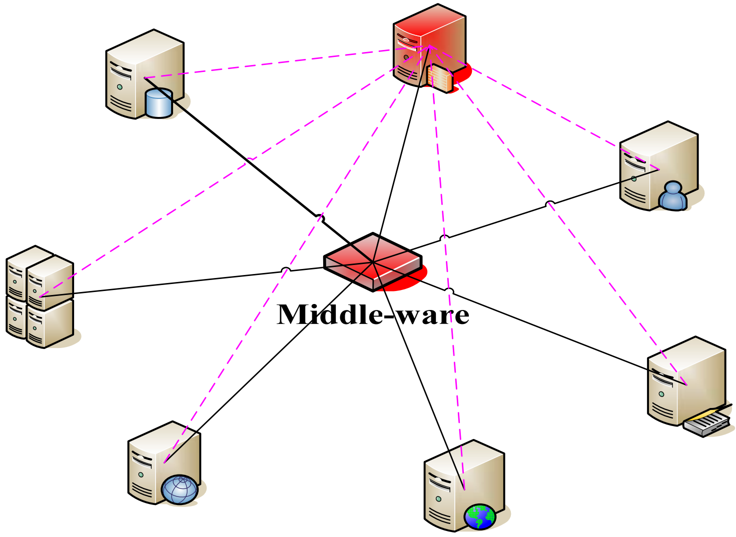


Fig.2. Adopting middle-ware to realize rapid changes

6. DISADVANTAGES OF ADOPTING SOA

Of course there are some disadvantages in the system. Firstly, to realize information synchronization, when integrated with the system time related, the server should read the information time to time. This increased the burden of the server and bring forward higher hardware requirement. For example, to realize the OA system's user name and password be accordant with other system, the method adopted is to share AD's information with organization structure module. And this asks for an agent read the information from AD through Websphare time to time. This not only increases the server's burden but cause the phenomenon of data can't take effect real time. Secondly, although from macro view the SOA system is independent from programming language, to every sub system, it is programming language related. Thus, when shared, the data must be transferred into proper format. For example, the information read from AD must be transferred to the OA system's format and stored in organization structure module; only in this way can the information be used by OA system. Lastly, since the system was asked to serve outside ones by SOA, when facing various requirement outside it should contain a series of interfaces to supply kinds of data. In order to ensure the effectiveness of information, it is necessary to store the same data in different forms, and thus causes redundancy. For example, in the organization management module, the relationship among personnel, organization and leader is defined in two ways. The first is making personal ID as key word and the second organization. When to query some leader, according to the fact either personal ID or organization can be used as parameter. Though, even if the data is stored in only one form, the process can also be finished, experiences prove that it is worth of wasting some space to attain efficiency.

7. CONCLUSIONS

SOA is a better way to build information systems. The one month's test run proved that the system fit the requirements effectively and could realize rapid changes when facing various new or changed requirements. So, it can be concluded that the system built under the direction of SOA is successful.

REFERENCES

[1] Jian Bin, Yan GuangRong and Zhu XinXiong, "SOA-Based Business Process Integration System for Small and Medium-Sized Manufacturing Enterprises", *Journal of computer-aided design & computer graphics*,Vol119 , No11, Jan1 , 2007

[2] Huang JunBin and Tang DeYou, "Research on SOA-Based data changed middle-ware", *Journal of ZhuZhou Institute of Technology*, Vol20, No6, Nov. 2006

[3] Chen Peng and Li GuangYao, "Software development based on SOA", *Journal of Henan University of Science and Technology: Natural Science*, Vol27 No.5, Oct. 2006

Distributional and Parallel Processing Database System Based on EJB

Yetian Li, Qingping Guo
School of Computer Science and Technology, Wuhan University of Technology
WuHan, HuBei 430063, China
Email: liyetian1982@126.com, qpguo@mail.whut.edu.cn

ABSTRACT

Introduce the framework of J2EE, the conception and the composition of EJB. Expound the conception of distributed and parallel processing technology and system at last, put forward a new model to construct a kind of distributed and parallel processing database system with J2EE and EJB technology.

Keywords: J2EE, EJB, Distributed, Parallel, Database System

1. INTRODUCTION

With the development of the computer network technology, it has been unavoidable that the collections, saving, processing and dissemination of the data with database technology change from concentration and occluded type to distributed and opening type.

The distributed database makes the system have appropriate data redundancy through replicating to increase the credibility and usability of the system and provides the partial autonomic data sharing and the coordination between places to increase the data processing ability.

Moreover, through the combination of the database technology and the distributed parallel processing technology, making use of the scale performance produced by parallel transaction of the multiprocessing machine, the fast reaction ability of the system could be improved largely.

Recently, more and more business enterprise own many geographically dispersed subsidiary companies and it directly results in the business data scattered. The parent company and the subsidiary companies have been placed in different cities or different region in the same city, so they need data communication and transaction in business one another in addition to their own data processing.

The J2EE frame structure, especially the emergence of the EJB, provides an approach for the business enterprise to set up the distributed and parallel processing database system.

2. BRIEF EXPLANATION OF J2EE PLATFORM

J2EE(Java 2 Platform, Enterprise Edition) technology simplifies enterprise applications by basing them on standardized, modular and re-usable components Enterprise JavaBeans (EJB), providing a complete set of services to those components, and handling many details of application behavior automatically. By automating many of the time-consuming and difficult tasks of application development, J2EE technology allows enterprise developers to focus on adding value. That is, enhancing business logic, rather than building infrastructure[1].

And the J2EE platform uses a multitiered distributed application model for enterprise applications. Application logic is divided into components according to function, and the various application components that make up a J2EE application are installed on different machines depending on the tier in the multitiered J2EE environment to which the application component belongs. Fig 1 shows the J2EE applications divided into four tiers described in the following list[1]:

- Client-tier components run on the client machine.
- Web-tier components run on the J2EE server.
- EJB components run on the J2EE server.
- Database system runs on the database server.

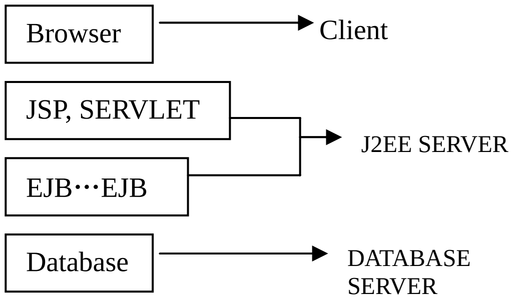


Fig.1. the architecture of J2EE

But J2EE multitiered applications are sometimes considered to be three-tiered applications because they are distributed over three different locations: client machines, the J2EE server machine, and the database or database systems at the back end.

What's more, J2EE takes advantage of many features of J2SE such as "Write Once, Run Anywhere" portability, JDBC API for database access, CORBA technology for interaction with existing enterprise resources, and a security model that protects data even in internet applications. Building on this base, J2EE adds full support for Enterprise JavaBeans components, Java Servlets API, JavaServer Pages and XML technology. The J2EE standard includes complete specifications and compliance tests to ensure portability of applications across the wide range of existing enterprise systems capable of supporting J2EE.

3. BRIEF EXPLANATION OF EJB TECHNOLOGY

EJB (Enterprise JavaBeans) is the most important part of J2EE.The EJB architecture is component architecture for the development and deployment of component-based distributed business applications. Application developers do not need to understand low-level transactions and state management details, multi-threading, connection pooling, and other complex low-level APIs. Applications written using EJB architecture are scalable, transactional, and multi-user secure. These applications may be written once, and then deployed on any server platform that supports the EJB specification. Fig 2 shows the architecture of EJB, it mainly contains the following parts[2][3]:

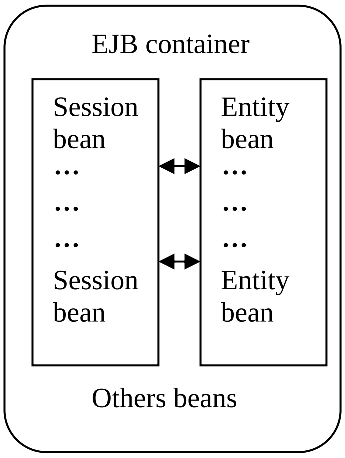


Fig.2. The architecture of EJB

(1) Stateless Session Beans:
A Stateless session bean does not maintain a conversational state for a particular client. When a client invokes a method of the class of a stateless bean, the bean's instance variables may contain a state, but only for the duration of the invocation. When the method is completed, the state is no longer retained. Except during method invocation, all instances of a stateless been are equivalent, allowing the EJB container to assign any instance to any client. Since stateless session beans can support multiple clients, they can offer better scalability for applications that require a large number of clients. A typical example is the beans that cope with complicated mathematic calculation. We use a stateless session bean to implement matrixes multiplication calculation in the numerical experiments.

(2) Stateful Session Beans:
The state of an object consists of the values of its instance variables. In a stateful session bean, the instance variables represent the "state" of a unique client-bean session. Since the client interacts (talks) with its bean, this state if often called the conversational state. The state (the bean’s instance variables) is retained during the client-bean session. If the session terminates or the bean terminates, the session ends and the state disappear. The transient nature of the state is not a problem, however, because when the conversation between the client and the bean ends there is no need to retain the state.

(3) Entity Beans:
Entity beans represent objects that persist through a server shutdown. The data representing an instance of an entity bean is typically stored in rows and tables of a relational database, which can be accessed using a JDBC data store. These tables can also span multiple databases. Some of the common examples of an entity bean are customers, departments, orders, and inventory products. There are two types of entity beans: bean-managed Entity beans and Container-managed beans.[4]

4. THE DISTRIBUTED PROCESSING SYSTEM AND THE PARALLEL PROCESSING SYSTEM

The distributed processing system and the parallel processing system are two different kinds in the computer system structure.

The distributed processing system linked many computers having different function or different location or different data by correspondence networks and complete the information processing task in phase under the unite management of the control system.[5]

And that the parallel processing system is a kind of system that makes use of several function components or more than one processing machine working at the same time to improve its capability and reliability. The system includes parallel processing of instruction-grade or above at least and its research and development refers to computing theory, arithmetic, system structure, software and hardware aspects etc.[6]

The distributed processing system has the close relation to the parallel processing system. With the development of the corresponding technology, the boundary between them is more and more misty. From broad sense, the distributed processing could be considered to be a kind of parallel processing.

Generally thinking, the tight coupling multiprocessor system concentrated in the same machine cabinet or the same location and the large scale parallel processing system could be called the parallel processing system. And the computer system connected by LAN or WAN is called the distributed processing system.

The distributed processing system includes hardware, control system, interface system, data, applications and person six factors. And the control system includes distributed database system, distributed operating system, along with the communication protocols[11].

The system as followed will add the parallel processing factors into the distributed database system and put forward a new model — the distributed and parallel processing database system.

5. THE DISTRIBUTED PARALLEL PROCESSING SYSTEM

The distributed and parallel processing system is composed of many stations which can be called nodes. The nodes are connected by correspondence networks and each of them is a self-govern database system owning respective database, central processor, terminal and local database management system. So the system could be considered to be the combination of a series of concentrated database systems which belong to the same system logically but distribute in the different physical structure.

The distributed and parallel processing system has become an important field in the information processing subjects and is under rapid development. The reasons are as follows[7][10]:

(1) It can resolve the problem that the organization is dispersed but the data between them often need to be communicated. Such as the bank system, the head office and the branch office which located in the different cities or the different regions in one city. They need to transact respective data in business and also need the communication and the process each other.

(2) With the extension of the enterprise and the aggrandizement of the business data, the requirement for the rapid processing ability is more and more high.

(3) If an organization wants to increase new relatively independent units to extend its scale, the distributed and parallel processing system has the lowest influence to the current system.

(4) The demand for the equipoise loading. The decomposition of the data results in the huge part application. The loading shared by several processors could avoid appearing the critical bottleneck problem.

(5) The distributed and parallel processing database system is not lower than the concentrated database system with the same scale in the broken- down appearance rate. But its reliability is much higher from the aspect of whole system.

(6) If several database systems had existed in the current framework and the necessity of implementing the whole application is increased, we could make use of them to compose a distributed and parallel processing database system from the bottom to the top.

The characteristics of the distributed parallel processing database system are as follows[8][9]:

(1) The system doesn't emphasize the conception of the concentrating control. It has a delaminating control structure based on the whole database, but each local database system has the high independence.

(2) The distributed transparence. We will feel that the data is not dispersed and the diversion doesn't influence the validity of the procedure.

(3) Although the huge data is distributed in the different nodes or computers, the parallel processing mechanism increase the speed of every component.

(4) The data redundancy is considered to be the necessary characteristic because we can increase the local application through duplicating the data in some nodes and increase the validity through operating the replicated data in other nodes when one node is damaged.

6. THE IMPLEMENTING OF THE DISTRIBUTED PARALLEL PROCESSING SYSTEM

As figure 3 showing, there are many subsidiary companies with their own WEB server and local database system. And there existing a business server in the information center for the long-range data being transferred. Because the companies and the database system distribute the different places and each of them has necessary and enough data processing logic, so we can use a distributed parallel processing database system to implement it.

The following is the brief analysis of the working process

(1) For the local data (the local database system) transferring, the browser will send the request to the local WEB server and then the SERVLET or JSP existed in the WEB server will call the related EJBs in the local EJB server according to the request. The EJB server may include several session beans and entity beans and then it will choose the most appropriate session bean or beans to deal with the request. Then the session beans will call the relevant entity beans through the local home interface, get the final results and write them into the JSP pages to return to the client.

(2) For the remote data transferring (the long-distance database system), after receiving the request from the WEB server, the local session beans will connect to the business server located in the information center and send the request to its relevant session beans. So the business server could contact other company's EJB server in the light of the request. The process in other EJB server is similar to term (1). At last, the request will be returned to the initial WEB server and client.

(3) For both the local data and remote data transferring, term (1) and term (2) will be all executed. The local final data and the long-distance final data will be combined to return to the client.

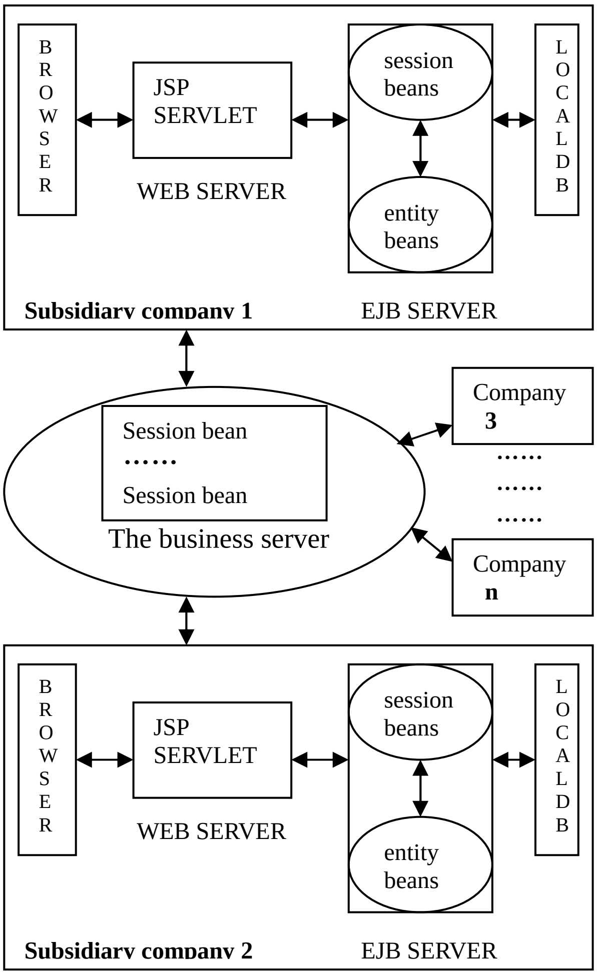


Fig. 3. the module of distributed parallel processing DB system

7. CONCLUSIONS

The appearance of the J2EE framework particularly the EJB architecture provides a very good tool to construct the distributed and parallel processing applications. And the database system based on it has resolved many problems in data, computing and resource which are difficult to settle in the past. The parallel processing mechanism assures the speed of it and the distributed processing mechanism adapts to the requirement of the dispersed and huge data quantity. So the distributed and parallel processing system will be the direction of the database system and applications in the future.

REFERENCES

[1] http://java.sun.com.
[2] http://mars-sim.sourceforge.net.

[3] Andrzej Duda, “Analysis of Multicast-Based Object Replication Strategies in Distributed Systems,” In Proceedings of the 13th International Conference on Distributed Computing Systems, Pittsburgh, USA, pp311-318, 1993.

[4] George Coulouris et al., “Distributed Systems Concepts and Design (3rd ed.),” Addison-Wesley, 2000.

[5] Wei JIE Wentong CAI Stephen J. TURNER “Dynamic Load-Balancing Using Prediction in a Parallel Object-oriented System” 2001 *IEEE.*

[6] K. Fritsch, J. Power, and J. Waldron. “A Java distributed computing library.” In 2nd International Conference on Parallel and Distributed Computing Applications and Technologies (PDCAT2001), pages 236–243, Taipei, Taiwan, July 2001.

[7] M. Wang, S. Zheng, W. Zheng. JDCS: “A Distributed Computing System Implementing High Performance Computing. Computer Engineering and application” Vol.21, 2002. (in Chinese).

[8] LI Xiaozhou, LI Qinghua.” Implementation of Matrix Multiple Cannon Parallel. ComputerEngineering,” June 2002 (in Chinese).

[9] Zhang Ri, Zhang Xiang. “The data replicated technology in the distributed database system.” Journal of Wuhan University of Technology (Information & Management Engineering) pages 24–26, May 2003.

[10] Wang Yijie, Jiang Xueyang, “Research of Replication Techniquesin Internet Distributed Storage System, “Journal of Computer Research and Development, 40(Suppl.):30-35, 2003.

[11] T. Keane, R. Allen, T. J. Naughton, J. McInerney, and J. Waldron.” Distributed Java platform with programmable MIMD capabilities.” In N. Guelfi, E. Astesiano, and G. Reggio, editors, Scientific Engineering for Distributed Java Applications, volume 2604 of Springer Lecture Notes in Computer Science, pages 122–131, Feb. 2003.

Yetian Li is a master degree candidate in the School of Computer Science and Technology, Wuhan University of Technology. He majors in computer application and his research interests are J2EE and network security.

Applying Web Services-based SOA to XML-based Network Management *

Shisong Xiao, Yan Xu, Hui Xu, Zhixia Zhao
Dept. Computer Science and Technology
HuaZhong Normal University, Wuhan, 430079, P.R.China
Email: xuyan720@mails.ccnu.edu.cn

ABSTRACT

Since the network management research and standardization in late 1980s, several approaches have been applied, from protocol-based ones, to Web-based ones, and recently, XML-based ones. These days, Web services, in particular, have been emerging as a promising XML-based architecture consisting of several XML-based technologies, and it seems that Web services may be used in the field of network management. In this paper, we survey the key aspect of XML-based approach to network management. We then examine Web services as a XML-based approach to network management, and provide standardization for management information definition and access offered by XML-related technologies. Since Web services are a Service-Oriented Architecture (SOA) more than just a set of technology, Web services could be used in XML-based network management at the SOA level recur to the concept of Peer-to-Peer. Correspondingly, a prototype is provided to further demonstrate the potential of Web services-based SOA in XML-based network management.

Keywords: Network Management, XML, Web Services, SOA

1. INTRODUCTION

Network management research and standardization started in the late 1980s, but now, 16 years past, there is still no such a framework and technology that satisfies the general needs of network management. Initial approaches were protocol-based, such as open system interconnection systems management (OSI-SM), which is too complicated to implement such a system, and Simple Network Management Protocol (SNMP), which is too simple to realize some basic needs of network management. With the popularity of World Wide Web, Web-based approaches were then adopted in network management. Recently, more attention is paid to XML-based approaches, and in particular, as a standard based on XML, Web services seem to be also appropriate for network management.

Some organizations have participated in the research of using Web services in network management, such as the Organization for the Advancement of Structured Information Standards (OASIS) [1], which is a consortium that produces more Web services standards than any other organization along with standards for security, e-business, etc., and the Network Management Research Group (NMRG) [2] of the Internet Research Task Force (IRTF), which has discussed web services technologies, and compared them with SNMP. In addition, several researchers have also studied the standardizations and prototypes for applying Web services to network management [3], [4]. And only few investigators have ever considered its potential in management [5].

In most studies, the real potential of Web services in the network management domain has rarely been discussed. However, as a XML-based standardization, Web services have many advantages offered by XML and its related technologies, all of which facilitate its use in the field of network management. What is more, since Web services is an Service-Oriented Architecture (SOA) more than just a set of technologies, a study at the SOA level may seem more suitable in order to make full use of it. The aim of this paper is to provide the advantages of Web services in network management from the XML's point of view, and based on the current development of XML-based network management, to apply Web services at the SOA level recur to the concept of Peer-to-Peer.

The organization of the paper is as follows. First, an overview of XML-based network management will be given in Section 2. Then in Section 3, we will present a brief introduction of Web services, which includes its definition and architecture (a kind of SOA). Section 4 will describe the advantages of Web services in network management, focusing on the benefits offered by XML and its related technologies, including Web Services Description Language (WSDL) and Simple Object Access Protocol (SOAP) over HTTP. Sequentially in Section 5, we will discuss XML-based network management using Web services at the SOA level, recur to the concept of Peer-to-Peer, and a prototype will then be given. In addition, the benefits of the prototype will be analyzed. Finally, we conclude our work and discuss directions for future in Section 6.

2. AN OVERVIEW OF XML-BASED NETWORK MANAGEMENT

Today, XML-based network management, which applies XML technologies to network management, has been regarded as an alternative to existing network management, especially to SNMP.

Extensible Markup Language (XML) [6] is a meta-markup language standardized by the World Wide Web Consortium (W3C) for document exchange in the web. Nowadays, XML-based specifications, such as XML Schema [7] [8] [9], Document Object Model (DOM) [10], Simple API for XML (SAX) [11], XML path language (XPath) [12], Extensible Stylesheet Language (XSL) [13], XSL transformations (XSLT) [14], Simple Object Access Protocol (SOAP) [15] [16] [17], and Web Services Description [18] [19], are widely applied in network management. What is more, XML is now a standard that is supported and accepted by thousands of vendors as well as a lot of related technologies and tools, which are illustrated in Fig.1.

*This study is supported by the key project funded by the Ministry of Science and Technology of Hubei (2004AA103A01).

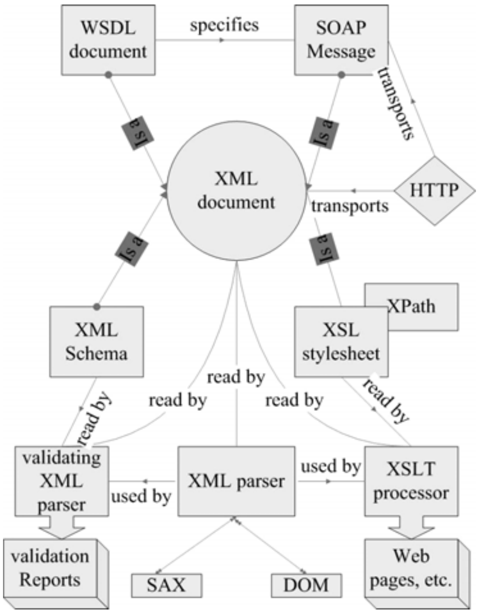


Fig.1. XML technologies and tools

The advantages offered by the use of XML in network management are presented detailedly in [20]. Some key points are listed as follows:

·Management data can be represented as XML documents.

·The structures of management data can be expressed as XML Schemas.

·The DOM and SAX APIs can be used to access management data from applications.

·XSLT can be used to process management data.

·Widely deployed protocols, such as HTTP, can be used to ship the data.

·High-level management operations can be defined through WSDL and called via SOAP.

3. WEB SERVICES BACKGROUND

XML-based technologies are now widely used in network management, and particularly, Web services have been emerging as a promising Internet-oriented technology for network management [21].

3.1 Web Services Definition

Web services is developed and standardized by the World Wide Web Consortium (W3C), which gives the following definition [22]: "A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards".

The beginning of that definition, "Web Services are software applications", conveys a main point: Web services are software systems available on the Web that perform specific functions. And its purpose is to support interoperable machine-to-machine interaction over a network, in other words, the machine-aware part of the network, especially the Internet. In order to implement its designed purpose, a machine-processable format (specifically WSDL) is provided for the description of a Web service, and SOAP-messages, typically conveyed using HTTP, are used for interaction between a Web service and another system (or another Web service). In addition, other Web services technologies such as Universal Description Discovery, and Integration (UDDI) [23] and electronic business XML (ebXML) [24] registries, allow applications to dynamically discover information about Web services.

3.2 Service-Oriented Architecture (SOA)

The word "services" in Web services refers to a Service-Oriented Architecture (SOA) [25]. In fact, SOA is a recent development in distributed computing, in which applications call functionality from other applications over a network. In an SOA, functionality is "published" on a network where two important capabilities are also provided – "discovery", the ability to find the functionality, and "binding", the ability to connect the functionality. So when considering a SOA, these three parts must be take into account, which are briefly presented as "publish", "find", and "bind".

In the Web services Architecture, three important roles are Web service provider, Web service requester, and Web service register, which correspond to the "publish", "find", and "bind" aspects of a SOA. The Web services-based SOA, combined with related technologies, is depicted in Fig. 2.

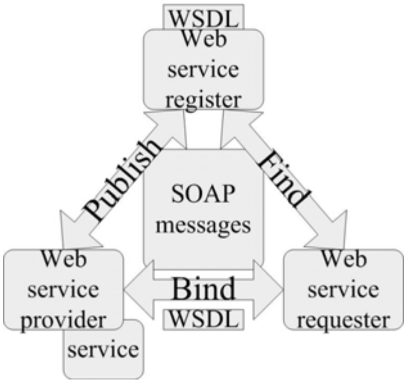


Fig.2. The Web services-based SOA

4. THE POTENTIAL OF WEB SERVICES IN XML-BASED NETWORK MANAGEMENT

As one of the emerging standards based upon XML, Web services are a generic technology, for the simple reason that it is wholly XML-based. And the use of standard XML protocols or technologies makes Web services platform-, programming language-, and vendor-independent. The support of XML technologies, such as WSDL and SOAP, provides the capability for the standardizations of management information definition and access, which are very important in network management. Thus in this way, Web services provide a distributed management capability for monitoring the services of applications on the Internet or intranet using standard XML protocols and formats.

4.1 Standardization of Management Information Definition by WSDL

A Web service is described in a WSDL document. In order to easily use Web services for network management， standardization of management information definition is needed. On the other hand, with the mechanism provided by WSDL to describe a Web service in a modular manner using the elements <import> and <include>, modularization can be achieved.

With regard to the division of WSDL documents, we propose the “three separate WSDL definitions” suggested by J.Sloten et al [3], for it seems more reasonable according to the functionalities of each element. It contains an abstract part: the messages and interfaces (the “what” part) and two concrete parts: a binding (the “how” part) and a service (the “where” part) with WSDL import mechanism.

The <types> element can be omitted in the WSDL document and if not, there are two types by the means of XML Schema: <simpleType> and <complexType> elements. The <message> element is used to describe the information being exchanged between a Web service provide and a Web service request. It consists of zero or more <part> elements, which can be associated with the type that defined in the <types> elements. The <part> elements describe data for exchange, mainly used in two ways: (a) the parameters of a Web service including transparent ones (parameter transparency [3] makes the management information abstracted from protocol level) and non-transparent ones, which make operations easy to understand and develop; and (b) the return data. The <operation> element defines the input and output messages with <input> and <output> elements, while the <interface> element is just the set of <operation> elements. Operation granularity [3], the level of variation between very coarse and very fine operations, is always needed to consider in the practice of network management, and it is, in essence, just a “tradeoff” problem.

The <binding> element provides concrete information on what protocol is being used for the Web service, and how data is encoded and transported [4]. Since it offers the “style” functionality of a Web service, the standardization of which can make a default protocol available for operations and a standard encoding for messages, it may be more appropriate to be split up from the “message and interface” part. The <endpoint> element specifies the position to access a particular Web service as well as the protocol used for this purpose, while the <service> element is just a set of <endpoint> elements, which means that, each Web service can be accessed by endpoints.

4.2 Standardization of Management Information Access by SOAP over HTTP

According to related specifications, SOAP focuses on the basic forms of transporting messages, regardless of the transfer protocols. Since most of the common transfer protocols on the Internet adopt the message intercommunication model, a natural thought is to regard SOAP messages as the contents of these protocols, which transports these SOAP messages.

Considering the fact that most applications are through HTTP, the combination of SOAP messages and HTTP is undoubtedly the most practical and widely used means to implement interconnection of services. In this way, SOAP over HTTP, which supports its own RPC interfaces, becomes a natural application protocol for network management and this default transport scheme provides a standardization of management information access. In fact, it has been gradually used in the current network management, such as Netconf [26]

5. A PROTOTYPE AT THE SOA LEVEL

5.1 Some Considerations

XML-based network management systems have become more and more popular these days, for it applies XML technologies to network management. However, these systems differ much in the extent of using these XML technologies. Most of these systems just use a few simple XML technologies, while some do make a better use of the XML technologies, such as the system presented in [27].

As to current attempts to use Web services in XML-based network management, the usage level of Web services must be taken into account. Since Web services are a SOA more than a set of XML-based technologies, it is reasonable to distinguish the XML-based network management using Web services at two levels: the technology level and the architecture level. Recent studies focus more on the technology level, but in order to make full use of Web services, it seems to be more appropriate to focus on the architecture level, or more exactly, the SOA level.

5.2 The Prototype

In Section 3, Fig. 2 has offered the Web services architecture as a Service-Oriented Architecture (SOA). Bear in mind, SOA and web services are not just abstract concepts, but are real approaches to solving network management problems. A number of widely adopted web services technologies are available today, such as XML, UDDI, SOAP and WSDL.

As a seamless integration of XML-based network management and Web services-based SOA, Fig. 3 illustrates a prototype at the SOA level, which each entity in the distributed environment, can act as a manager role (a Web service provider) or an agent role (a Web service request), or both.

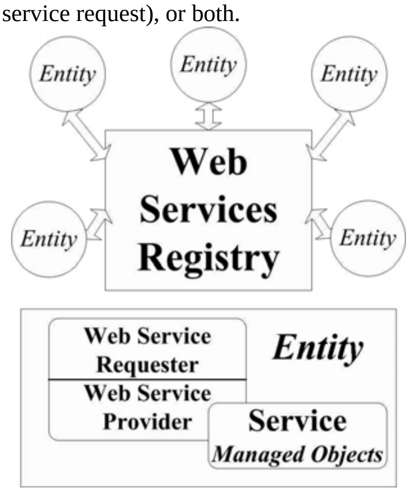


Fig.3. A prototype at the SOA level

As is shown in Figure 3, two main components in the prototype are the Entity and, the Web Services Registry, which can be a private one for a particular network management task. The use of entities based on the Web services-based SOA, in fact, utilizes the concept of Peer-to-Peer [28].

Peer-to-peer communication is a type of person-to-person communication, which is distinguished from the client-server communication. In this form, individuals who form a loose group can communicate with others in the group. What makes peer-to-peer systems interesting is that they are totally distributed and all nodes are symmetric. In a typical peer-to-peer system, the users each have some information that may be of interest to other users. If there are large numbers of users, they will not know each other and will not know where to find what to find what they are looking for. One solution is a big central database, which in this prototype, is the Web Services Registry. In addition, point-to-point link provides an effective connection means for peer-to-peer communication.

With the very use of the concept of Peer-to-Peer, the working flow of this prototype is as follows.

First of all, these entities, each of which acts as a Web service provider in this model, register their own services of some managed objects to the Web Services Registry. When another entity, acting as a Web service requester, wants to acquire the information of one or more managed objects provided by one entity, it just needs to connect the Web Services Registry to get the access information to that entity, the procedure of which is based on the Web services and the peer-to-peer communication. To accomplish the point-to-point link, the entity can acquire the service through SOAP over HTTP or another transfer protocol.

5.3 The Benefits from the Prototype

In the following, the main advantages of this prototype will be discussed in detail.

As is known to us, a management process in OSI-SM architecture can be configured to act in either or both roles. So the use of Web services can bring XML-based network management closer to OSI-SM, which is so far the strongest network management architecture that supports all the essential features in any management framework.

What's more, Web services-based SOA [29] can largely and effectively improve the current network management systems.

On one side, SNMP and some other network management approaches often suffer from potential scalability problems for large managed object populations. However, one primary feature of Web services-based SOA is just scalability, which can properly solve the underlying problem in network management. Since services in an SOA are loosely coupled, applications that use these services tend to scale easily and certainly more easily than applications in a tightly coupled environment. The reason for this is obvious that there are few dependencies between the requesting application and the services it uses.

On the other side, flexibility is always very important in the network management related applications. Loosely coupled services in the Web services are typically more flexible than tightly coupled applications. In a tightly coupled architecture, different components of an application are tightly bound to each other. This makes it difficult to update the application to meet the changing requirements. In contrast, the flexibility of Web services-based SOA, which is provided by the loosely coupled, document-based, asynchronous nature of services, allows applications to be flexible, and easy to evolve with changing requirements. Thus in this way, the network management systems using Web services can be flexible in dealing with ever changing management needs.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented issues related to the potential of Web services-based SOA in network management with an emphasis on its advantages as a standard based on XML with the help of WSDL and SOAP over HTTP, which realize the standardization of the network management information definition and access respectively. Then, believing that Web services are a SOA more than a set of XML-based technologies, the paper has provided a prototype of XML-based network management using Web services at the SOA level recur to the concept of Peer-to-Peer.

Our study indicates that Web services are suitable in network management, and with the popularity of XML in the field of network management, one reasonable way to use its capability is XML-based network management using Web services-based SOA. In order to make full use of Web services, it seems that in XML-based network management, using Web services at the SOA level would be better than that at the technology level.
Further work is needed to implement XML-based network management systems using Web services at the SOA level and performance tests are needed to compare its capability with that of currently popular XML-based network management systems.

REFERENCES

[1] "Organization for the Advancement of Structured Information Standards (OASIS)," http://www.oasis-open.org/home/index.php.

[2] *IRTF*, "Network Management Research Group (NMRG)," http://www.ibr.cs.tu-bs.de/projects/nmrg/.

[3] J.Sloten, A.Pras, M.Sinderen, "On the Standardisation of Web service management operations", Proc. 10th Open European Summer School (EUNICE 2004) and IFIP WG 6.3 Workshop, June 2004, pp. 143-150.

[4] T.Drevers, R.Meent, A.Pras, "Prototyping Web Services based Network Monitoring", Proc. 10th Open European Summer School (EUNICE 2004) and IFIP WG 6.3 Workshop, June 2004, pp. 135-142.

[5] G.Pavlou, P.Flegkas, S.Gouveris, A.Liotta, "On Management Technologies and the Potential of Web Services", *IEEE Communication Magazine*, July 2004, pp. 58-66.

[6] W3C, "Extensible Markup Language (XML) 1.0 (Third Edition)",*W3Recommendation*, http://www.w3.org/TR/2004/REC-xml-20040204, February 2004.

[7] W3C, "XML Schema Part 0: Primer", *W3C Candidate Recommendation* http://www.w3.org/TR/2000/CR-xmlschema-0-20001024/, October 2000.

[8] W3C, "XML Schema Part 1: Structures Second Edition", *W3C Recommendation* http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/, October 2004.

[9] W3C, "XML Schema Part 2: Datatypes Second Edition", *W3C Recommendation* http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/, October 2004.

[10] W3C, "Document Object Model (DOM) Level 1 Specification", *W3C Recommendation* http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001/, October, 1998.

[11] Simple API for XML (SAX), http://www.saxproject.org/

[12] W3C, "XML Path Language (XPath) 2.0", *W3C Recommendation,* http://www.w3.org/TR/2005/CR-xpath20-20051103/, November 1999.

[13] W3C, "Extensible Stylesheet Language (XSL) Version 1.0",*W3CRecommendation,* http://www.w3.org/TR/2001/REC-xsl-20011015/, October 2001.

[14] W3C, "XSL Transformations Version 1.0", *W3C Recommendation,* http://www.w3.org/TR/1999/REC-xslt-19991116, November 1999.

[15] W3C, “SOAP Version 1.2 Part 0: Primer”, *W3C Working Draft,* http://www.w3.org/TR/2001/WD-soap12-part0-20011217/, December 2001.

[16] W3C, “SOAP Version 1.2 Part 1: Message Framework”, *W3C Working Draft,* http://www.w3.org/TR/2001/WD-soap12-part1-20011217/, December 2001.

[17] W3C, “SOAP Version 1.2 Part 2: Adjuncts”, *W3C Working Draft,* http://www.w3.org/TR/2001/WD-soap12-part2-20011217/, December 2001.

[18] W3C, “Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language”, *W3C Working Draft,* http://www.w3.org/TR/2005/WD-wsdl20-20050803, Aguest 2005.

[19] “Web Services Description Language (WSDL) Version 2.0 Part 2: Message Exchange Patterns”, *W3C Working Draft,* http://www.w3.org/TR/2004/WD-wsdl20-patterns-20040326, March 2004.

[20] F. Strauss, T. Klie, “Towards XML Oriented Internet Management,” *Proc. IFIP/IEEE* Int’l Symp. on Integrated Network Management (IM 2003), Colorado Springs, USA, Mar. 2003, pp.505-518.

[21] J. Schonwalder, A. Pras, J.P. Martin-Flatin, "On the Future of Internet Management Technologies", *IEEE Communications Magazine*, October 2003, pp. 90-97.

[22] W3C, “Web services Architecture (8 August 2003)”, *W3C Working Draft,* http://www.w3.org/TR/2003/WD-ws-arch-20030808/\#whatis, August 2003.

[23] UDDI Spec TC, “UDDI Version 3.0.2 UDDI Spec Technical Committee Draft”, *UDDI Spec Technical CommitteeDraft,*http://uddi.org/pubs/uddi-v3.0.2-20041019.htm, October 2004.

[24] “Electronic business XML”(ebXML), http://www.ebxml.org/.

[25] “Service-Oriented Architecture (SOA)” http://www.service-architecture.com/.

[26] T.Goddard, “NETCONF over SOAP”, Draft-ietf-netconf-soap-06, September 2005, work in progress.

[27] M.Choi, W.Hong, H.Ju, “XML-Based Network Management for IP Network”, *ETRI Journal*, Volume 25, Number 6, December 2003, pp. 445-463.

[28] “Peer-to-Peer Architecture,” http://www.webopedia.com/TERM/p/peer_to_peer_architecture.html.

[29] Ed Ort, “Service-Oriented Architecture and Web Services: Concepts, Technologies, and Tools”, http://java.sun.com/developer/technicalArticles/WebServices/soa2.

An Effective Approach of Web Crawling for Deep Web

Shunyan Wang, Binghua Wu, Luo Zhong
Department of Computer Science and Technology, WHUT
Wuhan, Hubei 430070, China
Email: busy_1@163.com

ABSTRACT

One of the most promising new ways of large-scale data analysis is the large and growing collection of Web-accessible database known as the Deep Web - the part that is typically hidden from Web search engines. Unlike the traditional Web, where Web pages are accessible by traversing hyperlinks from one page to another, Deep Web data is accessible by interacting with a Web-based query interface. Considering the problem of poor information coverage in web data mining, the pager proposes a configurable web crawling method for deep web which can improve the results performance of a general search engine significantly. It classifies web pages and manipulates key information of page content in order to make sensible queries. The experimental results also prove it.

Keywords: Web Crawler, Deep Web, Script Crawler

1. INTRODUCTION

With the development of Internet, the number of websites is keep growing day by day. And more and more companies and persons have websites with their own. As a result, the scale of available information is expanding and larger than ever before. In contrast to several years ago when we were living in a state of longing for information, now we are in an era of excessive information, or, the so-called information overload.

When enjoying the benefit and convenience from the Internet, we are also facing a lot of problems. How to find useful and valuable portion of information from those massive web pages is a big challenge. Therefore, both web data extraction and utilization are hot frontiers in information technology nowadays. Meanwhile, big giants such as Google, Yahoo, Baidu, and other search engines are harvesting as much content as possible with which they offer free search service for market preeminence.

2. MOTIVATION

Searching service is the function of Full-text Indexed Search Engine, search engine for short, software who always tries its best to find the right information for you. Fig. 1 illustrates how a general search engine works.

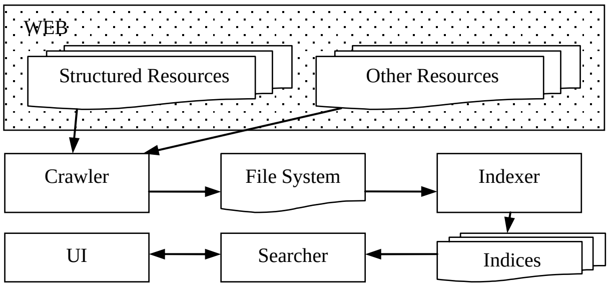


Fig.1. Process Model of General Search Engines

In Fig.1, the web crawler keeps wandering on the web and saves every web page visited as files in a certain format. With these files, the indexer parses and indexes them into indices files for searching subsequently. After that, everything is done and the engine is ready for performing searching. When the user post a query term against the engine, the query analyzer exams the term and looks for matches in the indices files generated by indexer.

The key issue involved in search engine implementation is how to design a smart web crawler since it harvests web pages from the Internet for searching. It's a piece of software which keeps running all day long to visit and collect web pages on the Internet. The page content and other related information is the data which a search engine searches against. So, the quality of results returned by a search engine lies significantly on the eyeshot of a web crawler.

Search engines are friends we can rely on when we are looking for information; however, they are not silver bullet for us. Web pages keep increasing every day, and many pages cannot be found by web crawler. Unfortunately, most of these pages contain valuable information according to investigations. In other words, the low coverage of a web crawler can have a bad effect on the quality of a search engine. The reasons why a crawler unable to visit the entire web are summarized as follows:

First, a web crawler cannot catch up with the growth of resources available on the Internet. There are millions of new pages generated and updated every day. It is too expensive for the crawler to figure out each page. In real-life application scenarios, there is time span between two rounds of crawling. So, it's impossible for a search engine to know every single web page on the Internet.

Second, not all web pages are visible to a web crawler, say; the crawler cannot "see" some pages. Google and other search engines can index the page via crawlers if a link is found to the resource from another web page. The crawler can only discover new pages by extracting uniform resource locators (URLs) for a found page. If there are no links navigating to a page, the page cannot be visited. Generally, crawlers prefer to collecting static web pages, such as HTML page. Although popular search engines are trying to find and index pages generated by script, they are ignorant of most websites containing generated pages.

3. PRELIMINARIES

The Web is becoming a complex entity that contains information from a variety of source types. It is much more than fixed Web pages. In fact, the part of the Web that is not fixed, and is served dynamically "on the fly", is far larger than the fixed documents that many associate with the Web. In addition, many types of data resources are not visible or not identifiable to Web crawlers. We refer to this content as the "Deep Web", or "Hidden Web". When we refer to the deep Web, we are usually talking about the following:

(1) The content of databases accessible on the Web. Databases contain information stored in tables created by such programs as Access, Oracle, SQL Server and DB2. Information stored in databases is accessible only by query. This is distinct from static, fixed Web pages, which are documents that can be accessed directly. A significant amount of valuable information on the Web is generated from databases. In fact, it has been estimated that content on the deep Web may be 500 times larger than the fixed Web.
(2) Pages in non-HTML format. Non-textual files such as multimedia files, graphical files, software, and documents in formats such as Portable Document Format (PDF).
(3) Script-based pages, whose links contain a "?" or other script coding, such as Active Server Pages. These pages do not contain page content, but they make query against database to dynamically generate HTML pages. Data is stored as forms in databases which can be obtained through some queries. This is a clear distinction between fixed pages and dynamically generated pages.
(4) Content available on sites protected by passwords or other restrictions. Some of this is fee-based content, such as subscription content paid for by libraries and available to their users based on various authentication schemes.

The categorized resources above compose a deep Web. We are interested in the second type of deep Web which is script-based pages.

The deep Web concept was initially proposed by Dr. Web. Jill Ellsworth in the last century which refers to those Web pages cannot be found by general search engines easily. Deep Web is defined as: the resource including text, documents and other high-quality information which is accessible on the Internet but not searchable by ordinary search engines for technical limitations or other reasons. Here are some figures about Deep Web:
(1) In 2004, there were more than 3 million of Web sites and 40 thousand of databases for generating HTML pages. This figure keeps growing and it grew by 3-7 times from 2000.
(2) Now, Web crawlers have overcome many technical barriers and made it possible for them to find and provide some deep Web pages. Some popular search engines may cover about one-third of the entire deep Web.
(3) Although some indexing service for deep Web is available, however, it has low coverage of 0.2% to 15.6%.

Compared with fixed web pages, deep web pages contain more high-quality contents which are pulled from databases. At present, many popular search engines are able to index and search against only a part of the deep web. There are many reasons for this, one obstacle is that Search engines still cannot type or think. If access to a web pages requires typing, web crawlers encounter a barrier they cannot go beyond. They cannot search our online catalogs and they cannot enter a password or login.

Unlike pages on the surface Web, Deep Web pages are generated in response to a query submitted through a Web interface. There are millions of new pages generated by a variety of script languages every day, such as Active Server Pages (ASP) and Personal Home Pages (PHP). Both of them are popular server-side scripts nowadays. This paper studies how to design a script-crawler for web pages which are generated by ASPX script. The crawler tries to make reasonable queries against certain pages to collect more pages according to some calculated results.

4. SCRIPT-CRAWLER

One main objective in crawler design is to make the crawler visit and save web pages as more as possible. For the reason that script-based pages can be accessed by query only, the script-crawler should be able to perform queries like human. Two problems should be considered carefully before moving on. First, it must be able to identify pages which may contain potential hidden web based on collected pages. Second, it should submit correct inquiries to visit the hidden website and collect pages subsequently.

In order to identify the web hidden behind normal pages, we define some key characteristics for those pages in which hidden web may exist. Meanwhile, reorganization and analysis of page content can help to calculate suitable query parameters.

In this paper, we use configuration files to describe the pattern of those pages which may have potential web pages. When digesting a normal page, the crawler uses regular expressions in the identification and extraction of query data. After that, the new query data is posted and new pages are returned as page source for the next round of crawling.

4.1 System Framework

In addition to performing some basic actions and functions, the script-crawler can also recognize and analyze some pages generated by scripts.

We now present the overall framework for supporting large-scale data analysis of the Deep Web, with an emphasis on the core data preparation modules, as illustrated in Fig. 2.

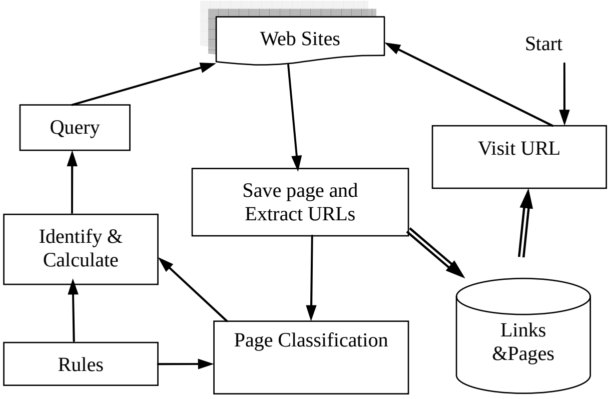


Fig.2. Overall Framework

The crawler starts with an initial URL, which is called root URL, then it visits and saves this page in local file system. After that, the crawler extracts all links from the saved page and update links database for next crawl. This is the way in which a general crawler works. However, the script-crawler does not visit all of the links in the page. Instead, it tries to parse the page content and look for Deep Web characteristics.

As shown in Fig.2, the script-crawler classifies the saved pages according to page content, and then it calculates the query parameters and submits them to web server. The calculation process may vary with page content. If the crawler succeeds to make the calculated query, then it will receive result pages which are generated scripts running on the web server. Otherwise, it will abort the operation and try the next link.

4.2 Page Classification

For efficiency concern, the crawler should categorize pages before calculating required parameters for query. This operation refines the returned page content and filters useful pages which may contain Deep Web. The process of grouping and digesting Web pages is illustrated below:

First of all, to group pages correctly and effectively, the page should be well-formed. The requirements of a well-formed page include, but are not limited to, the following: all start tags, including standalone tags, must have a matching end tag; all attribute values must be in quotes; tags must strictly nest. Pages that do not satisfy these criteria are automatically transformed into well-formed pages using JTidy(http://tidy.sourceforge.net).

Second, the crawler will extract all possible links out of the saved Web pages to look for candidate URLs according to a configuration file. The file consists of a group of regular expressions which denote the pattern of links from which the crawler will look for Deep Web. For instance, the regular expression “http://news.xx.com/show.aspx.*” represents all links which start with the prefix “http://news.xx.com/show.aspx”.

4.3 Definition And Extraction of Page Pattern

The script-crawler can identity Deep Web according to the pattern found in saved Web pages. So, it is very important to define the pattern of Web pages properly which may contain Deep Web. We collect a set of Web pages from different Web sites which contain Deep Web pages, and then examine the structure of page content to extract a pattern. The pattern is refined and saved into pattern database for future use. Experimental results show that it works fine.

The pattern database stores necessary parameters required to define a page pattern. The parameters describe HTML fragments in a Web page which query data is calculated against. For example, a pattern may define the structure and position of a piece of HTML tags within a document. These patterns are collected manually. We found that pages generated by a certain script have something in common. For example, Web pages generated by ASPX script usually have some well-known hidden fields. In addition, it is more extensible and flexible to make the pattern configurable. For instance, most of ASPX generated pages have a JavaScript function named __doPostBack, and there is a record in pattern database which represents the HTML fragment containing this function. And more importantly, this can be changed easily anytime as needed. The sample configuration file demonstrates the pattern for identifying some pages which contain paging-enabled datagrid control in ASPX.NET. ASP.NET contains three data Web controls—the datagrid, datalist, and repeater—each designed to allow for rich data display. The most commonly used of the three data Web controls is the datagrid, due in large part to its handy built-in feature set. A pageable datagrid posts back to the server every time when the user navigates from one page to another through a pager element. The trick is that by clicking on the pager you originate a postback event which consists merely of a form submission. The submitted arguments contain information about the current page index of the datagrid and the next page requested. For it is versatile in its features and popularity in Web development, we are interested in the Web pages generated by ASPX.

```
        <suffix>aspx</suffix>
        <category>paging</category>
        <tag>table</tag>
        <offsets>
                <offset value="1"/>
        </offsets>
        <targetCtrlRegex>
                <startBlock></span>&nbsp;<a
href=\"javascript:__doPostBack('</startBlock>
                <endBlock>','')\"></endBlock>
                <ctrlStart>DataGrid1$_ctl44$_ctl0</ctrlStart>
                <pager>DataGrid1$_ctl44$_ctl</pager>
        </targetCtrlRegex>
```

Fig.3. A Sample Configuration File

The file in Fig.3 defines the pattern for HTML fragment with its location and content information. This XML formed file denotes which level the pattern tag, which is the <table> here, is located and what the target tag is. This information is used by script-crawler to calculate query parameters used to request for a specific page of content.

4.4 Perform Query

The script-crawler will calculate suitable query string when it succeeds to find the fragment of HTML with the help of a pattern. However, it is difficult to submit a perfect query. Authors[3] proposed a subject based method to generate keywords for query. This approach does not work here for there is no keyword available. The script-crawler extracts data from saved pages and calculates suitable values for next query. It will set proper values for hidden fields in the page and submit the form.

The script-crawler looks for HTML fragments in saved Web pages with a given pattern. When it finds a match, the crawler will extract all of the values and calculate correct new values for the next query. The algorithm is illustrated as below.

```
procedure submit()
while(Q is available AND the last query succeeded) do
R = sendRequest(Q)          //submit the query
saveResponse(R)       //save the page in local file system
Q = parseQuery(R)           //analyze the page
Q = nextQuery(Q)            //calculate parameters for the
next query
done
```

Fig.4. Algorithm of Query Calculation

Based on the algorithm shown in Fig. 4, the crawler is able to know the page index of the datagrid control in the context of this HTML fragment. With this in mind, it calculates and sets values of controls required for the next query. In general, the next query will bring back more results from Web server and sustains the data mining process. When it is failed to make the query, the crawler will ignore the error and abort the process.

5. EVALUATION

We have tested the crawler against a list of Websites which may contain Deep Web. Experiments show that the proposed crawler can effectively improve the coverage of general search engines. Take the WHUT News (http://www.wutnews.net/) as an example – most pages are generated by ASPX script and a number of pages cannot be visited through navigating, say, there are no in-links for these pages. These pages are only available by issuing queries against search interfaces provided by the Website itself. Meanwhile, the paging-enabled datagrid control is widely used in this site which results in more post-back operations required when the client interacts with the server.

For ease of comparison, we conducted full-text crawling and indexing against the whole site with a general crawler, which is borrowed from Nutch project (http://lucene.apache.org/nutch/), and the script crawler separately. With the generated indices, we perform searches and examine returned results to measure the recall and coverage of the two crawlers. The indices harvested by two crawlers are listed in Table 1.

Table 1. Comparison of Indices

crawler	time cost	size of indices
nutch crawler	~12hours	16,995,925 Bytes（~16.2MB）
script crawler	~16hours	24,054,577 Bytes（~22.9MB）

Table 1 shows that the index file generated by script crawler is 1.4 times that of Nutch crawler. Test on search also proves that the search engine using script crawler generated indices returns more results with given query words. Table 2 and Figure 3 are results for testing search performance of the two cralwers.

Table 2. Comparison on Search Results

crawlers / query	nutch crawler	script crawler
education	3198 records	4401 records
activity	6615 records	9259 records
festival	1427 records	1912 records
intern	1017 records	1366 records
meeting	3126 records	4313 records
Total	**15383**	**21251**

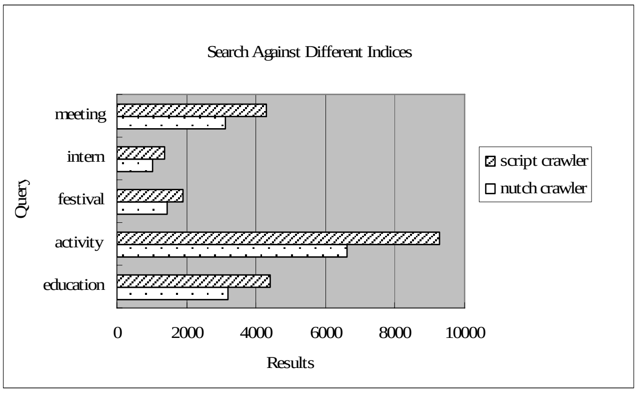


Fig.5. Search Against Different Indices

Fig.5 shows the script-crawler outperformed nutch crawler in success rate (or coverage). And the script-crawler can sniff more Web pages than nutch crawler.

6. CONCLUTIONS

Web crawling has succeeded to date in part because it provides best-effort coverage of a Web site's resources while requiring little to no cooperation from the Web site being indexed. However, search engines must become increasingly complex and aggressive to discover the growing number of resources in the Deep Web. To address this problem, we're developing a variety of tools and techniques to communicate to Web crawlers the existence of structured access to a Web site's resources. We are looking for the weakness of general existing Web crawlers and developing a script crawler which can greatly improve the search engine coverage.

REFERENCES

[1] Laura Cohen.The Deep Web.Internet Tutorial.2006.
[2] http://www.internettutorials.net/deepweb.html.
[3] Zheng dong-dong,Cui zhi-ming. "On research of deep web crawler's crawling strategy".*Computer Engineering and Design*,2006,27(17):3154
[4] ChengFei,Wang Jian-hai,Luo Jian. "MIU:A Web Crawler Method of Increment Updating".*Computer Engineering & Science*.2006,28(12):28
[5] UC Berkeley - Teaching Library Internet Workshops. Invisible Web: What it is, Why it exists, How to find it, and Its inherent ambiguity.
[6] http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html
[7] Erik Hatcher, Otis Gospodneti. *Lucene in Action*. US: Manning, 2004
[8] WU lihui,WANG Bin,YU zhihua. "Design and Realization of a General Web Crawler".*Computer Engineering*

Integrating J2EE Multi-patterns in Development of Enterprise Applications[*]

Jingli Zhang, Xuezhong Qian, Li Mao
School of Information Technology, Southern Yangtze University
Wuxi, Jiangsu 214122, China
Email: wwddyyzzjjll@163.com

ABSTRACT

For most developers of enterprise applications, there seems to be a lack of understanding of how to integrate J2EE patterns in best practices to design enterprise applications. In this paper, an example was chosen to discuss how to apply J2EE patterns in combination to form a better, more robust solution. Result shows that using J2EE patterns can optimize system performance remarkably and enhance numerous aspects of the system, including maintainability, extensibility, and reusability.

Keywords: J2EE, J2EE Patterns, Enterprise Application, Business Component, Online Ordering System

1. INTRODUCTION

The Java 2 Platform, Enterprise Edition (J2EE) provides the standards and technologies for building and deploying multi-tiered enterprise applications. The J2EE platform simplifies enterprise applications with standardized, modular components, services and containers, and improves development productivity by handling many details of application behavior automatically without complex programming.

J2EE is certainly an important platform, enabling teams to build some very powerful enterprise applications. However, reality is, there is still a wild semantic gap between the abstractions and services that J2EE provides and the final application that a team must build. J2EE patterns represent solutions that appear again and again in filling that gap[1]. Rather than applying J2EE patterns in isolation, a complete enterprise application can be composed of J2EE Multi-patterns. Appling J2EE multi-patterns to the system can significantly ease the development process of J2EE applications and improve the quality of the produced software. This paper chooses a B2B online ordering system to discuss how to integrate J2EE multi-patterns in best practices to design robust, efficient, reusable enterprise applications.

2. OVERVIEW OF J2EE PATTERNS

The design patterns[2][3] offer flexible solutions to common software development problems. Each pattern is comprised of a number of parts, including purpose/intent, applicability, solution structure, and sample implementations.

The J2EE patterns describe typical problems in software design for the J2EE platform, and provide solutions that have been used over and over again to solve similar problems at different times in different projects for these problems. Thus, J2EE patterns provide a powerful mechanism for reuse, helping developers and architects to avoid reinventing the wheel. J2EE patterns also allow better communication between software designers, because they provide a common vocabulary and format. Since the design pattern itself already encapsulates the structure and properties of the design, and comprehensive documentations are also available. This makes expressing the system design to other designers more clearly. It also helps less experience designers to speed up their learning curve by reading those design pattern literatures.

A J2EE platform is a multi-tiered system, which has been modeled as five tiers shown in Fig.1:

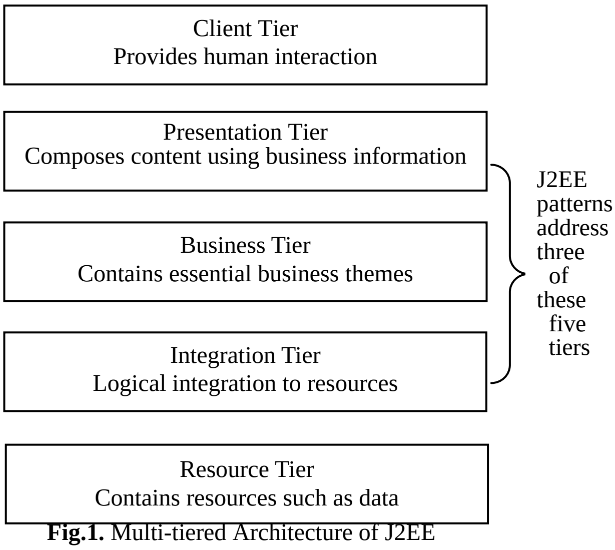

Client Tier
Provides human interaction
Presentation Tier
Composes content using business information
Business Tier
Contains essential business themes
Integration Tier
Logical integration to resources
Resource Tier
Contains resources such as data
J2EE patterns address three of these five tiers

Fig.1. Multi-tiered Architecture of J2EE

1) The client tier, which represents all device or system clients accessing the system or the application, can be a Web browser, a Java or other application etc.
2) The presentation tier, which encapsulates all presentation logic required to service the clients that access the system, is implemented using Servlets, JSPs and HTML/WML pages.
3) The business tier, where core business mechanisms are implemented, is usually encapsulated in enterprise bean components.
4)The integration tier, which is responsible for communicating with external resources and systems such as data stores and legacy applications, can work with the resource tier by using JDBC, J2EE connector technology, or some proprietary middleware.
5) The resource tier contains the business data and external resources.

The J2EE patterns have been categorized into three of these

[*] This work has been supported by Wuxi Technology Burean under grant number CI03003.

five tiers: presentation, business and integration. The presentation tier patterns contain the patterns related to servlets and JSP technology. The business tier patterns contain the patterns related to the EJB technology. The integration tier patterns contain the patterns related to JMS and JDBC.

3. INTEGRATED APPLICATION OF J2EE PATTERNS

The B2B online ordering system is a subsystem of Wuxi public online information services developed specifically for textile industry. The online ordering system provides fast, effective and safe solutions for ordering textile products for a large number of the enterprise customers. The system contains on-line shopping, backstage management, negotiating and on-line help subsystems.

A multi-tier J2EE deployment configuration shown in Fig.2, which contains separate Web and EJB servers, is selected in this system.

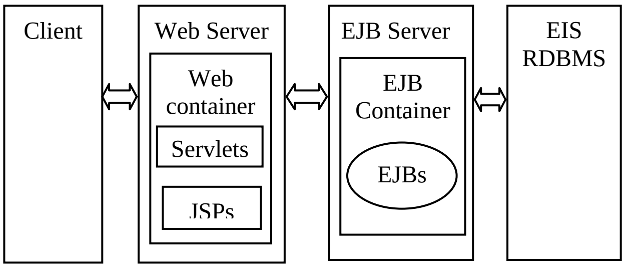


Fig.2. A Multi-tier J2EE Deployment Configuration

We use Service to Worker pattern in presentation layer, use Data Access Object pattern in integration layer, and in business layer, we use the following patterns: Business Delegate pattern, Session Facade pattern, Application Service pattern, Service Locator pattern, Transfer Object pattern and Business Object pattern. Fig.3 shows the J2EE multi-patterns framework for the B2B online ordering system.

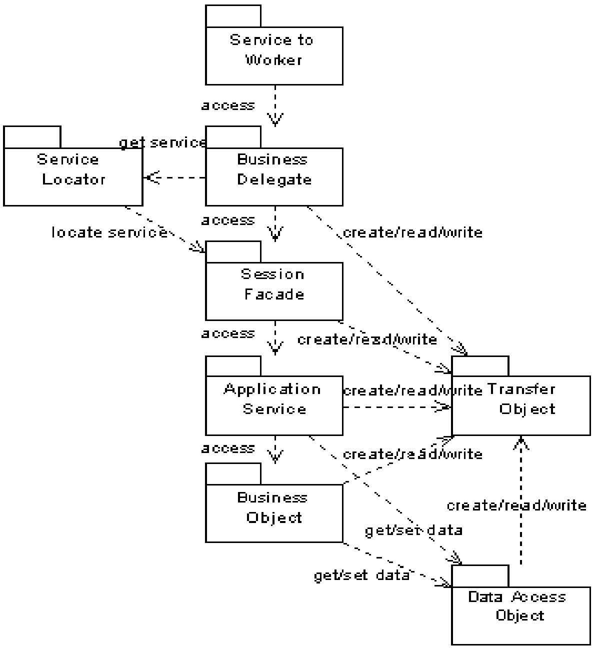


Fig.3. The J2EE Multi-patterns Frameworkfor the B2B Online Ordering System

Due to the length limit of this paper, only J2EE business tier patterns are discussed here.

3.1 Introduce Business Object

We use the term conceptual model to mean the abstract model, which mainly describes domain entities, their relationships and business rule. To describe a concrete object-oriented implementation model of a conceptual model, we use the term object model, which describes the classes and relationships used to realize a conceptual model.

However, if a conceptual model includes a variety of business behavior and relationships, implementing such applications using a procedural approach causes the following problems:

1) Reusability is reduced and business logic code gets duplicated.
2) Bloated procedure implementations that become lengthy and complex.
3) Poor maintainability due to duplication and because business logic is spread over different modules.

Business Objects are used to solve these problems by separating business data and logic using an object model. For example, in this system, Customer, Product and Order can be recognized as Business Objects. Using Business Objects has the following benefits:

1) Centralizes business behavior and state, and promotes reuse.
2) Avoids duplication of code and improves maintainability of code.
3) Separates persistence logic from business logic.

In J2EE applications, Business Objects are usually implemented either as POJOs or as Enterprise Entity Beans.

3.2 Introduce Session Facade

In multi-tiered J2EE applications, some server-side components (e.g. Entity Beans, Session beans, and Business Objects) encapsulate business logic and data. Exposing these components to clients can cause the following problems:

1) Tight coupling between client and server components causes direct dependency and poor extensibility. Any change in business component requires corresponding code changes on client side.
2) Too many method invocations between the client and the server cause poor network performance.
3) Lack of a uniform client access strategy may lead to business objects to misuse.

To solve these problems, Session Facade is used to encapsulate business-tier components and expose a coarse-grained service to remote clients. Clients access the Session Facade instead of accessing business components directly.

Using a Session Facade has the following benefits:

1) Decouples the business components from the clients, and
2) reduces tight coupling and dependency between the
3) presentation and business tier.
4) Centralizes security management and transaction control.
5) Reduces fine-grained methods between the client and the
6) server, and improves the network performance.

3.3 Introduce Application Service

Applications implement use cases that coordinate multiple Business Objects and services. However, we cannot implement use case-coordinating behavior specifically within

Business Objects, because it increases coupling and reduces cohesion between these Business Objects. Likewise, we cannot add this business logic to a Session Facade, because the business logic potentially gets duplicated among different facades, reducing the reusability and maintainability of common code.

The Application Service becomes a better home for such business logic and results in simpler and more elegant and maintainable Session Facade implementations. Application Service also centralizes reusable business and workflow logic, and improves reusability of business logic. Application Services are implemented as POJOs in this system.

3.4 Introduce Business Delegate

By using the Session Facade pattern, we did not rule out tight coupling between the presentation and business tier. We do have a centralized access to the business logic but still the Session Facade itself is exposed to the client. To achieve loose coupling between clients at the presentation tier and the services implemented in the enterprise beans, Business Delegate Pattern is used. Business Delegate, as a proxy of Session Facade, encapsulates the underlying implementation details of the business service, such as lookup and access mechanisms. It hides the complexities of the services and acts as a simpler uniform interface to the business methods.

For example, the business tier of the on-line shopping subsystem provides a Session Facade named ProductSearchFacade for inquiring product module. ProductSearchFacade is implemented as Session Bean, which encapsulates some methods (such as searching products by name, etc). The presentation tier provides a Business Delegate named ProductSearchDelegate, which is implemented as a Java class. ProductSearchDelegate communicates with ProductSearchFacade and maintains a one-to-one relationship with that facade. ProductSearchDelegate has the same interfaces as the facade it delegates to.

3.5 Introduce service locator

In J2EE applications, clients need to locate and interact with the business components consisting of session and entity beans. The lookup and the creation of a bean is a resource intensive operation.

In order to reduce the overhead associated with establishing the communication between clients and enterprise beans (clients can be other enterprise beans), the Service Locator Pattern is used. This pattern hides the implementation details of the lookup mechanism and encapsulates related dependencies. It is designed as a single point to access the business components to improve the reuse of complicated lookup operations. The Service Locator is implemented as a Singleton in this system.

3.6 Introduce Transfer Object

J2EE applications implement server-side business components as Session Facades and Business Objects, and some of their methods return data to the client. These components are typically implemented as remote objects, such as session beans and entity beans. When these business components expose fine-grained get and set methods, a client must invoke several getter methods to get all the attribute values it needs. So it causes efficiency problems because every method call to an enterprise bean is potentially remote, and such remote calls create a network overhead.

The solution to this problem is to use a Transfer Object to encapsulate the business data transferred between the client and the business components. The Transfer Object is a serializable POJO that contains several members to aggregate and carry all the data in a single method. When the client requests the enterprise bean for the business data, the enterprise bean can construct the Transfer Object, populate it with its attribute values, and pass it by value to the client. When clients require more than one value from the business services layer, it is possible to reduce the number of remote calls to the Session Facade and to avoid overhead by using Transfer Objects to transport the data from the enterprise bean to its client.

3.7 Apply J2EE Multi-patterns to the Business Tier

We organize the system business tier as a set of business components. A business component (BC), as defined by Herzum and Sims[4], focuses on a business concept as "the software implementation of an autonomous business concept or business process. It consists of all software artifacts necessary to represent, implement, and deploy a given business concept as autonomous, reusable element of a larger distributed information system" [5].

Some general guidelines for identifying business components are as follows [6]:

1) Divide the entity classes into groups such that members of
2) a group can be managed together by a business component that handles responsibilities involving manipulating these entities;
3) By looking into analysis-level realizations of interrelated use cases, find and group similar responsibilities of the participating control classes in those realizations, and use a business component to handle these responsibilities;
4) Consider evolving boundary class representing a passive actor into a business component. This happens if a boundary class handles nontrivial connection logic.

The internal business accessing process of business component shown in Fig.4 is as follows:

1) Clients access a business component via one or more Business Delegates;
2) Business Delegate uses service locator to find and instantiate Session Facade, then transfer the client's service request to this Session Facade;
3) The Application Service, which is invoked by the Session Facade, completes client's service request by interacting with several Business Objects;
4) During business data transfer process, the Transfer Object is used to transfer data over multi-layers.

We use a session EJB to implement the Session Facade of business component. Entity classes of analysis model are mapped into Business Objects of business component, implemented with entity EJBs or Java classes. Control classes of analysis model (responsible for interval interactions, processing the business logic, calculating, and interacting with entity classes) are mapped into Application Service of business component, implemented with session EJBs or Java classes. We also define additional design elements such as business delegate Java classes, service locator Java classes, value object Java classes, etc.

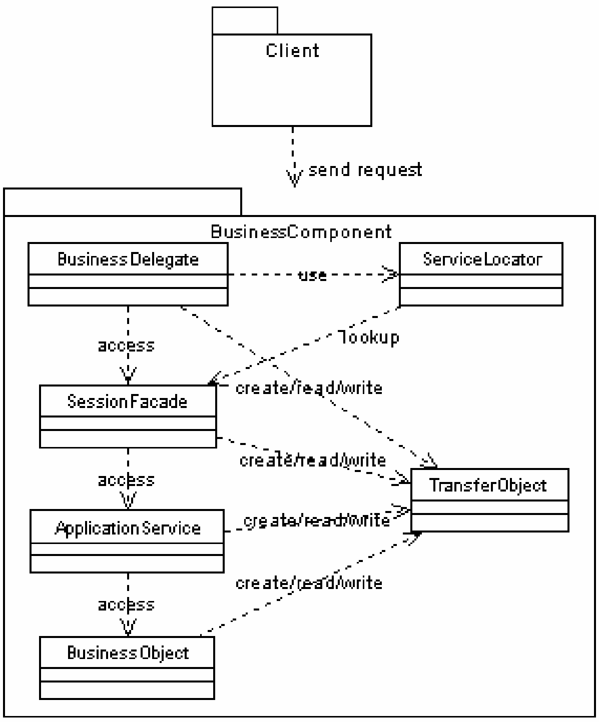


Fig.4. An Internal Perspective of a BusinessComponent

4. CONCLUSIONS

In this paper, we have briefly studied some proven design solutions to J2EE application known as “J2EE patterns”. These solutions could significantly improve the design and architecture of distributed enterprise applications. Using B2B online ordering system as example, we put these patterns together in best practices to design robust, efficient enterprise application. The result shows that using J2EE multi-patterns for enterprise applications can optimize system performance remarkably and enhance numerous aspects of the system, including maintainability, extensibility, and reusability, and also can abridge the developing time greatly.

REFERENCES

[1] D. Alur, J. Crupi, D. Malks, *Core J2EE^{TM} Patterns: Best Practices and Design Strategies*, 2nd Ed., Prentice Hall, 2003.

[2] E. Gamma, R. Helm, R. Johnson et al, *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison Wesley, 1994.

[3] F. Marinescu, *EJB design patterns.* John Wiley & Sons, 2001.

[4] P. Herzum, O. Sims, *Business Component Factory: A Comprehensive Overview of Component-Based Development for the Enterprise,* John Wiley & Sons, 2000.

[5] S. Arch-int, D. N. Batanov, “Development of industrial information systems on the Web using business components”, *Computers in Industry*, Vol.50, February 2003, pp.231~250.

[6] P. Eeles, K. Houston, W. Kozaczynski, *Building J2EE Applications with the Rational Unified Process,* Addison Wesley, 2002.

Jingli Zhang is a lecturer and assistant director of Computer Foundation Department in School of Information Technology, Southern Yangtze University. She graduated from Nanjing University of Science and Technology in 1994, visited and studied in famous universities of America such as Duke University, George Mason University etc. in December 2006. She has published five papers and several textbooks. Her research interests are in distributed parallel processing, software engineering and e-commerce.

Implementation of Navigation System based on User Interest in Active Service

Jingling Yuan[1], Yang Yu[2], Xuan Xiao[3]
Computer Science and Technology School, Wuhan University of Technology
Wuhan, Hubei, China
Email: yjl@mail.whut.edu.cn

ABSTRACT

A new dynamic modeling method based on user interest is proposed. Firstly, user information document is built by XML, the procedure of user interest modeling is mapped into the maintenance procedure of a class, which can describe various interest features and trace changes of user interest. Then, user requirements are stated classified and intelligent navigation is provided according to statistic results. Finally, navigation system based on user interest in active service is implemented.

Keywords: Active Service, User Interest, Navigation System, Interest Class, Classified Statistic

1. INTRODUCTION

The continuing developed electronic commerce and other applications lead the higher demand on the intelligence state of the Internet implementation. People hope the services provided by the network are able to change according to the applications; each user can enjoy personalized services. Active service is based on web services, adding users' requirement identification and processing function, making the users can choose appropriate function set according to their need and create new service and applications. This active service made Internet possess the capacity of providing services oriented to requirement.

In order to provide different service navigation to different users, the navigation system in active service has to building a user model first. The most common used modeling techniques such as vector space model, user appraisal matrix all have problems: It's difficult to describing user's various interest feature effectively and adequately by using single vector to generalize all users' interest; disabled from tracing the change of users' interest in time and implementing the real-time update of model, especially some frequently updated short-time interest change.

According to the drawbacks in the existed modeling methods, in this system, we build models based on user interest class dynamically. This kind of modeling method maps the procedure of establishing into that of a class maintenance, each class presents a kind of user interest .The addition of new interest and cancellation of old interest are implemented by the addition of new class and cancellation of old class.

According to user interest model, in the navigation based on classified statistic we will recommend the most welcomed service in every interest class to users. In other aspect, we may commend the proper interest class to users of different age, sex and career by classification of users

2. THE MODELING BASED ON USER INTEREST

In the active service navigation system, a user information document will be build the first time he loaded with the purpose to acquire their information and interest domains, then record their requirement, update and improve user document, provide better service to users.

2.1 The Acquisition and Matching of User Interest

To build user interest model, we have to extract some features from the information that the user interested in. Suppose there already had information set, we can extract features of user-preferred information from it as the following steps:

At first, a lexical analysis based on the requirement input by users will be done. ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) provides a complete set of dynamic linked library *ICTCLAS.dll.* By using the library we will get the result of Chinese words segmentation, pos tagging and unloaded word identification. Then employ *stoplist* technique to remove classification information of words, such as the, is, and but. A demand key word table will be obtained after this step.

The next step is to build text representation for key word table. At present, we employ the Vector Space Model (VSM, Gerard Salton and Mc Gill,1969) to represent text. Its basic idea is to use vector to express text: $(w_1,w_2,w_3,\ldots,w_n)$, w_i is the weight of attribute i. Generally we chose word or phrase as attribute and phrase as the component of text vector and do quantification on the text according to the use frequency of word. The computing method of weight based on VSM is based on this idea. In the active service, users' requirements are usually short without frequency problem, so the weight value can be assigned in accordance with the information gain after extracting the key words. Information gain refers to the amount of information that a phrase can apply to a type of class.

Whether the requirement accords with the user interest or not is judge by weighting the similar degree between a document (represented by vector) and the features of user interest. User interest information is described in natural language. The way to denote document vector can be used to denote user feature vector: $D=((y_1,d_1),(y_2,d_2),\ldots,(y_n,d_n))$, (y_i,d_i) means the weight of key word y_i is d_i .

For one of user interests can be described as a user eigenvector in the same space, we usually account the inner product or include angle cosine to denote it. The way to figure up the similarity quotient of document and user feature is shown as following:

$$Sim(X,D)=\frac{\sum_{i=1}^{n} xi*di}{\sqrt{(\sum_{i=1}^{n} xi^2)(\sum_{i=1}^{n} di^2)}}$$

2.2 Building User Interest Class Set

Each kind of user interest is a class. Accounting the similarity between the eigenvector of each user demand and the central vector of all user interest class .If the similarity is large than the predetermined similarity clique value LGate, the requirement will be classified into the most similar interest class (winner class) according to the basic competition rule.

If the similarity degree between input and all central vector is less than the clique value LGate, it shows the input demand is different from the existed classes, so it represent a new class which is one of user new interests and a new class need to be added. At this moment match it with the new interest class allocated from the interest class library. First identity the domain that requirement belongs to in the interest class library. Then judge the requirement in which interest class. In the next step identity a interest class by accounting the degree of similarity .Finally add the class into user interest class set.

Another kind of adjustment procedure is to delete the classes that users no longer interested in. The basis of deletion is: if a class has never won in the competition in a sufficient number of consecutive input mode, it means that users have no information of interest to the ownership of this type of information. Therefore, this information is the type of information users are no longer interested in and should be removed from the list of classes.

The delete method of the two counter classes is defined here. One is the Input Counter Service, which records the number of import demand in counterpart is the number of input. Another is Winning Counter ServLast, which records the recently input number for each interest class. When the margin of input counter with the winning counter of certain class Dis surpasses a certain clique value DisMax, then the certain class will be deleted from the user interest class.

Here adopt the rule trigged the additions, deletions of interest class controlled by clique value to track user interest changes This allows the user model to update and users interested in the changes to keep pace with it, which means that user model can immediately update once changes were detected. User's interest is in a good fit, especially the nature of dynamic changes of some frequent change in short-term interest.

Algorithm is as follows:

1. Initialization:

1) According to the table filled by users first time they registered establish the initial user interest model .The model including basic information (Age, Sex, etc.), and m user interest classes.
2) The number of demand services: Service=0. The service number of each interest class: ServNum[i]=0(i∈[1,m],i∈m).

2. User's demands:

1) Break down the user needs into a vector x , set
   1. Service=Service+1
2) Account the similarity of vector x which describes demand with the central vector of all interest class.

L[i] = $$Sim(X,D)=\frac{\sum_{i=1}^{n} xi*di}{\sqrt{(\sum_{i=1}^{n} xi^2)(\sum_{i=1}^{n} di^2)}}$$

3) If L[max]=Max(L[1],L[2],…,L[m])< LGate、(max ∈[1,m],max∈m), establish a new interest class. Otherwise considered this demand belongs to the *max* interest class:
   i. ServNum[max]=ServNum[max]+1
   ii. ServLast[max]=Service

3. Introduce a new interest class

In order to determine the demand doesn't in any existing user interest class, it should in match with a new interest class allocated from the interest class library interested.

[1] Comparing the central keyword of vector x to that of Area in interest class library .Forward maximum matching algorithm can be apply to ascertain the respective needs of the respective interest class of Area, the Area amount is AreaNum . AreaNum is a predetermine value which value is between the two to five and being determined according to the actual conditions.

[2] Apply forward maximum matching algorithm on the central keyword of each interest class in each Area, the result of this is a proper interest class set I .The set contains n interest classes which can not be deleted.

[3] To every interest class in I, computing the similarity of its central demand vector and demand vector.

L[j]= $$Sim(X,D)=\frac{\sum_{i=1}^{n} xj*dj}{\sqrt{(\sum_{i=1}^{n} xj^2)(\sum_{i=1}^{n} dj^2)}}$$

[4] if L[Newmax]= Max(L[1],L[2],…,L[n]) (Newmax ∈[1,n],Newmax∈n), we consider the demand belongs to the Newmaxth interest class in I.

[5] Add Newmax into the user interest class set ,then define the new class Server Counter ServNum[m]=1, ServLast[m]=Service

4. Delete user interest class

Delete the classes that user no longer interested in after completing the classification of demand.

[1] Calculate margin between the frequency of services for users and the latest server offered to user of a certain interest class
   i. Dis[k]=Service-ServLast[k] (k∈[1,m],k∈m)

[2] If Dis[k]>DisMax, delete the corresponding interest class.User interest class established by this algorithm has a good adaptability. In general its size is not large. The domains that user are favorable not more than 20.

2.3 User Information Document

User information document maintains the personal information of users, interest set and the historical record of their needs. To facilitate management and expanding, XML is used as storage. Fig 1 shows the basic framework:

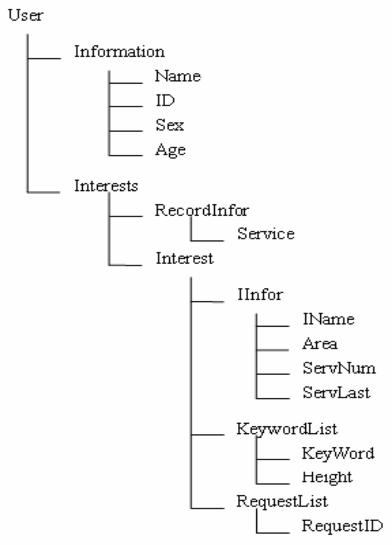


Fig.1. User Information document

Personal Information (Information) preserve the factors like age and gender which may affect customer demand.

User interests table (Interests) storage the latest and most concentrated interest class of users and the frequency of services Service. Each class contains the information like its name, the document it belongs to, the frequency of service ServNum, the serial number of the latest service ServLast, central keyword table and demand table. The central keyword table is used to preserve the central keywords of this interest class and its corresponding weight that is eigenvector in counterpart. The demand of each user is kept in the demand table which only records the demand number. Query the demand records library through s serial number when the specific information is required.

The keyword table of user interest and that of class in the corresponding interest class library is the same. This set this for the update algorithm of user interest class deliberately, which means when the user needs that can be attributed to the existing interest class they do not need to visit user interest class library.

The basic steps to establish and maintain the user information document on the basis of needs of users is shown in Fig 2:

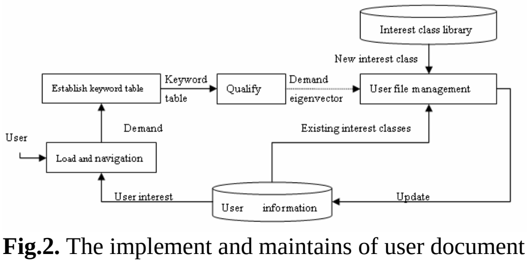


Fig.2. The implement and maintains of user document

3. CLASSIFIED STATISTICS

The way to establish user interest class set which contains their interests and demand records has been mentioned above. Besides recording their historical demand, in order to provide intelligent navigation, analysis demand is to be conducted on the basis of historical record, then recommend the network services they maybe interest in to users. A classified statistic library will be building with the purpose of record the results of a statistical.

The main purpose of classified statistics:
1) Derive the interests that user might have from their personal information.
2) Derive the demands user might have from their interests.

Normally personal interests can be attributed to many factors, such as gender, age, family environment, education and so on. To obtain the interests they might have through their personal information, users should be classified. Then static the possible interest of a certain kind of users, such as gender and age of only 15 to 25 years old as an example of statistics:

If Sex=male And 15＜Age≤20
Then Interest∈TypeA

If Sex=male And 20＜Age≤25
Then Interest∈TypeB

If Sex=female And 15＜Age≤20
Then Interest∈TypeC

If Sex=female And 20＜Age≤25
Then Interest∈TypeD

Type expresses the corresponding interest class of a particular type of users; Interest is a counterpart of one of user interests. We can get their information in the user information library, including age ,sex, and the classes of user interested in. According to the above algorithm, certain type of interest can be classified into the class set that different types of users interested in. For example, both men aged 15-20 and women aged 20-25 may have interested in “film”. Through adding other classified factors to refine the types of user and recording the number of a certain interests, a class set of one type user is set up. The more a certain interest appeared the more likely it being interested in, so the number of it shows up can presence its weight.

When one user is loading, the system can recommend the class he might be interested in by judging the types of users he belongs to (The recommended class should own the highest weight value in the interest class of this type of users).

To offer users their referral service on the basis of their interest, just need to record the most popular service currently in a certain class. The procedure as follow: first classified the demand of users in the demand record library based on the interest in the interest library, then record K items of demands about a certain kind of interest in the statistic library recorded recently , storied as a FIFO queue.

By using the above two classified statistics methods, user navigation information can be obtained. The process is shown in Fig 3:

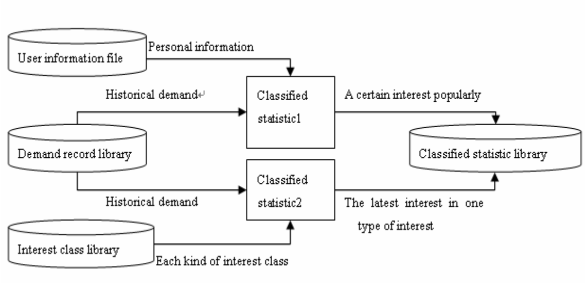


Fig.3. The process of classified statistic

4. THE IMPLEMENT OF NAVIGATION SYSTEM IN ACTIVE SERVICE

When serving users with active service, users first load on the navigation interface through its own user account. There are two optional services provide in the interface, one is a direct import demand, another one is the system automatically provides the services that might be of interest to the user.

When user imports the needs directly, the navigation system will do lexical analysis on demand and establish a keywords table. Keywords table will be handed over to other subsystems of active service (Functional decomposition and component assembly system), then provide the assembled service to user. On the other hand, the corresponding eigenvector can be gained by quantify the keywords table which will be submitted to the user's information document library and update the parameters of the user's information. This is the core part of the system: the management and maintenance categories of user interest class.

Several categories of interest are provided in navigation interface for users to choose, such as movies, food, football. They provide the possible interests according to the user interest class and classified statistics library. Provide the popular services of that type or the service had been used by the user before when users have chosen a particular interest.

The interactive of navigation interface, the user interface and other subsystems of the active service are shown in Fig.4:

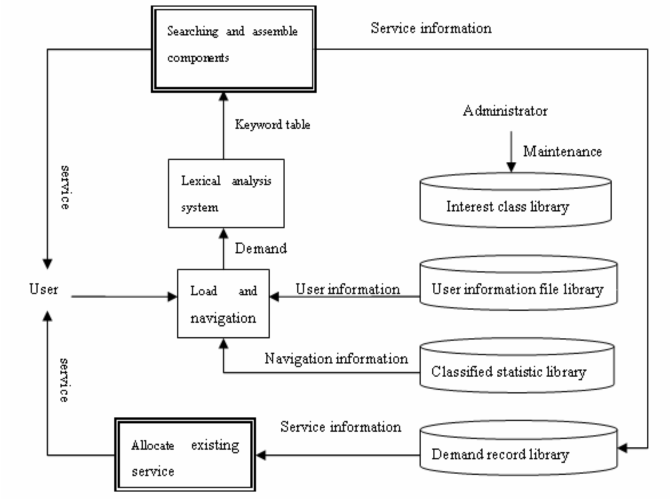


Fig.4. navigation interface and the database external interface

5. CONCLUSIONS

The navigation system has basically implemented the function of user intelligent interface in active service. It has achieved the demand of intelligent navigation by set dynamic user interest model and classified statistics.

The interest class library in this system, which basic contents is essentially the same after being defined ,is a static library that only managed by the system administrator. The advantage is that this will provide some unified alternative interest classes when establishing user document. The inadequate lies in the essentially fixed contents that can not do intelligent update according to the user's new demand. The study will focus on design an effective intelligent algorithm to the update of interest class library in the future.

REFERENCES

[1] Zhang Yaoxue, Fang Cunhao, “Active Service: Concepts, Architecture and Implementation”, Beijing: Science Publishing Company

[2] Guo Guoqiang, Zhang Yaoxue, Wei Zizhong, “Process of Program Mining”, *Changde Normal College Transaction (Natural Science Edition)*, , Vol.13 No.4, 2001.12, 45-48

[3] Zhang Raoxue, Dou Yuhong, Chen Songqiao, Li Xing, “Acquisition and Decomposition of Users’ Requirements in Program Mining”, *Computer Engineering and Science,* Vol.25 No.1, 2003, 1-5

[4] Wu Lihua, Liu Lu, Wei Kun, Wu Juhua, “User Interst Modeling Method Based on Dynamic Self-organizing Map Neural Network”, *Computer Integrated Manufacturing Systems*, Vol.12 No.8, Aug.2006, 1183-1210

[5] Liao Zhuhua, Liu Jianxun, Yi Aiping, “Web Service Discovery Based on User Interest”, *Micro-electronics and Computer (Supplement)*, Vol.23, 2006, 23-25

[6] Wang Junying, Guo Jingfeng, Huo Zheng, “Design and Implementation of Chinese Text Categorization System”, *Micro-electronics and Computer (Supplement)*, Vol.23, 2006, 262-265

[7] Liu Haifeng, Wang Yuanyuan, “Research of Several Problem in Text Retrieval Based on VSM”, *Journal of Information No.10,* 2006, 59-62

[8] Liu Qun, Zhang Huaping, Yu Hongkui, Chen Xueqi, “Chinese Lexical Analysis Using Cascaded Hidden Markov Model”, *Journal of Computer Research and Development*, Vol.41 No.8, Aug.2004, 1421-1428

Intelligent Web Information Categorization and Description Based on FCA *

Jun Ma^{1,2}， Fang Wang^{1}， Ming Chen^{1}
[1]College of Computer and Information Engineering, Henan University
[2]Institute of Data and Knowledge Engineering, Henan University
Kaifeng, Henan 475004, China
Email: mj@henu.edu.cn

ABSTRACT

First, this paper presents a method for extracting formal concept from web information characteristic and constructing formal context automatically. Then, it is expounded that how to use concept lattices to describe extracted results. After these it is gives two optimum strategies for browsing concept lattices, one is key-lattices and another is concept clustering based on similarity of attributes. Finally, it is expatiates how to apply these strategies on the practical software realization through specific examples.

Keywords: FCA, Concept Lattices, Information Characteristic, Key-Lattices, Product Information Search

1. INTRODUCTION

When the user searches a product through the internet, the present searching engine will show the result directly according to the keywords without further classification and general processing. With the rapid increasing of web information this practice which simply show the direct searching result can not satisfy the users' demands for retrieval of features of products. Firstly, the uses often need to shift different website to skim through the information and filter the searching result manually due to there is no further aptitudinal classification and processing. Secondly, the users can not know the relations and differences of products based on the searching results, because that the searching engine does not process secondary induction and integration. To solve these problems, the author proposes a new method to search, classify, and process products, i.e. makes use of FCA (Formal Concept Analysis) to extract with theory, integrate and structuralized search some information about features of products. After classified processing, shows to the uses in the form of concept lattices, as to simplify the difficulty of manual filter, and improve the efficiency of retrieval of features of products.

FCA (Formal Concept Analysis) is a method that concept knowledge is formalized. Its main idea is to extract concept and hierarchy from the context composed of binary relation, and build concept lattices based on data sets. For the practical demand of searching features of products, firstly we extract relative information from internet based on keywords, and transform the extracted result to the form of standard context. Then, construct concept lattices based on the above. A concept lattice is a structuralized description built on order relation. It has particular advantages on showing the classifying rules between concepts. It can display the order relation of concepts clearly and visually. The structure is suitable to integrate similar information and comparatively skim over it.

The paper begins from the practical point of view, firstly introduces the method to extract formal concept from the internet and construct context; then explain how to use concept lattices to describe extracted result, and build up the model of concept lattices based on it; and then proposes the method of using structure and concept of key-lattices to optimize integrated information, and provides the main idea of confirming structure of key-lattices, and arithmetic of concept clustering confirmed by similarities of attributes; Finally, expatiates how apply these strategies on the practical software realization through specific examples.

2. WEB INFORMATION EXTRACT AND DISPLAY

The primary task of displaying web information by concept lattices structure is extracting formal concept from internet. We extract needed characterized information by mode matching, and construct standard context based on extracted information. Finally, uses structure of concept lattices to integrate and describe these pieces of web information.

2.1 Extract Formal Concepts from Internet

There are two stages to extract formal concept. First, use present multi-searching engines to inquire information simultaneously. This method can make full use of present resource, and make the searching wider; then, analyze searching result and extract needed characterized information by mode matching. This mainly analyzes the HTML pages feed backed, and make certain and draw out needed content. Then it installs these contents in certain form to prepare constituting context.

2.2 Express Features of Information by Context

Data analyzed by FCA is usually marked by context. In standard context, suppose an object g has binary relation with an attribute m, which means object g has the attribute m. In the real world, one object not only has or has not one "attribute", for example, "color", "weight" have attributes more than one, so multi-context description is needed. It is not convenient for users to skim if lattices structure is too large; as a result, the classification of graduation about multi-context attribute becomes the most important problem. Due to different users have different demands on classification precision of object concerned, we will not adopt fixed classification degree to process the attributes, but roughly classify the attributes marked by specific numerical value into n sub-attribute fragments, and set proper range for each sub-attribute fragment. If an attribute is in a certain range, set the value of this sub-attribute fragment as "1" and others as "0". Fig.1 shows the process of transforming the specific attribute into multi-context.

The context constituted by this way is make up of ternary group （G,M,R），where $G=\{g_1,g_2,g_3\cdots g_n\}$is object set，i.e. One specific product collection, $M=\{m_1, m2, m3\cdots m_n\}$ is attribute set，which is the feature collection of products，R is the binary relation between G and M. Then we can construct concept lattices as along as the context is confirmed.

* Supported by Henan Innovation Project for University Prominent Research Talents of Henan Province (No. 2007KYCX018) and Natural Science Foundation of Henan University (No.05YBZR008).

	Price
1	480
2	670
3	1100
4	2500
5	3600
6	870
7	1350
8	2100
9	2800

⇨

	below500	500 - 1499	1500 - 2500	above2500
1	1	0	0	0
2	0	1	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	1	0	0
7	0	1	0	0
8	0	0	1	0
9	0	0	0	1

Fig.1. Transform the specific attribute into multi-context

2.3 Integrate and Describe Products Information by Concept Lattices

During the process of construct concept lattices, we use Godin [2] algorithm to form a concept lattices model based on hierarchical layout. In this concept lattices model, the whole lattices structure is a form of structure of information classification. Each node of lattices represents a piece of information, which means each node of lattices is an object. The connotation of node is the description of features of information, i.e. the attribute of an object.

The aim of displaying products feature information is providing users a clear relationship among concepts and brings convenience to skim and choose when people skim over and inquire some relative products information. Users can see relationship between information visually through hiberarchy of concept lattices and number of node in same degree. They can also obtain specific information through node. When people choose a node, the description of represented concept is show for reference.

In a concept lattice each node signifies a piece of information. Only when there is a rising route from the node *c2* to *c1*, the information represented by *c1* is subset of the information represented by *c2*, and *c2* is called the low-neighbor to *c1*, and the upper-neighbor *c1* to *c2*. If a node has many upper-neighbors that means the attributes comes from different node of upper-neighbor. For instance, a user inputs a keyword “Nokia” to search, and the sort of searching is “cell phone”, and the aim of searching is “features comparison”, so as to expose created concept lattices .In Fig.2, Nokia1110i and Nokia3250 have the function of GSM and Chord simultaneously, so these two types of cell phone represented by a same node. If users want to find a cell phone more suitable to them, all they need do is to find the lower-neighbor of this node. The next node represents the cell phone has more function than the above two. In Fig.2, the node represents the cell phone Nokia1110i which has the features of GSM, Chord and CandyBar.

3. THE STRATEGIES OF DISPLAYING AND OPTIMIZING LATTICES STRUCTURES INFORMATION

Before showing the web information by concept lattices structure, we need to design a method to make the best improvement of information accuracy within the minimum range. Under the usual case, users skim over information with certain purpose that means they concern about some attribute and neglect others irrelevant. If the lattices structure is over-complex, it has little significant for users, and it’s hard to use in practical way.

An optimized lattices structure not only can show the relevance of information, but also eliminate irrelevant information. It only shows the relevant key-lattices structure, and reduce browse complexity, so we must adopt certain strategies to control the complexity of lattices structure. According to these demand, we propose two optimized method: key-lattices structure and concept clustering, the former is used to decrease the complexity and the later to control the size of lattices structure.

3.1 Optimized and Integrated Information by Key-Lattices Structure

When we use concept lattices structure to show the relationship between information, if the size of lattices structure is proper, its advantage is obvious, but if it is too large, it will lose the meaning in practical browse. As a result, to control the size of lattices structure becomes the key point of showing the searching result. In real cases, users only show interest to some certain object or attribute, what they concerns is the partial information. We propose the strategies to optimize the relation of information by using key-lattices structure. The main ideas of making sure calculating way of key-lattices are described in below.

(1) Begins from the original node, visit the next lower-neighbor node which is not visited orderly. The line which links the original node and the node that share same features with it is named as “1”; the line which links the original node and the node that does not share same features with it is named as “0”.

(2) Begins from node that is named as “1”, visit next lower-neighbor node which is not visited orderly. The line which links the node and the lower-neighbor node that share same features with it is named as “1”; the line which links the node and the lower-neighbor node that does not share same features with it is named as “0”.

(3) Repeat step 2 until all the nodes which contains concerned attribute are found. Finally the chart which is constituted by lines named “1” and the nodes linked is the key-lattices structure.

Take cell phone Nokia information concerned by a user for example, the concept lattices (Fig.2) constituted by context (Table 1), user use right key of the mouse to click the node color “message”, “chord”, key-lattices structure(Fig.3) is shown if choose “only show relevant information” in shortcut menu. As the key-lattices structure is certain, all the object collection that users may be interested in is shown and the further inquire can be carried out based on it.

Table 1. Formal Context

	a	b	c	d	e	f	g	h
1		√		√			√	√
2	√			√				√
3		√			√		√	√
4		√		√			√	√
5		√	√	√				√
6	√		√			√	√	√

1-Nokia6088, 2-Nokia1110i, 3-Nokia2635i, 4-Nokia2865, 5-Nokia2875i, 6-Nokia3250, a-GMS, b-CDMA, c-Digital Camera, d-CandyBar, e-Clamshell, f-Slide, g-MMS, h-Chord.

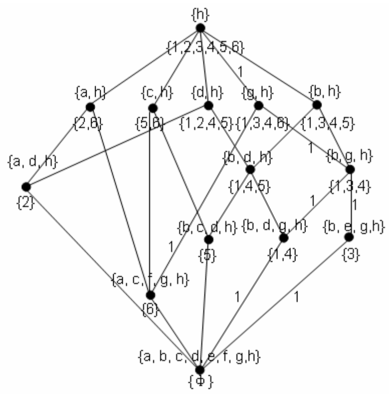


Fig.2. Concept Lattices

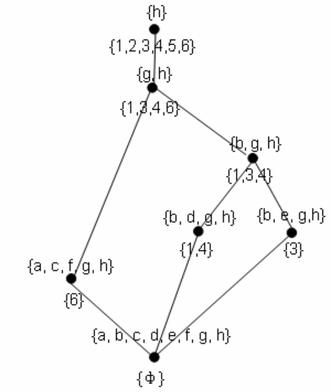


Fig.3. Key-Lattice

3.2 Clustering Integration Result by Using Attributes Similarities

Products features information which lattices structure is optimized by strategies of key-lattices structure is suitable to the case that search specific product according to feature information, but sometimes product feature information might be the common attribute of all products, so the key-lattices structure is totally same to original structure. Obviously, strategies of key-lattices structure can not arrive at the purpose to show the products features information which has been optimized. Therefore, the paper proposes another strategy--concept clustering. Its main idea is to measure the similarities among concepts by a similar scale, and finish the classification based on attribute belongingness.

Due to there exist a feature of order relation between object and attribute, we must consider about similarities of attributes when we discuss the concept similarities. Because all objects are formed by attributes in same range, the similarities of object can reflect similar scale. More similar the attributes are more similarities between the two objects. Confirm the similarities of concept object according to attribute, and make clear the relation of objects according to similarities, classify the concept, consider the more similar concept collection as a cluster, and then consider the cluster as node to make lattices. As a result, users only need to find the attribute cluster they are interested in. Definition1 is a method to compare the similarities of objects based on similarities of attributes obtained from quoted passage [3].

Definition 1 $$Sim(m,n)=\frac{f(M\cap N)}{f(M\cap N)+\alpha\cdot f(M-N)+\beta\cdot f(N-M)}$$

M and *N* are attribute set of object *m* and *n*, *f(M-N)* means the attribute set which is contained in attribute set *M*, but not in *N*; $f(M\cap N)$ signifies the set contains the attribute of both *M* and *N*. α and β are set parameter.

According to definition1, that calculates the similarities of two objects by attributes similarities can be defined as:

$$Sim(ob1,ob2)=\frac{\left|(m1\vee m2)_{LA}\right|}{\left|(m1\vee m2)_{LA}\right|+\alpha\left|m1_{LA}-m2_{LA}\right|+(1-\alpha)\left|m2_{LA}-m1_{LA}\right|}$$

where *Ob1* and *ob2* are two objects needed to compare; *m1* is the attribute set of *Ob1* and *m2* is the attribute set of *ob2*; $\mathrm{m1}_{LA}\vee\mathrm{m2}_{LA}$ is the set of supremum-irreducible elements which are in both *m1* and *m2*; $\mathrm{m1}_{LA}-\mathrm{m2}_{LA}$ is the set of elements which are in *m1* but not in *m2*; Also, $\mathrm{m2}_{LA}-\mathrm{m1}_{LA}$ is the set of elements which are in *m2* but not in *m1*; α is criterion to measure the similarities of attributes, and can set some for users' choice.

During the realization of software, the two objects which similarities are more than criterion can be classified in a same cluster, in the same way, we can also compare all the objects and classify them, and finally we can get concept clustering confirmed by attributes similarities. The method to calculate is described as the following:

Input: a group of objects would be compared: *ob1* and *ob2* and coefficient
Output: produced concept clustering collection S_C
The process:

```
    $S_c := \phi$
    for (p=1; p<n; p++)
    {
        for (q=p+1; q<n; q++)
        {
        if
        $Sim(ob1,ob2)=\frac{|(m1\vee m2)_{LA}|}{|(m1\vee m2)_{LA}|+\alpha|m1_{LA}-m2_{LA}|+(1-\alpha)|m2_{LA}-m1_{LA}|}>\Phi$
        then    $S_C \leftarrow S_C \cup \{p, q\}$
        else    $S_C \leftarrow S_C \cup \{\phi\}$
        }
    }
```

Compare similarities of object Nokia2865 and object Nokia2875i according to the main idea of concept clustering calculation method based on attributes similarities (set coefficient for 0.5).

$$\mathrm{Sim}(\mathit{Nokia}\ 2865,\mathit{Nokia}2875\mathrm{i})=\frac{5}{5+0.5\cdot 1+(1-0.5)\cdot 1}\approx 0.83>0.5$$

It can be seen that Nokia2865 and Nokia2875i have high similarities, and can be classified into a same cluster. Similarly, to obtain the object similarities through measuring the attributes similarities between two concepts, then make a clustering process according to the similar relation among concepts. Its realistic significant lies in: classify the cell phones which function has high similarities into one concept cluster, so as to provide convenience to users' general grasp, and users can amply analyze and compare the function of cell phones within a cluster. Finally they can find a type suitable to their taste.

Table 2 shows the clustering result based on attributes similarities of six types of cell phone Nokia.

Table 2 Cluster Result

Cluster1	{Nokia6088， Nokia2865， Nokia2875i}	Cluster2	{Nokia2635i}
Cluster3	{Nokia3250}	Cluster4	{Nokia1110i}

It is obvious that the purpose of controlling the size of lattices structure can be arrived at by using the strategies of concept clustering, and the clustering result is fit to real one, also a good effect can be obtained. One should be paid attention is that when the number of lattices node is in the range of 10—15, the effect of using key-lattices structure is quite good. When it is beyond 20, concept clustering is better. Tin the real use, which strategy is better depends on specific case or combine both of them.

4. SOFTWARE REALIZATION

In the process of software realization, we make use of ASP.NET and C# language to realize products feature information searching system, and display the processed result; Fig.4 shows one of the pages. The tree view part in the left is area can be chosen under character view, and cell phones are classified to several categories according to “brand”, “mode”, “function”, and “price”. Users can choose according to different classifying ways. The form of cell phone line is shown after users’ choosing on the right of the webpage. Users can have a detailed comparison through choosing two or more types of cell phone. In the middle of the Fig.4, there is correspondent concept lattices structure, and users can also use lattices structure to examine the mode of cell phone which has certain functions. It is easy to say that the combining use of concept lattices structure and tree view structure can reflect more clearly the relation and differences of products, and play an important role of guidance for users’ choice.

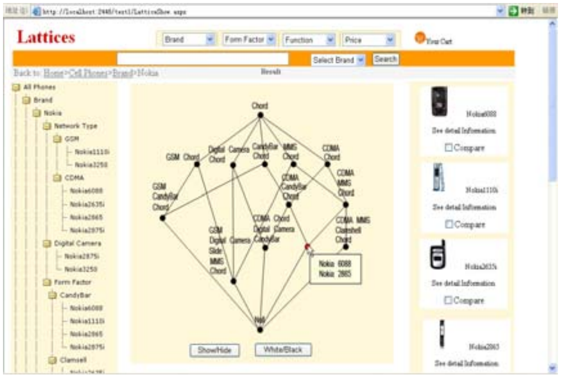


Fig.4. Part of results web page

5. CONCLUSIONS

The paper begins from strengthen the searching function of traditional searching engine, makes use of FCA theory, discusses the searching products feature information, integration and browse, and simplify the browse complexity by using key-lattices, and classify the objects by concept similarities measuring method and concept clustering idea. From the structural observing point of view, lattices structure obtained by concept clustering is more visual, easier to understand, and more convenient to observe than marking by character and tree-like structure. At the same time, because products information has been classified and integrated process, it is more beneficial for users to find rapidly the needed information, and it has greater significant in the real application.

REFERENCES

[1] Ganter B, Wille R.Formal “Concept Analysis: Mathematical Foundations.” *Berlin: Springer-Verlag,* 1999.

[2] Robert Godin, Rokia Missaoui, “Hassan Alaoui. Incremental concept formation algorithms based on Galois(concept)”,Appeared in *Computational Intelligence*, 1995 ,11(2),246 ~ 267.

[3] Yi Zhao, Xia Wang, Wolfgang Halang. “Ontology Mapping based on Rough Formal Concept Analysis. Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services” (*AICT-ICIW'06*), 2006, 19(25):180 ~ 180.

[4] Yu QIANG, Zong-tian LIU, “Research on Fuzzy Concept Lattice in Knowledge Discovery and a Construction Algorithm,” *Chinese Journal of Electronics*, 2005,33(2):350 ~ 353.

[5] Juan M. Cigarr á n, Julio Gonzalo, Anselmo Peñas, Felisa Verdejo. “Browsing Search Results via Formal Concept Analysis: Automatic Selection of Attributes.” *Springer Berlin*,2004, 2961:74 ~ 87.

[6] Petko Valtchev, David Grosser, Cyril Roume, Mohamed Hacene1.Galicia: “an open platform for lattices. In Using Conceptual Structures: Contributions to the 11th Intl.” *Conference on Conceptual Structures* (ICCS’ 03), 2003, 241 ~ 254.

[7] Andreas Hotho, Gerd Stumme. “Conceptual Clustering of Text Clusters.” FGML Workshop, Hannover, 2002.

[8] Richard Cole, Peter Eklund, Gerd Stumme. CEM-A program for visualization and discovery in email. The 4th European Conf on Principles and Practice of Knowledge Discovery in Databases, *Berlin: Springer2 Verlag*, 2000.

[9] Thanh Tho Quan, Siu Cheung Hui, Tru Hoang Cao. “FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web,” In Proceedings of the Knowledge Discovery and Ontologisms Workshop, September 24, 2004, Pisa, Italy.2004.

Jun Ma(1964-), male, an associate professor of Henan University, director of network engineering and software engineering staff room, college of computer and information engineering. He graduated from Tianjin University in 1986. He has published five books. His research interests are in distributed parallel processing, and knowledge discovery.

Fang Wang (1980-), female, student of computer and information engineering department, Henan University. Her research interests are in software engineering, network application and knowledge discovery.

Enterprise Information System Integration Technology Based on SCM

Lifang Kong[1, 2], Hong Zhang[1]
[1]School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221008, China;
[2]Basic Courses Department, Xuzhou Air Force College, Xuzhou 221000, China
Email: klf030@163.com

ABSTRACT

With the prevalence of internet and the coming into being of knowledge economy, enterprises are developing step by step towards a so-called Enterprises′ Combination, forming "Association of Economic Resources"-the Supplying Chain which is constituted by suppliers, manufactures, wholesalers, retailers and customers involved in production and circulation so as to cater to the needs of times, markets and competition .Starting from the essence of the Supplying Chain, we have put forward a model of Enterprises′ Information System Forming Frame based on web service and workflow towards the management of the Supplying Chain and we also have stated its functional constitution and analyzed the key technology to forming.

Keywords: Management of the Supplying Chain (SCM), System Integration, Workflow, Web Service

1. INTRODUCTIONS

With the taking shape of global economy and the development of internet ,the market competition of 21st century will be promoted from among the enterprises themselves to a more highly level, "Joint System of Enterprises"---competition among the Supplying Chains. Just as what Martin Christopher who is an expert on the Supplying Chain had said "Only is there the Supplying Chain in the market, but no enterprises.", "The real competition lies in the Supplying Chain, but not the enterprises themselves". The Supplying Chain, emphasizing on system perfect and operation entirely, will become the key ability for competition. The Management to the Supplying Chain is a shape of management, which needs the realization of system, integration, promptness, including the new thoughts and technology of modern management. Enterprise have to get their interior and exterior resources combined because of the business circumstance, so as to more perfectly meet the clients' requirements, therefore more value can be gained by the clients and more profit can be obtained by the enterprise. The process among research, suppliers, manufactures, retailers and customers in the Supplying Chain are becoming frequent and their combination will be tougher due to the features of market's active change. Above all, so the following problems must be solved at once:

System Structure. The supplying Chain needs not only to clear up the obvious, effectiveness, no fortune-creating actions, but also to make a series of fortune-creating actions more effective through system-integration, reducing delay and making much progress; Supporting the re-establishment of enterprise' business in order to make the purposes between enterprise' inner production activity and the Supplying Chain same.

Management-Exchanging Available. The Integrated Information System needs great compatible ability so as to realize the operation beyond language and platform, providing the applied programmer of the web with reliable visits.
Management-Changing. Considering the activity of every enterprise's co-operation relationship, the integrated information system must have the great ability of tenderness, looseness, tightness and sub-assembling in moving so as to support the requirements of enterprises' fast catering to market needs.

2. FUNCTION FRAME

As a kind of extended enterprise, the flowing of information and means to get in the Supplying Chain are different from a single enterprise. Node enterprise of the Supplying Chain must collect and spread information from the two resources interior and exterior, seizing the means, technology and method of handling which can mostly create fortune. The system must realize the counting and controlling of enterprise' inner resources, the analysis and comparison of distributing plans among enterprises, the conveying and share of information with other enterprises as well. Internationally, many enterprises which have already put SCM into practice successfully have already set up a system used for coordinating resources interior and exterior, for instance, Citroen's system of SCM' system is based on EDI, the SCM' system which is the combination between K.B company and electronic business. In view of our country's current condition, the most direct way for realizing business electronically among the enterprises is the entering into the global information expressway-Internet. The information tools that are for putting SCM into practice should be established on Internet/ Intranet. Fig.1 has described the function frame of enterprises' information-itegrating system in the Supplying Chain based on Internet/ Intranet.

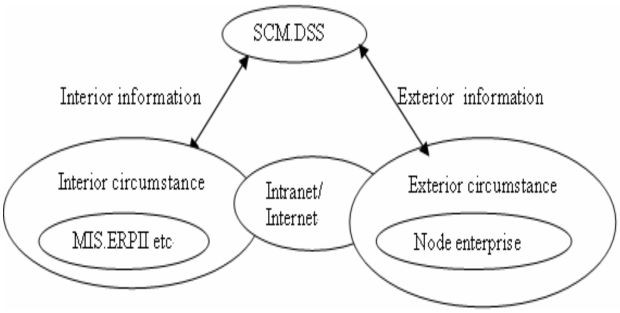


Fig.1. Frame of Enterprise Information System Integration Technology based on SCM

We can conclude from Fig.1 that the system comes from branch system which has 3 levels. they are: interior information management System, exterior information exchange system, SCM decision support system.

2. 1 Interior Information Management System

Interior Information Management System involves all the business processes of every department inside enterprise. Its main function is the accomplishment of data-handling, state-counting, tendency-analyzing and so on. The Supplying Chain inside enterprise can see the coordination by the establishment of this information system. The scheme available is the MIS(Management Information System) based on Internet, i.e IBMIS (Intranet Based Management Information System)which cores Internet Technology and bases Web Technology, is an application system based on network object Joint technology, multimedia document framework, span multi-operating system and multi- database platform. It is made up of network application support platform, information resource management platform, information transfer platform, working flow management platform and transaction processing platform .Every application of MIS is based on these platforms.

2.2 Exterior Information Exchange System

The functional model of Exterior information exchange system can be seen in Fig.2.It concretely does two assignments: one is the information for share and exchanging between Supplying Chain enterprise and upper or lower enterprises; the other is the business system on internet which directly faces final users, including business and clients' services on internet and so on .These work will be done by Internet Information –Exchanging outside, so an agreement between exchangers and enterprise is necessary ,regulating the kinds styles and standards of information–exchanging .The integration of enterprise's exterior Supplying Chain can see its realization after the establishment.

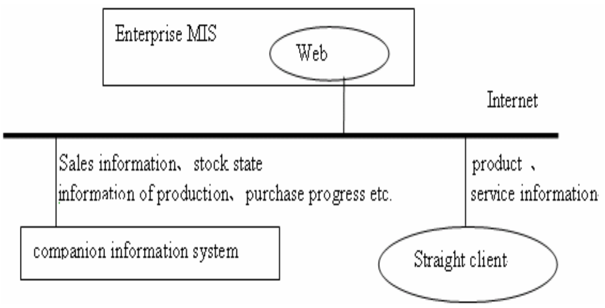

Enterprise MIS
Web
Internet
Sales information、stock state
information of production、purchase progress etc.
product 、
service information
companion information system
Straight client

Fig.2. information exchange system of function model

3. THE PRINCIPLE OF WEB SERVICE

Web service is a series of standards and developing standards. which are designed and named by W3C ,is used to promote communication beyond medium between programmers .These standards include XML, HTTP, UDDI, WEDL, SOAP and so on . Web service is the structure of normal use towards service which is the finding and releasing of Web service.

It includes three joiners and three basic operations: service-provider, service-requester, service-actor and releasing, looking for, binding, Service-provider release its content to service-actor; when service-requester is eager to get served, firstly he will reach service-actor and gain the related information for how to use it, then goes to coordinate and use the service released by service-releaser; service-acting has provided service with a mechanism which can release (register classify) themselves and their service, it has also provided request with a mechanism which can search for the service they want. Web service use general description, searching and UDDI when it is releasing service; uses UDDI and WSDL which is descriptive language of web service when looking for service; uses SOAP which is visiting agreement of simple objects when binding service. Its working mechanism is as following (Fig.3).

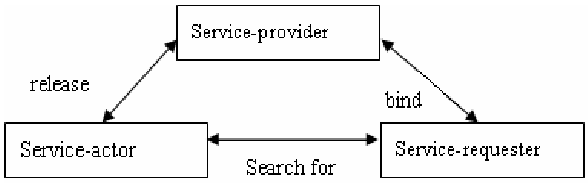


Fig.3 Principle of Web Service

4. ENTERPRISES'INFORMATION SYSTEM-INTEGRATING WHICH FACES THE SUPPLY CHAIN BASED ON WEB SERVICE

4.1 The Structure of Enterprises' Information System-integrating Which Faces the Supplying Chain

Starting from the requirements of information system integration which facing the Supplying Chain, we have launched an information system-integrating structure which is a distributing pattern, but open, and moving in internet towards the supplying Chain based on Web Service and Working Flow(Fig.4).The Design Purpose of the Structure: Develop software system structure which makes all kinds of information and applied resources involved in business processes become one information system, make it as an integrated whole operation. We can get idea from Fig.4 that "the Structure" is mainly made up of service-acting of the Supplying Chain, service-provider and service-requester. The centre of service-acting is a versatile system, realizing the main function as following: (1) Draft the web service needed by suppliers and clients in the Supplying Chain; (2) Write down and check the web service needed by every member through Servicing Machine of UDDI; (3) Service safely to controlling visits between members of the Supplying Chain and enterprises. And information is conveyed between members and enterprises, (4) Estimate and record the credit of members. Service-providers are responsible for coordinating and using its own information system in the light of WSDL files supplied by the centre of Service-acting in order to realize corresponding function, registered and released at the Acting Center so that they can be visited, accepted by other application and coordinated, used by requesters for dealing with Web Service.

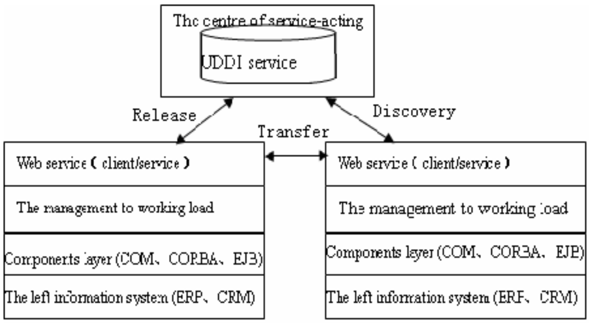

The centre of service-acting
UDDI service
Release
Discovery
Transfer
Web service (client/service)
The management to working load
Components layer (COM、CORBA、EJB)
The left information system (ERP、CRM)
Web service (client/service)
The management to working load
Components layer (COM、CORBA、EJB)
The left information system (ERP、CRM)

Fig.4. Frame of Enterprise Information System Integration

According to enterprises' own needs, Service-requesters are in charge of looking for web service they want through the Centre of Service-acting, then bind and coordinate, use web service in the light of the definition.

4.2 The Structure of Enterprise's Information System

To every member, he is not only service-provider, but also service-requester; client, service machine. Four levels are designed to every member: the web service level, the management to working flow level, the components-layer and the left information system level. Related business logic is inside the business logical level in the Structure. The business logical level will coordinate and use web service programmer of clients when we want to contact other enterprises and coordinate, use other enterprises' web service. While other enterprises want to coordinate and use the web service of this enterprise, after the realization of forwarding request from the Service Machine of web service to web service, web service realization program itself has nothing to do with the handling of business, after its accomplishment of coordination and using of operation components of logic-layer, then gives the results back to other enterprises. Logic-components can be based on CORBA, EJB or DCOM. The connecting level is responsible for the contact and mapping between the level of business logical and the left information system. The level of management to working flow is cored by execution service of the working flow, taking advantages of WSFL to realize the making up between web service and commercial process. It is mainly make up of execution service of working flow, packing tools of service, model-building tools of working flow, giving notice to users' task and tool-submitting, management to working flow and tools for supervisation. We can realize the execution and controlling of commercial process, and use, manage the information, application etc. related to commercial activities through realizing management of requirement, production, purchase, research on new product, feedback of process by the level of management to working flow and coordinating the joins among these processes. In this way, the intactness, continuousness and coincidence will be well guaranteed and the effectiveness of the Supplying Chain will also be promoted rapidly.

1） Carrying out Service of Working Load, its function includes the living examples of model, the coming out of working table, coordination and use of applied objects outside, controlling to the route of activity, recording events and handling the abnormality etc, and investigating information, sending information as well which support the communication between the model of execution service and the model outside.
2） The Tools to model-building of Working Flow is a model which backs up users to establish working process on the platform through illustration.
3） Supervisation to Working Flow and Tools to Management, its function includes management to working flow diary, management of edition to working flow model and the watch to executing state of working flow.
4） The Notice of Users' Task and the Tools for Submittion. Realize the executing state of noticing ways which support all kinds of tasks. support users' task-submitting by all kinds of means; support the example of the users' searching for specific working load.
5） The Tools to Service-Packing is the key tool integrated by working flow level and B2B. Its main function is drawing put operation which can be touched by information outside from the model of working load and waiting for corresponding affair outside, thereby WSDL for business service is come into being. Put the formed service into web service, then supply externally certain operation and can affect the affairs outside.

4.3 The Process of Coordination and Use to Web Service

The process is explained as following based on SUN ONE (Open Net Environment):

1） The service content which is needed to supply is to be registered into the Service Machine of UDDI by service-supplier's using JAXR(Java API for XML Registries);
2） The service-requester chooses web service supplier who is interested to him by skimming over machine, then search for the service in the Service Machine of UDDI by using JAXR;
3） After finishing downloading Document WSDL, which is needed service，the service-requester can get the connection and semantic ness of coordination and use for service.
4） SOAP Request Information which is produced by JAX RPC or JAXM will be sent to service-suppliers through Web Servicing Machine in accordance with the information from USDL.
5） The operations are as following when the request reaches service-supplier. The files are anti-serialized into the object of JAVA through integrated SAX and JAXP of DOM after receiving the request of File XML by servlet. These JAVA objects will be the method name of coordition and use request if it is RPC request. JAVA objects will be the handled information name if it is information service. The content of information will be changed into parameter of these methods so that the business data can be handled through one or more EJB components according to these JAVA objects by Servlet. The left information system presently will be probably coordinated and used when handling JMS or JAVA spaces can be used as connecting machine when coordinating and using.The results will be returned to Servlet after EJB has finished handling .The returned data will be serialized into File XML by Servlet ,and then the file is returned to clients.

5. CONCLUSIONS

Web service bases on open standard, providing enterprises with a leveled calculating circumstance which is flexible loose but tight; meanwhile, it has shielded applied platform of low-leveled enterprises and it hardly has anything to do with platform. The data for communicating are normal XML data, so it has noting to do with language. Besides, it has also provided the applied integrating of enterprises a convenient and swift channel. People on developing software can easily solve the problem of system-integrating which is absolutely hard to be solved before. It has perfect quality in integrating and elasticity as well. Web service has removed the problem of mutual operation in solving method of EAI at present and has provided a best solving method of "Use available at once after insertion". It is one of the best technologies presently in the realization of system-integrating, though it needs to be perfected on safety, working load, the quality of service and management etc.

REFERENCES

[1] *Developing Technology of Web Service on Spanning platfor, translated* by Tao Yang and Xiaoyun Yang, Beijing. Machinery Industry Press.
[2] *Service Frame and Developing Open Mutual Technology*, translated by Xiaolu , Qinghua University Press.
[3] *Agile Management System of the Supplying Chain Based on Web Service*, translated by HaoYin and Huiyou Chang. Integrating and manufacturing System of computer.

BRIEF INTRODUCTION TO THE AUTHOR

Name: **Lifang Kong**, Year of Birth: 1972, Sex: Female
Nation: Han, Degree: Dr, Direction of Research: the handling of pictures and signals, communication of computer.

The Research of Distributed Parallel Computation Based JXTA

Chun Liu, Qingping Guo
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430063, China
Email: liuchun@whut.edu.cn

ABSTRACT

JXTA provides an underlying framework for the executing and developing of P2P application. The paper analyzes the essence of the JXTA virtual network, proposes a distributed parallel computing model based JXTA, and addresses how to use this model to develop the JXTA application, after the deep study of concepts and protocols about JXTA.

Keywords: JXTA, P2P, Parallel Computation

1. INTRODUCTION

With the traditional network framework Client/Server model, the client applies for services and the server provides them. The more the clients, the more pressure the server will stand. At the end the server will invariably becomes resource bottleneck limiting the expanding of the network. While the novel P2P network is totally different, which is characterized by decentralized control. All the peers in the network hold the equal status, which means that every peer can act as both server and client. Due to this equal model, the core of net application is transferred from central server to network edge devices, which eliminates the great dependence of server.

Because of the brilliant prospects of P2P techniques, many software companies develop the P2P software productions in competition. But most of them only suit to some certain platform, which can’t communicate and share data with each other. For example, Napster mainly provides the query and download of music files, while Gnutella only offers the share of normal files. These systems are not compatible account for lacking of the common infrastructure. Project JXTA has joined the P2P family with a novel approach. Instead of providing a solution for a specific application domain, JXTA offers a complete infrastructure for the development of P2P applications, from collaboration to parallel computation, which enables developers to create P2P services more quickly and easily.

The paper addresses how to use JXTA platform to develop distributed parallel computation application, after deeply studying of concepts and protocols about JXTA platform.

2. JXTA OVERVIEW

The JXTA protocol defines a suite of six XML-based protocols that standardize the manner in which peers self-organize into peergroups, publish and discover peer resources, communicate, and monitor each other.

This means that JXTA is a standard framework which supports P2P application. The JXTA protocol helps to discorver ，communicate between peers and manage the P2P applications. The six JXTA protocols are Peer Resolver Protocol, Peer Discovery Protocol, Peer Information Protocol, Pipe Binding Protocol, Endpoint Routing Protocol and Rendezvous Protocol. Any peer who implements the protocol can enter the JXTA network to get the peer computing services.

2.1 JXTA Virtual Network

We will start with the JXTA virtual network in order to have a better understanding of JXTA. In traditional net, every peer has its own unique identity that is known as IP address, which determines the route of sending messages to the peer. Since the reliability of the P2P network is low, any device often breaks down from the network and reenters the network with another new IP address. The changes of peers in P2P network are so frequent that central naming services cannot handle it. So JXTA introduces an overlay on top of the existing physical network. Any device in JXTA network not only has its own unique identity to label its physical network address but also has the virtual identity to label its position in JXTA network, which is illustrated in Figure 1. The dots in graph identify the virtual identities of peers in JXTA network. JXTA net offers services to map the virtual identity to physical network address, which makes the virtual net is able to cross barriers like firewalls and Network Address Translation (NAT), and establishes peer communities spanning any part of the physical network.

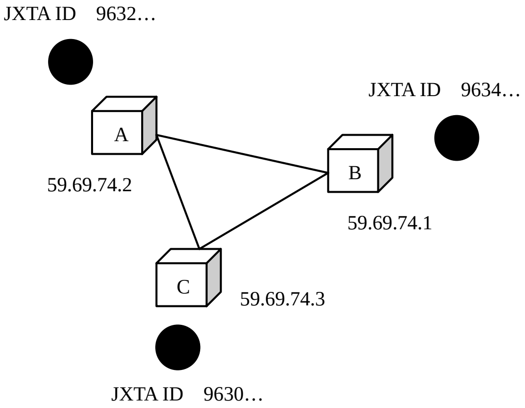


Fig.1. duple network formulated by JXTA virtual network and physical network

2.2 JXTA Core Entity

Project JXTA References introduce some novel concepts such as peer, peer groups, pipe, endpoint and advertisement, which are pivotal to understand and master the JXTA protocols.

Peers are any devices that run some or all the JXTA protocols. A peer can be anything with a digital heartbeat that supports the JXTA core, including servers, PCs, cellular phones and so on. There are two kinds of peers in JXTA virtual network. Most of the peers are simple or edge peers, usually desktop computers connected by a LAN or modem to the Internet. Rendezvous peers are usually more powerful peers, which facilitate search and discovery and provide resolving operation. The rendezvous peers can also act as proxy or relax peers to provide routing information and pass messages between peers separated by firewalls and NAT.

A Peer Group is a collection of peers that have agreed upon a common set of rules to publish, share and access codats, and communicate between themselves. Each peer group can establish its own membership policy from open (anybody can join) to highly secure and protected (join only if you have sufficient credential).

Endpoint is the address that peer use to implement some certain communication protocol (HTTP, TCP, BEEP). A peer may have several endpoints, which use different protocols to communicate with other peers.

Pipes are an abstraction used for inter-peer communication, which are a mechanism for establishing unidirectional, asynchronous communication between peers. Peers bind to one end of the pipe and when both ends are bound, messages can be passed. At runtime, a pipe end is bound to an endpoint address instead of physical location and hat is why the JXTA network is the overlay of physical network.

Advertisements are XML documents of a well-defined format, which are used to represent all entities in JXTA, including peers, groups, pipes and services.

3. A DISTRIBUTED PARALLEL COMPUTING MODEL

In distributed computation fields, many problems are solved through a kind of parallel computing method, which divides a large compute task into several small sub-tasks, dispatches them to other computers located far away to fulfill, and collects results from them and assembles to give the final results. Such parallel computing model is widely used.

3.1 Server Mode/Client Mode

The unusual character of such kind of application is that every peer acts as coordinator and subordinate at the same time. The role of coordinator is to load required class library and make the peer to work correctly, while the role of subordinate is to fulfill the child mission that the coordinator assigned. However the subordinate can also divide the child mission into even small ones, grandchild missions, and dispatch them to the active peers on the network. In this situation, the subordinate acts as the role of coordinator. In addition, one thing should be pointed out that for the adapter between coordinator and user interface, the coordinator acts as the role of subordinate to receive the mission the user assigned and compute it. The server mode/client mode (SM/CM) is an important design mode in P2P application, which is shown in Fig.2.

The traits of above parallel computing model are shown as following. Every peer has the same code. Different peer use different data to fulfill the mission, and the data is distributed stored in the network. Collect the result according to the mission identity. In general, generic distributed computation can be solved with this model.

3.2 Typical Peer Operations

In the progress of implement, two suites of codes, server mode and client mode, should be provided. According to JXTA framework, no matter what mode it is, the typical peer operations are similar, which can be concluded as following.

(1) Start the JXTA platform
Starting the JXTA platform loads a class library, which depending on the configuration of the peer and makes the peers work correctly.

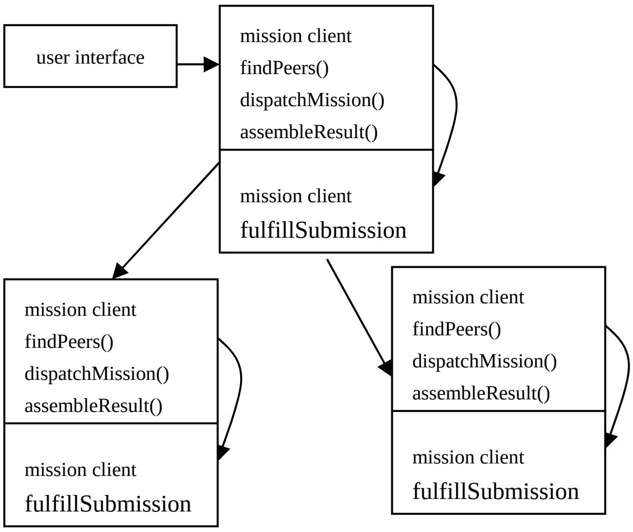


Fig.2. SM/CM of peers

(2) Join a peer group
Joining a peer group can give the peer more secure and efficient services. It is performed by loading a peer group advertisement from cache, instantiating a peer group object, and then applying for membership using a credential.

(3) Publish own advertisements
Peers announce its availability and resources to other peers in JXTA network by publishing its advertisement. JXTA allows for two types of publishing, local executed by method “publish” and remote executed by method “remotePublish”.

(4) Open an input pipe
Opening an input pipe refers to the creation of a class instance that represents an input end of a unidirectional pipe. After another peer creating the class instance of the output end of the pipe, messages can be passed from that peer to this peer through the pipe.

(5) Learn about other peers
Using JXTA peer discovery protocol (PDP) to find other peers information in the peergroup. For a shared computing system, it may be important to know who is active in the group to distribute mission properly and efficiently.

(6) Obtain pipe advertisements
Obtaining pipe advertisement let us know the existence of other peers’ pipes, which is the first step to send messages. The pipe advertisement can be discovered from network, retrieved from local cache, loaded from a user specified files, etc.

(7) Open output pipe
Opening the output pipe means binding to the output end of a pipe for which another peer has opened the input end. And in this way the peer can send messages to others through the pipe, which is illuminated in step4.

The operations discussed are typical of group structured or hierarchical P2P networks and systems, and also apply to the well-known file-sharing system. Since JXTA is designed and implemented with such structure in mind, especially group-based peer communication, it is desirable that the peers follow this idea.

4. CONCLUSIONS

Distrusted computing and parallel computing have the common characteristic of dividing large mission into smaller sub-tasks, and dispatching them for fulfilling, but one difference should be pointed out. In distrusted computing, every task is individual and the results of other tasks usually have no effect on it. While in parallel computing, there is great relationship between the parallel tasks. Every computing task is essential, and the result associates with each other. So confirming the computing mission property is first important, which determines the way to solve.

JXTA is an open P2P network platform, which provides an underlying framework for the executing and developing of P2P application. The paper here pays more attention to how to build distributed parallel computing model based JXTA. In the next research, the load of dispatching mission will be taken into account.

REFERENCES

[1]. Robert Flenner etal, *Java P2P UNLEASHED*, Sams Publishing, 2003

[2]. Daniel Brookshier etal, *Java P2P Programming*, Sams Publishing, 2003

[3]. *JXTA v2.0 Protocols Specification.Project JXTA*.http://www.jxta.org

Chun Liu is a postgraduate student studying for doctor degree in the area of Computer Application at Wuhan University of Technology with research interests of distributed parallel computing. She also graduated from the same university in 2005 with the Master degree, and during the process of graduate study, she has published 3 Journal papers.

Research on Distributed Web Services-Based Web Applications

Guangming Wang, Shuliang Tao
College of Computer and Information Engineering, Zhejiang Gongshang University,
Hangzhou, Zhejiang, 310035, China
E-mail: g.m.wang@hotmail.com, slingHappy@gmail.com

ABSTRACT

This paper first discusses the development, characteristic and architecture of Distributed Systems, and latter it moved on to Web Services technology architecture for details. The article find out the limits of the traditional distributed technology when apply to Internet, it also analyses the advantage of Web Services in Web based applications. In the final part, the author developed a User Management system which is based on Axis2.

Keywords: XML, UDDI, WSDL, SOAP, Web Services, Axis2

1. INTRODUCTION

Distributed technology usually refer to the distributed applications that development, deployment, management and maintenance on the network computing platform, with the aim of sharing resource and cooperative work. The distribute system mainly experienced following four stages: ①The first generation of distributed computing technology which is process-oriented, the core of this distributed computing technology is supporting for sharing information and application requirements. ② With the push of the first generation of distributed computing technology, it comes the hot movement from centralized computing model to distributed computing model, so called 3-tier Client/Server(C/S) architecture. It isolates the business logic from the client side, and the business logic becomes an independent tier. The 3-tier C/S architecture implements the tight client. ③ In early 1990s, with the combination of object and distributed technology, distributed object technology become popular. The typical representatives of distributed object technology is COM/DCOM(Component Object Model/Distributed Component Object Model), CORBA(Common Object Request Broker Architecture), JAVA/RMI(JAVA/Remote Method Invocation). These technologies are widely used in enterprise application development. ④ Web Services. With the popularity of Internet, Web based application become fashion. Traditional distributed object technologies have its limits in interoperation, because it is platform and implement standard dependence. Web Services perfectly solves these problems, it is the core of the next generation of distributed systems.

2. TRADITIONAL DISTRIBUTED OBJECT MODEL

In the domain of distributed computing, there are many distributed application solutions, the mainly used technology are DCOM(Distributed Component Object Model), CORBA(Common Object Request Broker Architecture), JAVA/RMI(JAVA/Remote Method Invocation), EJB(Enterprise Java Bean) and so on. They most run on commercial applications, which usually have tight coupling sub systems. The disadvantage of the traditional distributed object model is that any changes in sub systems may lead to damager in the related applications.

2.1 Characteristics of Traditional Distibuted Object Model

1) Tight coupling. Whatever the traditional distributed application is based on, DCOM, RMI, or CORBA, the server and the client are tight coupling. The both side know well where each other is. This “hard-linkable” has its own advantage, for example, it is efficient when security is considered. The disadvantage of the tight coupling system is obvious, that any changes in sub systems may lead to damager in the related applications. It also bring in tremendous difficult in maintenance and upgrade system.
2) Dependence on a single factory. CORBA and DCOM are already implemented on various platforms. Any solutions that based on these technologies need a single factory based technical solution. DCOM based applications must run on Windows platform, and CORBA based applications must use a particular ORB(Object Request Broker) products.

3. WEB SERVICES TECHNOLOGY

Web Services is an application that can be invoked through Internet standards and norms. It is based on HTTP, XML and open standards such as SOAP, It is the basic building block of Web-based distributed applications. In other words, the Web is to provide programming services to the users. URL through which users can visit it to obtain the necessary information. How to use different programming languages, different operating systems in different object model to build applications integrated together, and to transform them into an easy-to-use Web application software development process is currently facing a major challenge. Web Services is the response to this challenge. It is actually the components used in the development of Internet technology.

Table 1 Web Services technology architecture

(Null)	To be determined	Management	Quality of Service	Security
Routing, Reliability and Transaction	To be determined			
WSFL	Workflow			
UDDI	Service Discovery and Integration			
WSDL	Service Description			
SOAP/XML	Message Protocol			
HTTP, FTP, SMTP….	Transfer Protocol Networks			

3.1 Web Services Technology Architecture

Table1 describes the Web Services Technology Architecture. Among them, the bottom is the definition of good and has been widely used by the network layer and transport layer standards, such as IP, HTTP, SMTP. Web Services is among the four agreements related standards, Calling services, including SOAP(Simple Object Access Protocol, SOAP), WSDL(Web Service Definition Language, WSDL), UDDI(Universal Description, Discovery, and Integration, UDDI). And Web Services Flow Language(WSFL). Construction of the Web Service, the system does not necessarily require that all technical standards.

3.2 XML

XML (Extensible Markup Language; extensible markup language. 1998) [2] is a simple, flexible text format language. It comes from SGML (Standard Generalized Markup Language; standard generic markup language). Another widely used markup language HTML (Hypertext Markup Language; HTML) is also from SGML. HTML and XML can be simply seen as a subset of SGML. XML inherited the powerful SGML function also learned from HTML, after several years of vigorous development, XML has become the common language of computer systems, Internet data exchange standards. Meanwhile XML is the core content of Web Services technology. The entire Web Services, both UDDI and WSDL are Building on top of XML.

3.3 UDDI

UDDI(Universal Description, Discovery, and Integration, UDDI) is a Web Service-oriented information register center, which implement the standards and norms. The purpose of creating an UDDI is to public and discover web services. People use the UDDI specification to establish discovery service on web, these services provide a consistent service interface to all user. Web services that have been published need to be found through the program. UDDI standard defines the entrance station to support the operation of the UDDI API interface and API using XML description of the specific definition of the data structure. UDDI is a general designation of all public UDDI register service endpoint. It is an entia in logic, and it implemented with distributed system architecture in physics.

3.4 WSDL

Like DCOM, Web Service also requires a description file of how to use. The Web Service clients only know how to call that can correctly use it. Otherwise, the service can not be achieved by outside customers. WSDL (Web Service Description Language, WSDL)[4] is used to resolve this issue. As information format and communication protocols have been standardized, which is likely to be some kind of structured way to describe communication, Moreover, it has become increasingly important to achieve. WSDL defines an XML grammar. Web Service can be described as the gathered endpoint of communication for the exchange of information. WSDL provide documents for distributed system, and may be used for communications applications as automatic implementation of the details involved. WSDL has been designed to describe the Web Service and all Web Service methods. WSDL is based on the XML, so there is no problem of heterogeneous platforms. WSDL files in all of the parameters have their way of elaboration, it also contains information on the Web Service itself. WSDL itself is a standard protocol as SOAP. Simply, SOAP and SOAP server complete the information transmission. The primary role of WSDL is to provide the user-friendly Internet service.

3.5 SOAP

SOAP(Simple Object Access Protocol, SOAP) [3] is a lightweight protocol, that is used in distributed environment for the exchange of structured data. It uses XML technology to define an extensible framework which provides a message structure that can be transfer on various bottom protocols. This framework is designed semantic irrelevant. SOAP defines a way that message transfer from A to B. The SOAP message transfer is showed in Fig.1.

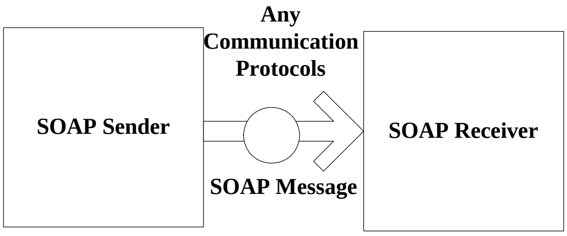


Fig.1. SOAP Message Transfer

3.5.1 SOAP Message Framework

The SOAP Message Framework is defined in SOAP specification . The Framework defines a serial XML elements, which is used to package any XML message for the transfer on various systems. The core XML element of SOAP Message Framework compose of Envelope, Header, Body, and Fault. A simple SOAP message can be represented as below:

```
<soap:Envelope
xmlns:soap=http://schemas.xmlsoap.org/soap/envelope/>
        <soap:Body>
        <x:SomeElement xmls:x=http://myNameSpace.com>
                    <x:sub1></x:sub1>
                    <x:sub2></x:sub2>
        </x:SomeElement>
        </soap:Body>
</soap:Envelope>
```

Include the Body element is the content the SOAP message really want to represent, it may be difference according to difference applications. The latest version is currently SOAP1.2, the version can be identified by the different namespace SOAP message uses.

3.5.2 The Working Principle of SOAP

To better understand how SOAP works, we could compare it with DCOM. DCOM deals with details of bottom network protocols, such as the communication between PROXY and STUB, life cycle management, object identifier. DCOM uses NDR(Network Data Representation) for data representation which is irrelevant with bottom platform when the client communicates with the server. Fig.2 describes the working principle of SOAP and DCOM.

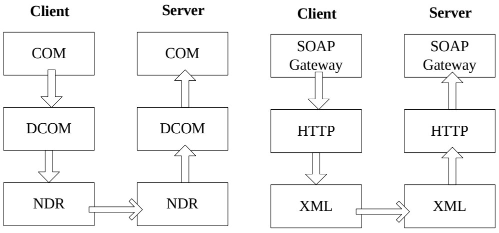


Fig.2. The Working Principle of SOAP and DCOM

SOAP and DCOM have similar working principle, simplicity is the main advantage of SOAP. SOAP use HTTP as a network communication protocol to receive and transmit data, it also use XML as a data format. SOAP replaces the format of NDR which DCOM uses, providing a higher level of abstraction, and has nothing to do with the environment and the platform.

When the client sends a request to a server, it first transforms the request into XML format, whatever the client platform is. The transformation is done by SOAP gateway. To ensure one and only of transmission parameters, methods, and return value, SOAP protocol uses a private mark table, so that server gateway can do correct analysis. After the request being encapsulated into HTTP request, it is sent to the server. The returned process is also similar, but opposite direction.

4. USER MANAGEMENT SYSTEM

4.1 System Overview

User management, authentication is indispensable in the enterprise application, the enterprise application integration subsystem might need all these functions. In traditional enterprise applications, each subsystem with a single database to support the functions of registration, cancellation and modification of users. If the subsystem need user information from another department, the user information must import to local database one by one, moreover, the data must be consistent with other departments. This consistency results in tight coupling between the systems, and is not conducive to enterprise application integration.

The User Management System developed by the author is implemented in Web Services. The system provides user registration, cancellation and modification within the enterprise. It solves the problem of asymmetric information, and provides a strong support for enterprise application integration.

4.2 System Analysis

User Management System is a part of enterprise application integration, the main function of the system is to provide users within the enterprise registration, inquiries, cancellation and modification and so on. The system can be divided into the following interfaces:

1) User register interface. This interface provides user registration within the enterprise.
2) User inquire interface. The inquire interface returns user information of the given department, and provides management supports for the manager.
3) User cancellation interface. This interface deletes a given user from the database.
4) User modification interface. The user modification interface modifies a given user’s information.

4.3 Database Design

The system has two tables, User and Department. See the following tables.

Table 2 User

ColumnName	DataType	Length	Null
UserId(pk)	Int	4	No
Name	Varchar	50	No
RegisterDate	Datetime	8	Yes
Password	Varchar	50	No
Department	Varchar	50	No
Age	Int	4	Yes
Sex	Bit	1	Yes

Table 3 Department

ColumnName	DataType	Length	Null
DepartmentId(pk)	Int	4	No
DepartmentName	Varchar	50	No
ManagerName	Varchar	50	Yes
FoundDate	Datetime	8	Yes
NumberOfPeople	Int	4	Yes

4.4 ADB(Axis2 Data Binding)

The objective of the ADB framework [1] is to provide a lightweight and simple schema compiler/ Java bean generator for Axis2. When the schema complier is invoked(one-way or another)it generates code depending on the following rules:

1) All named complex types become bean classes. Any attribute or element encapsulated in this complex type will become a field in the generated class.
2) All top level elements become classes.
3) Simple type restrictions are handled by replacing the relevant type with the basetype.

In the system, the type sub-element of the WSDL file defines the following data types:

```
<wsdl:types>
<xsd:schema
         targetNamespace="http://www.example.org/wsdl/">
  <xsd:complexType name="UserObject">
      <xsd:attribute name="UserId" type="xsd:int">
      </xsd:attribute>
      <xsd:attribute name="Name" type="xsd:string">
      </xsd:attribute>
      <xsd:attribute name="Password" type="xsd:Date">
      </xsd:attribute>
      <xsd:attribute name="RegisterDate" type="xsd:string">
      </xsd:attribute>
      <xsd:attribute name="Department" type="xsd:string">
      </xsd:attribute>
      <xsd:attribute name="Age" type="xsd:int">
      </xsd:attribute>
      <xsd:attribute name="Sex" type="xsd:boolean">
      </xsd:attribute>
  </xsd:complexType>
  <xsd:element name="User" type="tns:UserObject">
  </xsd:element>
  <xsd:element name="Result" type="xsd:boolean">
  </xsd:element>
 <xsd:complexType name="OldNewUserObject">
     <xsd:sequence>
      <xsd:element name="newUser" type="tns:UserObject">
      </xsd:element>
      <xsd:element name="oldUser" type="tns:UserObject">
      </xsd:element>
    </xsd:sequence>
 </xsd:complexType>
 <xsd:element name="OldNewUser"
              type="tns:OldNewUserObject">
 </xsd:element>
</xsd:schema>
</wsdl:types>
```

This XML Schema defines two complex types, one is UserObject which denotes a user, another is OldNewUserObject that denotes two users. The top elements the Schema defines ara OldNewUser, Result and User. The OldNewUser contains an OldNewUserObject object, and the User contains a UserObject. The Result denotes the return type that every interface returns. After the process of ADB, it produces some Java beans, UserObject.java, OldNewUserObject.java, OldNewUser.java, Result.java and User.java. The ADB also produces the server side Skeleton and client side Stub, which can be used in implementing Web Service. The client Stub makes the call of a services which is remote likes locally.

4.5 User Management System Architecture

The follow Fig.3 shows the architecture of the System:

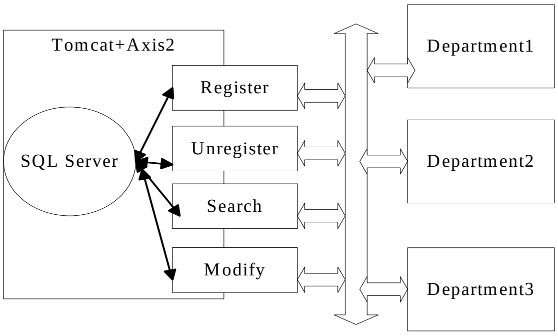


Fig.3. The Architecture of User Management

The system has four interfaces(Register, Unregister, Search, Modify)which implements the operations of registration, cancellation, searching and modification of user. The system architecture reflects the thinking of SOA framework. The architecture of SOA(Service-Oriented Architecture) shows in Fig.4 .

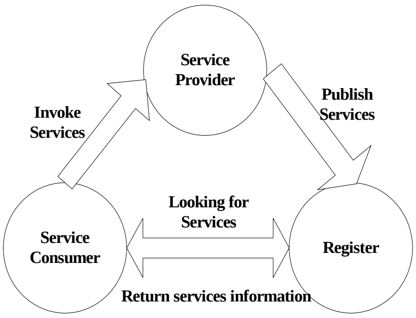


Fig.4. The Architecture of SOA

The services provider is the interfaces of the User Management System, the clients which use these interface are services consumer. The register center contains all information about the services that can be invoked by clients which obtain services information through inquiring the register center.

5. CONCLUSIONS

In this paper we develop a application using Web Services technology, it is a part of enterprise integration. Web Services technology is the future of software development. The greatness of SOAP lies in its standards, but not as a technology. The open XML message mechanism combined with problems of HTTP firewall, SOAP unifies them very cleverly, so that they will become the standard expression between the client and server, and makes distributed computing technology can be developed. Web Services is the building constructed in such technical.

REFERENCES

[1] Axis2, http://ws.apache.org/axis2/.

[2] W3C, “XML Schema Primer 0: Primer Second Edition,” http://www.w3.org/TR/xmlschema-0/.

[3] W3C, “SOAP Version 1.2 Part 0: Primer,” http://www.w3.org/TR/2003/REC-soap12-part0-20030624/.

[4] W3C, “Web Services Description Language (WSDL) Version 2.0,” W3C working Draft, November 2003, http://www.w3.org.

[5] James R. Groff, Paul N. Weinberg, *SQL: The Complete Reference, Second Edition*, Beijing: Publishing House of Electronics Industry, 2006.

[6] Joey F. George, Dinesh Batra Joseph S. Valacich, Jeffery A. Hoffer, *OBJECT-ORIENTED SYSTEMS ANALYSIS AND DESIGN*, Beijing Tsinghua University Press, 2004.11.

Research on Discovery Mechanism of Web Services Based on Ontology

Tianhuang Chen, Wei Zhang
Computer Applied Technology, Wuhan University of Technology
Wuhan, Hubei, China
Email: thchen@whut.edu.cn

ABSTRACT

With the fast development of Internet, the increasing availability of web services on the Web raises new and challenging search problems: efficiently locating functionality-desired web services among numbers of web services and selecting a best one among large numbers of functionality-similar web services. While these just are the task of web service discovering. This paper proposed an approach to support semantic web service discovery based on ontology. The approach combines with Web Services and Semantic Web and provides a solution for above. It constructs ontology with OWL-S and extends existing UDDI matching mechanism, in order to improve the result of service discovering.

Keywords: Web Service, Semantic Web Service, Ontology, OWL-S, UDDI

1. INTRODUCTION

As the rapid development of web services, web services system should be higher dynamic, hale, efficient and secure, traditional Web Services' localization is more and more evident. Traditional web service discovering technology is done by keyword match based on the syntactic description of web service. Such web service discovering technology can not capture the semantic information of web service and is lack of intelligence, so it can not meet the growing demand of people. In order to get a better result of discovering, a more efficient and perfect web service discovering technology is needed.

Web service is an important trend in the World Wide Web (WWW). Applying of Semantic Web technology on Web services gives birth to another new technology–Semantic Web Services. Ontology is the core technology of Semantic Web, and can be used to describe the semantics of Web services and make the service understandable by computers or agents, so as to enable agents or computers to discover, invoke and compose Web services. The motivation of this work is to introduce the background of web service and its discovery mechanism, semantic web, ontology and the services description language such as OWL and OWL-S, the services discovery mechanism UDDI.

The rest of the paper is structured as follows. Section 2 gives a brief introduce of concept of semantic web, UDDI Search Mechanism and how to extend UDDI with OWL-S. Section 3 discusses the basic concept of Description Logics, the Web Ontology Language and OWL-S. The ontology-based semantic matchmaking algorithm is presented in Section 4. The last section draws the conclusions and future work.

2. SEMANTIC WEB AND UDDI SEARCH MECHANISM

2.1 Semantic Web

Semantic Web is a new concept formally proposed by the WWW founder Berners-Lee in 2001[1]. The main goal of studying semantic Web is to expand current WWW, and the information is represented with the form that computers can understand and deal with, so that all information in network has semantics. The Semantic Web Services environment is built up, in order that computers can understand communication contents each other to realize the Web Services' automation. The Semantic Web approach is to develop languages and mechanisms for expressing information in machine understandable form. These languages emerge in form of ontology that will describe web resources easily processed by computer programs.

2.2 UDDI Search Mechanism

The Universal Description, Discovery and Integration (UDDI) is a Web-based distributed registry standard for SOA [2]. It is one of the central elements of the interoperable framework and an OASIS standard with major backers including IBM, Microsoft and government agencies. UDDI [3] is an industrial initiative aimed to create an Internet-wide network of registries of web services for enabling businesses to quickly, easily, and dynamically discover web services and interact with one another.UDDI allows a wide range of searches: services can be searched by name, by location,by business, by bindings or by TModels. The UDDI specification defines a set of data models for describing Web services. The core data model consists of hierarchical objects with static data fields for describing the service provider (businessEntity), the Web service (businessService), and the service binding (bindingTemplate). Current matching technology of Web Services is based on UDDI, which does not make any use of semantic information, the matching is based on key words and syntactic,it can not distinguish the information with the same syntactic and differentit semantic,it also can not distinguish the information with the same semantic and differentit syntactic.So it does not match the request with the capabilities.

For example a car selling service may describe itself as “New Car Dealers” which is an entry in NAICS, but a search for “Automobile Dealers” services will not identify the car selling service despite the fact that “New Car Dealers” is a subtype of “Automobile Dealers”. Such semantic matching problem can be solved if we use OWL, RDF etc instead of XML. The second problem with UDDI is the lack of a power full search mechanism. Search by Category information is the only way to search for services, however, the search may produce lot of results with may be of no interest. For example when searching for “Automobile Dealer”, you may not be interested in dealers who don’t accept a pre-authorized loan or credit cards as method of payments. In order to produce more precise search results, the search mechanism should not only take the taxonomy information into account but also the inputs and outputs of web services. The search mechanism resulted in combining the semantic base matching and the capability search is far more effective than the current search mechanism. OWL-S provides both semantic matching capability and capability base searching, hence is a perfect candidate.

2.3 Extending UDDI with OWL-S

The provision in the UDDI data structures to refer to external information is limited to the overviewDoc element that appears in the following places in the UDDI data structures:

a.the instanceDetails element that is part of the tModelInstanceInfo element
b. the tModel

The overviewDoc element consists of zero or more descriptions and an optional overviewURL. We make use of this overviewURL to refer to the external descriptions of services and requests. It is common for the overview URL to be dereferenced using HTTP GET. In the two data-structures that support capturing of external descriptions via overviewURLs, TModels are recommended for those instances of external descriptions that service providers/requesters would like to share with others (such as standardized representations). TModelInstanceInfos, on the other hand, are recommended for representing those instances of external descriptions that are specific to a given service/request. In this work, we have chosen the TModel approach (apart from keeping things simple, using Tmodels for external descriptions makes the inquiry API interfaces simple. It is also consistent with how WSDL definitions are published currently). An external description tModel should be categorized with the appropriate tModel to indicate the type of external description, such as OWL-S. For doing this, we have introduced new categorization, the DescribedUsing categorization (in this paper, for ease of explanation, using a convention such as XXXX…,YYYY… , CCCC…etc. for tModel Keys of tModels thatwe designed). It is as follows:

```
<tModel
tModelKey="uuid:XXXXXXXX-XXXX-XXXXXXXX-
XXXXXXXXXXXX">
<name>urn:x-ibm:DescribedUsing</name>
<description xml:lang="en">Used to categorise a tModel
by/with a particular external description
type/format.</description>
<overviewDoc>
<overviewURL>…</overviewURL>
</overviewDoc>
<categoryBag>
<keyedReference tModelKey="uuid:C1ACF26D-9672-
4404-9D70-39B756E62AB4"
keyName=”types”  keyValue="categorization"/>
</categoryBag>
</tModel>
```

The OWL-S/UDDI mapping is extended to reflect the latest developments in both UDDI and OWL-S. Fig.1 shows the resulting OWL-S/UDDI mapping. Furthermore we enhanced the UDDI API with the OWL-S/UDDI mapping functionality, so that OWL-S Profiles can be converted into UDDI advertisements and published using the same API.

3. DESCRIPTION LOGICS

3.1 Description Logics

Research in the field of knowledge representation and reasoning led to intelligent systems which have the ability to find implicit consequences of its explicitly represented knowledge. These systems are characterized as knowledge-based systems. One approach of these systems evolved into what we call today Description Logics (DL).Description Logics can be characterized by the following three properties:

a. The basic syntactic building blocks are atomic concepts, atomic roles and individuals.
b. Expressive power of the language is restricted by using a rather small set of constructors for building complex concepts and roles from existing concepts and roles.
c. Implicit knowledge about concepts and individuals can automatically be found with reasoning techniques.

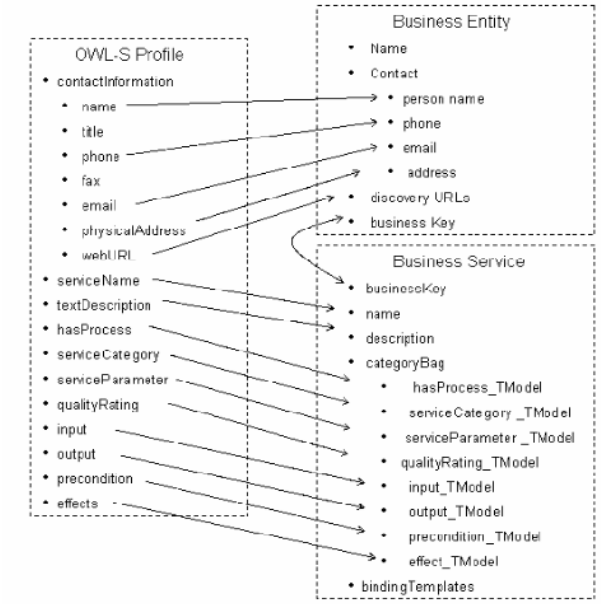


Fig.1. Mapping between OWL-S Profile and UDDI

3.2 Ontology

Ontologies are the basis for shared conceptualization of a domain [4], and comprise of concepts with their relationships and properties.

An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them.

Why would someone want to develop an ontology? Some of the reasons are[5]:

- To share common understanding of the structure of information among people or
  software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

Ontology is a rising branch of artificial intelligence. It can provide powerful supporting to semantic web and web service. It the basis of So the coding environment of ontology becomes more and more important. In the context of the Semantic Web, the term ontology occurs frequently. The term is borrowed from philosophy where ontology denotes a systematic account of existence. In the context of the Semantic Web, ontology denotes a description of the concepts and relationships that exist for a special domain. Therefore, ontology is nothing more than a terminology for a given domain of interest. These ontology allow computer agents and programs to unambiguously interpret the meaning of web resources.

3.3 Brief Overview of OWL-S

OWL-S is a Semantic Web Services description language that enriches Web Services descriptions with semantic information from OWL ontology and the Semantic Web[6][7]. It is put forward by many organizational researcher is a Web Services description language based on Ontology, it have a stronger information expression capability than WSDL, it support semantic consequence function and it make Web Services search more exact and comprehensive. OWL-S is a set of top-level ontology written in OWL specifically for the description of Web services. It is designed to enable the automation of service discovery, service invocation, and service composition and interoperation. The OWL-S Profile, Grounding, and Process ontology are created for these purposes respectively [8][9]. UDDI will only deal with the Profile and the Grounding. OWL-S organizes a service description into four conceptual areas: the process model, the profile, the grounding, and the service. (shown in Fig.2):

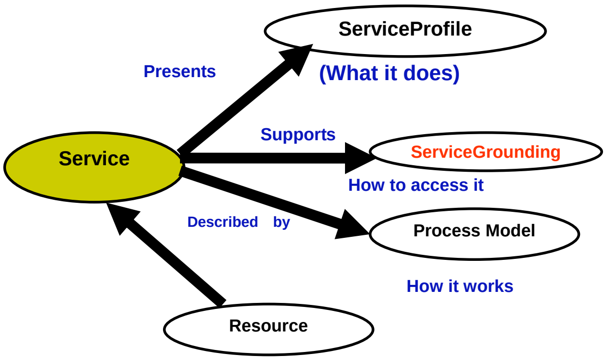


Fig.2. Top-level structure of OWL-S

A process model describes how a service performs its tasks. It includes information about inputs, outputs (including a specification of the conditions under which various outputs will occur), preconditions (circumstances that must hold before a service can be used), and results (changes brought about by a service). The process model differentiates between composite, atomic, and simple processes. For a composite process, the process model shows how it breaks down into simpler component processes, and the flow of control and data between them. Atomic processes are essentially ``black boxes" of functionality, and simple processes are abstract process descriptions that can relate to other composite or atomic processes.

A profile provides a general description of a web service, intended to be published and shared to facilitate service discovery. Profiles can include both functional properties (inputs, outputs, preconditions, and results) and nonfunctional properties (service name, text description, contact information, service category, and additional service parameters).The functional properties are derived from the process model, but it is not necessary to include all the functional properties from the process model in a profile. A simplified view can be provided for service discovery, on the assumption that the service consumer would eventually look at the process model to achieve a full understanding of how the service works.

A grounding specifies how a service is invoked, by detailing how the atomic processes in a service’s process model map onto a concrete messaging protocol. OWL-S provides for different types of groundings to be used, but the only type developed to date is the WSDL grounding, which allows any WS to be marked up as an SWS using OWL-S.

A service simply binds the other parts together into a unit that can be published and invoked. It is important to understand that the different parts of a service can be reused and connected in various ways. For example, a service provider may connect its process model with several profiles in order to provide customized advertisements to different communities of service consumers. A different service provider, providing a similar service, may reuse the same process model, possibly as part of a larger composite process, and connect it to a different grounding. The relationships between service components are modeled using properties such as presents (Service-to-Profile), describedBy (Service-to-Process Model), and supports (Service-to-Grounding).

4. MATCHING ALGORITHM

With the growing number of Web Services,the web is moving from being a collection of pages toward a collection of services that interoperate through the Internet. A fundamental step toward this interoperation is the ability of locating services. This needs matching engine to find services that satisfy request. Matching is very important in Web Services.
The matching algorithm we used is based on the algorithm presented in[10]. The algorithm defines a more flexible matching mechanism based on the OWL’s. When a request is submitted, the algorithm finds a appropriate service by first matching the outputs of the request against the outputs of the published advertisements, and then, if any advertisement is matched after the output phase, the inputs of the request are matched against the inputs of the advertisements matched during the output phase.

In the matching algorithm, the degree of match between two outputs or two inputs depends on the match between the concepts that represents by them. The matching between the concepts is not syntactic, but it is based on the relation between these concepts in their OWL ontology. For example consider an advertisement, of a vehicle selling service, whose output is specified as Vehicle and a request whose output is specified as Car. Although there is no exact match between the output of the request and the advertisement, given an ontology fragment as show in Fig.3, the matching algorithm recognizes a match because Vehicle subsumes Car. The matching algorithm recognizes four degrees of match between two concepts. Let us assume OutR represents the concepts of an output of a request, and OutA that of an advertisement. The degree of match between OutR and OutA is as follows.

exact: If OutR and OutA are same or if OutR is an immediate subclass of OutA.For example given the ontology fragment like Fig.3, the degree of match between a request whose output is Sedan and an advertisement whose output is Car is exact.

plug in: If OutA subsumes OutR, then OutA is assumed to encompass OutR or in other words OutA can be plugged instead of OutR. For example we can assume a service selling Vehicle would also sell SUVs. However this match is inferior than the exact match because there is no guarantee that a Vehicle seller will sell every type of Vehicle.

subsume: If OutR subsumes OutA, then the provider may or may not completely satisfy the requester. Hence this match is inferior than the plug in match.

fail: A match is a fail if there is no subsumption relation between OutA and OutR.

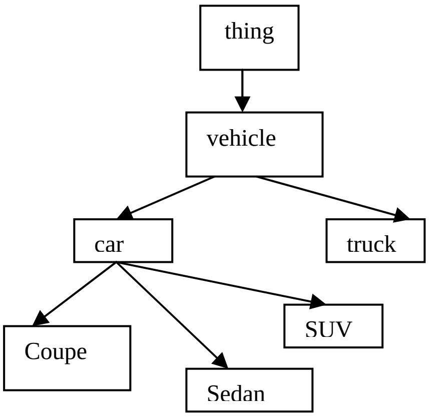


Fig.3. Vehicle Ontology

5. CONCLUSIONS

This paper deeply researched and discussed current web service discovery method and interrelated technologies, then emphasized the importance of combining web service with semantic web. At the same time,a solution is proposed. This solution imports semantic information into UDDI by defining special ontology and utilizes ontology technology to describe semantic web service and match the services.So, it can widly use a large number of existing UDDI to extend the range of web service,and make the web service discovery more effective.

REFERENCES

[1] Tim Berners_Lee, James Hendler and Ora Lassila. The Semantic web: A new form of Web content that is meaning to computers will unleash a revolution of new possibilities. Scientific American. http://www.sciam.com/2001/0501-issue/0501berners_Lee. 2001-05-01

[2] K. Sivashanmugam, K. Verma, A. Sheth, and J. Miller.Adding Semantics to Web Services Standards. http://lsdis.cs.uga.edu/lib/download/SVSM03-ICWS-final.pdf

[3] UDDI : The UDDI Technical White Paper , http://www.uddi.org, 2000

[4] Gruber, T.R. “A Translation Approach to Portable Ontology Specifications.” *Knowledge Acquisition,* 5(2), 199-220, 1993.

[5] Natalya F. Noy and Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology.http://protege.stanford.edu/publications/ontolgy development/ontology101.pdf

[6] W3C, "OWL Web Ontology Language Guide," 2004. http://www.w3.org/TR/owl-guide/

[7] W3C, "OWL Web Ontology Language Use Cases and Requirements,"2004.http://www.w3.org/TR/webont-req/

[8] Paolucci M., Kawamura T., Payne T., Sycara K. “Semantic Matching of Web Services Capabilities,” In:Proc.of *the 1st International Semantic Web Conference (ISWC)*, 2002.

[9] SWWS Home Page. http://swws. semanticweb.org

[10] Paolucci et al. “Semantic Matching of Web Services Capabilities.” In *Proceedings of the 1st International Semantic Web Conference* (ISWC2002)

Modeling Web Service Compositions with CSP

Wenhui Sun, Feng Liu, Jinyu Zhang, Gang Dai
Computer college, Beijing Jiaotong University
Beijing, 100044, P.R.China
Email: whsun1@bjtu.edu.cn

ABSTRACT

Web services composition is the hotspot research in the field of web services. BPEL4WS is a de facto standard description language for web services composition. Communicating sequential processes is a model for describing parallel composition of communicating sequential processes. In this paper we discuss a case study with BPEL. And then we use CSP to describe some part of BPEL such as port behavior, role behavior and processes interactions.

Keywords: Web Services, BPEL4WS, Communicating Sequential Processes

1. INTRODUCTION

Nowadays service oriented computing(SOC) is the hot research for the next tide of internet. IBM etc provokes that software is service. In service oriented computing (SOC), developers use services as fundamental elements in their application development processes. SOC thus offers three native capabilities, description, discovery, and communication[1]. Web services are a typical SOC example: developers implement SOC native capabilities using Web Services Description Language, Universal Description, Discovery, and Integration (for discovery), and SOAP (for communication).

Service composition requirements differ from those of mainstream component based software development. A composition mechanism must therefore satisfy several requirements: connectivity, nonfunctional quality of service properties, correctness, and scalability[2]. Our contribution is to use CSP to model web service compositions. And we compare it with BPEL.

CSP has an advantage over π-calculus in terms of its denotational semantics and refinement relations. These qualities make CSP a suitable process algebra for modeling complex service orchestration and choreography. Furthermore, CSP is supported by an industrial strength automated model checker FDR[8], which is crucial in workflow specification, refinement and verification.
The rest of this paper is organized as follows. Section 2 introduces web service composition approaches such as BPEL and CSP. Section 3 gives a case study with BPEL. And then we use CSP to describe some part of BPEL such as port behavior, role behavior and process interactions. Section 4 gives related work and future work.

2. WEB SERVICE COMPOSITIO APPROACHES

Once web services' native capabilities were fully developed, service composition approaches began emerging. Because the first generation composition languages IBM's Web Service Flow Language (WSFL) and BEA Systems' Web Services Choreography Interface (WSCI) were incompatible, researchers developed second generation languages, such as the Business Process Execution Language for Web Services (BPEL4WS, or BPEL), which combines WSFL and WSCI with Microsoft's XLANG specification.

A key research challenge concerns in web composition of web services, in order to construct new web services with desired properties or capabilities. The fundamental work[3] in this area has centered on three models, each coming with a different approach to the composition problem. One is OWL-S. The second is the "Roman" model[4] which model web services as finite state automata with transitions labeled by these activities. The third is the conversation model[5] which focuses on messages passed between web services, and again uses finite state automata to model the internal processing of a service. Here we use CSP to model web service composition. There is little literature to use this describing tool.

2.1 BPEL

BPEL[6] is an XML language that supports process oriented service composition. BPEL composition interacts with a web services' subset to achieve a given task. In BPEL, the composition result is called a process, participating services are partners, and message exchange or intermediate result transformation is called an activity. A process thus consists of a set of activities. A process interacts with external partner services through a WSDL interface.

BPEL has several element groups, but the basic ones are:

- Process initiation: <process>
- Definition of services participating in composition: <partnerLink>
- Synchronous and asynchronous calls: <invoke>,<invoke>…<receive>
- Intermediate variables and results manipulation: <variable>, <assign>, <copy>
- Error handling: <scope>, <faultHandlers>
- Sequential and parallel execution: <sequence>, <flow>
- Logic control: <switch>

As an example, we'll model the composition of three services. Service A is called synchronously and starts a process. Two asynchronous services, B and C, are then called in parallel using Service A's output as their input. The process waits for their completion and then makes a decision based on the results. The BPEL code for this composition follows.

```
<process name="test">
  <partnerLinks>
    <partnerLink name="client"/>
    <partnerLink name="serviceA"/>
    <partnerLink name="serviceB"/>
    <partnerLink name="serviceC"/>
  </partnerLinks>
  <variables>
    <variable name="processInput"/>
```

* This research is supported by the National High Technology Development 863 Program under Grant No. 2005AA147110.

```
    <variable name=”AInput”/>
    <variable name=”AOutput”/>
    <variable name=”BCInput”/>
    <variable name=”BOutput”/>
    <variable name=”COutput”/>
    <variable name=”processOutput”/>
    <variable name=”AError”/>
  </variables>
  <sequence>
    <receive name=”receiveInput” variable=”input”/>
    <assign><copy>
      <from variable=”processInput”/>
      <to variable=”AInput”/>
    </copy></assign>
  <scope>
    <faultHandlers>
      <catch faultName=”faultA” faultvariable=”AError”/>
    </faultHandlers>
  <sequence>
      <invoke name=”invokeA” partnerLink=”serviceA”
            inputVariable=”AInput”
          outputVariable=”AOutput”/>
  </sequence>
  </scope>
  <assign><copy>
    <from variable=”AOutput”/>
    <to variable=”BCInput”/>
  </copy></assign>
  <flow>
    <sequence>
      <invoke name=”invokeB” partnerLink=”serviceB”
                inputVariable=”BCInput”/>
      <receive                    name=”receive_invokeB”
                    partnerLink=”serviceB”
                    Variable=”BOutput”/>
      </sequence>
      <sequence>
      <invoke name=”invokeC” partnerLink=”serviceC”
                inputVariable=”BCInput”/>
      <receive                    name=”receive_invokeC”
                partnerLink=”serviceC”
                Variable=”COutput”/>
    </sequence>
    </flow>
    <switch><case>
      <!- assign value to processOutput->
      </case></switch>
      <invoke name=”reply” partnerLink=”client”
                  inputVariable=”processOutput” />
      </sequence>
    </process>
```

2.2 Communicating Sequential Processes[7] Concepts and notations

A command list specifies sequential execution of its constituent commands in the order written. Each declaration introduces a fresh variable with a scope which extends from its declaration to the end of the command list. For example, X(i:1..n)::CL stands for

$X(1)::CL_1||X(2)::CL_2||\ldots||X(n)::CL_n$

Where each CL_j is formed from CL by replacing every occurrence of the bound variable i by the numeral j. After all such expansions, each process label in a parallel command must occur only once and the processes must be well formed and disjoint. A parallel command specifies concurrent execution of its constituent processes. They all start simultaneously and the parallel command terminates successfully only if and when they have all successfully terminated.

Input and output commands specify communication between two concurrently operating sequential processes. Communication occurs between two processes of a parallel command whenever (1)an input command in one process specifies as its source the process name of other process; (2)an output command in the other process specifies as its destination the process name of the first process; and (3)the target variable of the input command matches the value denoted by the expression of the output command.

Examples:

(1) X?(x,y) from process named X, input a pair of values and assign them to x and y

(2) DIV!(3*a+b,13) to process DIV, output the two specified values.

A guarded command is executed only if and when the execution of its guard does not fail. An alternative command specifies execution of exactly one of its constituent guarded commands. A repetitive command specifies as many iterations as possible of its constituent alternative command. Consequently, when all guars fail, the repetitive command terminates with no effect. Otherwise, the alternative command is executed once and then the whole repetitive command is executed again.

Examples:

(1). [x≥y→m:=x□y≥x→m:=y]
If x≥y, assign x to m; if y≥x assign y to m; if both x≥y and y≥x, either assignment can be executed.

(2). *[n:integer;
X?insert(n)→INSERT□n:integer;X?has(n)→SEARCH;
X!(i<size)]

On each iteration this command accepts from X either (a) a request to “insert(n),” or (b) a question “has(n),” to which it outputs an answer back to X. The choice between (a) and (b) is made by the next output command in X. The repetitive command terminates when X does. If X sends a nonmatching message, deadlock will result.

P;Q represents sequencing execution processes. P∏Q represents inner choice of either P or Q.

3. A CASE STUDY WITH USING CSP

As show in Fig. 1, the travel agency’s BPEL work flow is composed of railway, ship, airline, finance and credit agency. Airline agency is a composite web services which is composed of several airline companies’ service. The other web services are atomic web services.

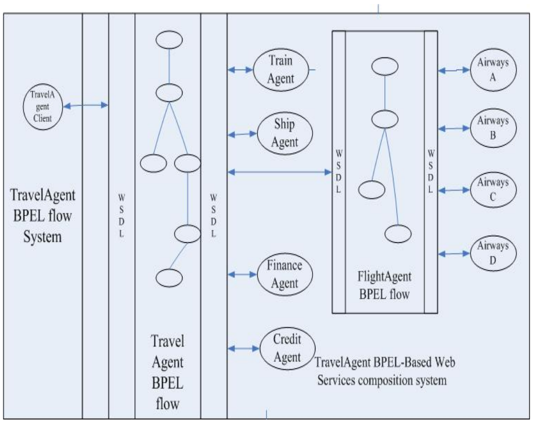


Fig.1. Composition system of TravelAgent

The travel agency receives client's ticket request. And according to the ticket type (railway, airline and ship), train agent, airline agent, and ship agent is respectively called. If the airline ticket request is received, a personal credit lookup service is provided by calling credit agent. If personal credit grade is good from the result returned from credit service, the ticket request is received. And then Airway agent is called. The behavior of travel agency has online payment operation. If the ticket request is permitted, the client may ask for online payment of ticket fee. At the same time, the travel agency calls finance agent' service and deduce the service fee of ticket ordering from the client's online finance account. And message interaction coordinated by TravelAgent BPEL flow is shown in Fig. 2.

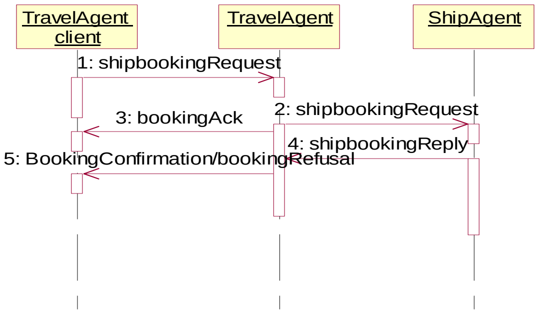


Fig.2. Message interaction coordinated by TravelAgent BPEL flow

At first we model the travel agency' port behavior:
TravelAgenEventSet={clientbookingRequest?(),clientbookingAck!(),clientbookingConfirmation!(),clientpayingAck!(), clientbookingRefusal!(), clientpayingRequest?(), clientpayingConfirmation!(),clientpayingRefusal!()}

Processes TravelAgentPort and PayingPort represent TravelAgentPort port behavior.

proc TravelAgentPort=clientbookingRequest?()→ clientbookingAck!()→(clientbookingConfirmatio !()→PayingPort∏clientbookingRefusal!()→ TravelAgentPort)

Proc PayingPort=clientpayingRequest?()→clientpayingAck!()→(clientpayingConfirmation!()→TravelAgentPort∏clientpayingRefusal!()→TravelAgentPor)

And then we model the flight agent's role behavior.
FlightAgentRoleEventSet={flightbookingRequest?(), flightbookingAck!(),flightbookingConfirmation!(), flightbookingRefusal!()}

proc FlightAgentRole=fightbookingRequest?()→flightbookingAck!()→(flightbookingConfirmation!()→FlightAgentRole∏flightbookingRefusal!()→FlightAgentRole)

Such other role behaviors as TravelAgentClientRole, ShipAgentRole and TrainAgentRole are defined as the same as the above.

And finally we model the five interactive processes between each role.

proc Glue=clientbookingRequest?()→clientbookingAck!()→(ClientFlightGlue∏ClientShipGlue∏ClientTrainGlue)

proc ClientpayingGlue=clientpayingRequest?()→clientpayingAck!()→financeRequest!()→financeReply?() →((clientpayingConfirmation!()→Glue) ∏ (clientpayingRefusal!()→Glue))

proc ClientFlightGlue=creditRequest!()→creditReply?()→((flightbookingRequest!()→flightbookingAck?()→((flightbookingConfirmation!()→clientbookingConfirmation!()→ClientpayingGlue)□(flightbookingRefusal?() →clientbookingRefusal!() →Glue)))∏(clientbookingRefusal!() →Glue))

proc ClientShipGlue=shipbookingRequest!()→shipbookReply?()→((clientbookingConfirmation!() →ClientpayingGlue)∏(clientbookingRefusal!() →Glue))

proc ClientTrainGlue=trainbookingRequest!()→trainbookingAck?()→((trainbookingConfirmation?→clientbookingConfirmation!()→ClientpayingGlue) □(trainbookingRefusal?()→clientbookingRefusal!()→Glu e))

4. RELATED WORK

Currently little research has been done into the application of CSP to BPEL. The only approach that has applied CSP in workflow process[12] do so as an extension of abstract machine notation for process specification within the domain of compositional information systems.

Other process algebras used to model workflow patterns include π-calculus[9] and CCS[10], a subset of π-calculus. These formalizations did not focus on formal verification and they did not demonstrate their applications. Moreover the semantics of π-calculus and CCS do not provide a refinement relation, which is crucial in formal specification and verification.

While Puhlmann et al.'s formalization is not oriented towards automated verification[9] and Stefansen's does not conform to standard CCS and is not machine readable[10], we have implemented our CSP models using standard CSP syntax; it is possible to translate our models directly into CSP$_M$, the machine readable dialect of CSP[11] , for model checking.

5. CONCLUSIONS

Besides the two above web service composition approaches, there are the other approaches. One is semantic web(OWL-S). OWL-S models services using a three part ontology:

- a service profile describes what the service requires from users and what it gives them;
- a service model specifies how the service works; and
- a service grounding gives information on how to use the service.

The second is algebraic process composition. Algebraic service composition aims to introduce much simpler descriptions than other approaches, and to model services as mobile processes to ensure verification of properties such as safety, liveliness, and resource management.

The third is Petri Nets. Petri nets are a well established process modeling approach. A Petri net is a directed, connected, and bipartite graph in which nodes represent places and transitions, and tokens occupy places. Other approaches for web service composition include model checking, modeling service compositions as Mealy machines, and automatic composition of finite state machines.

Here we use CSP to model web service composition. Our contribution is to use CSP to model web service composition with a case study.

REFERENCES

[1] M.P.Papazoglou and D.georgakopoulos, “Service OrientedComputing,”*Comm.ACM*,vol.46,No.0,Oct,2003

[2] Nikola Milanovic and Miroslaw malek, “ Current Solutions for Web Service Composition,*”in IEEE Internet computing*, Nov,2004.

[3] Richard Hull, “ Web Services Composition: A Story of Models, Automata, and Logics,” *in IEEE Proceeding, Service Computing,*2005.

[4] D. Berardi, D. Calvanese, G.De Giacomo, R.Hull, and M.Mecella, “Automatic Composition of e-Services that Export their Behavior,” ICSOC 2003, Vol.2910 of LNCS, 2003.

[5] T.Bultan,X.Fu, R.Hull, and J.Su, “Conversation specification: A new approach to Design and Analysis of E-Service Composition,” *in Proceeding WWW 2003, ACM*, 2003.

[6] BPELhttp://www.ibm.com/developerworks/library/ws-bpel.

[7] C.A.R.Hoare,“Communicating Sequential Processes,” *Communication of the ACM*, Vol.21,No.8, Aug,1978.

[8] *Formal Systems(Europe) Ltd. Failures Divergences Refinement,FDR2 User Manua*l, 1998, www.fsel.com.

[9] F.Puhlmann and M.Weske. “Using the π-calculus for Formalizing Workflow Patterns,”*in BPM 2005, LNCS*, Vol.3469, 2005

[10] C.Stefansen,”SMAWL: A SMALL workflow language based on CCS,” *Technical Report TR-06-05*, Harvard University, Mar. 2005

[11] A.Roscoe,*The Theory and Practice of Concurrency.* Prentice Hall, 1998

[12] S.Stupnikov,L.Kalinichenko,and J.S.Dong. “Applying CSP-like Workflow Process Specifications for their Refinement in AMN by Preexisting Workflows,” *in Proceedings of ADBIS’2002*, Sept, 2002.

Building Web Service with JAX-WS

Liang Huang, Qingping Guo
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei 430063, China
Email: whutdarren@whut.edu.cn

ABSTRACT

Web Services is a cool technology, which promises to make integration of applications both within the enterprise and between different enterprises significantly easier and cheaper than before. This paper is about how to build Web service with the new technology JAX-WS. The paper first introduces the Web Service and then talks about JAX-WS, JAXB. The latter part of this paper discusses the annotation-driven programming in detail and also gives some examples on how to write web service on both server side and client side.

Keywords: Web Service, JAX-WS, JAXB, Nnotation-Driven Programming

1. INTRODUCTION

Observing the Internet, the advanced computing and the advance processors that are being built today, you will find that there is a trend to carry forth the modularity principles that were outlined in computer science many years ago. We can realize that distribution of the components, loose coupling of modular components is the way that we will be building computer systems in the future. These components are sometimes called distributed, because they might be geographically distributed in terms of no temporal coupling between them. They may also have another characteristic that different components of these may not come from the same vendor. It's what we call heterogeneity, what means that you've got a variety of sources of the different parts, and that's just going to be a realistic thing[1]. For example, you can build a system partly using Sun products and partly using Microsoft products. In this case, Web service is a good choice.

The W3C defines that a Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.

Now, Web Service is neither a single large body of code nor a single large protocol. Web service architecture involves many layered and interrelated technologies.

2. JAX-WS

Now, much existing Java Web Service is built with JAX-RPC (Java API for XML-based RPC). But JAX-RPC has some problems such as Supporting only SOAP 1.1 over HTTP, Limited XML Schema support, Little control of mapping Java and WSDL/XML Schema, Large non-portable applications, Large runtime; too complex to development Web service and required Servlet container. JAX-WS (Java API for XML Web Services) 2.0 is our new Web service API and there are lots of improvements over what we had in J2EE 1.4. It's also notable that it's not only delivered as part of the Java EE platform, but also in the SE platform. An important change from the past is that JAX-WS has embraced plain old Java object (POJO) concepts and so feels much aligned with things like EJB 3.0. As a follow-on release of JAX-RPC 1.1, JAX-WS simplifies the task of developing web services using Java technology. It addresses some of the issues in JAX-RPC 1.1 by providing support for multiple protocols such as SOAP 1.1, SOAP 1.2, XML, and by providing a facility for supporting additional protocols along with HTTP. JAX-WS uses JAXB 2.0 for data binding and supports customizations to control generated service endpoint interfaces. With its support for annotations, JAX-WS simplifies web service development and reduces the size of runtime JAR files[2].

Fig.1 shows the JAX-WS layered architecture[3]. The application code at the top most of time calls into strongly-typed layer which is composed of Java classes and interfaces annotated with Web service annotations. At this level, all of the messaging features are effectively hidden from your application. Now, these classes are conceptually implemented on top of messaging layer which is a lower level API. In your applications, you are free to access this lower API, but you will find that working at the strongly-typed layer is much simpler. The upper layer using annotations extensively is easy to use. With annotation-based, tools can do a very good job in helping you customize the mapping form a WSDL and XML schema based contact to java classes. So, they can generate annotations in particular places, easily analyze what annotations are already in the class, and figure out how that relates to the contact.

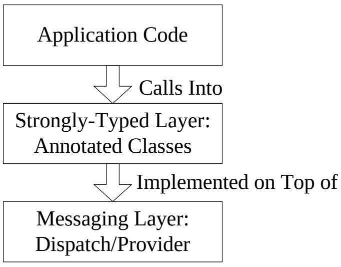

Application Code
Calls Into
Strongly-Typed Layer: Annotated Classes
Implemented on Top of
Messaging Layer: Dispatch/Provider

Fig.1. Layered Architecture

JAX-WS supports both SOAP+WSDL style Web Services and RESTful Web Services. REST (Representational State Transfer) basically means that each unique URL is a representation of some object. You can get the contents of that object using an HTTP GET. To delete it, you then might use a POST, PUT, or DELETE to modify the object. This paper mainly discusses the former one.

3. JAXB 2.0

The Java™ Architecture for XML Binding (JAXB) provides a fast and convenient way to bind between XML schemas and Java representations, making it easy for Java developers to incorporate XML data and processing functions in Java applications. As part of this process, JAXB provides methods for unmarshalling XML instance documents into Java content trees, and then marshalling Java content trees back into XML instance documents. JAXB also provides a way generate XML schema from Java objects.

Fig.2 shows the components that make up a JAXB implementation[4,5].

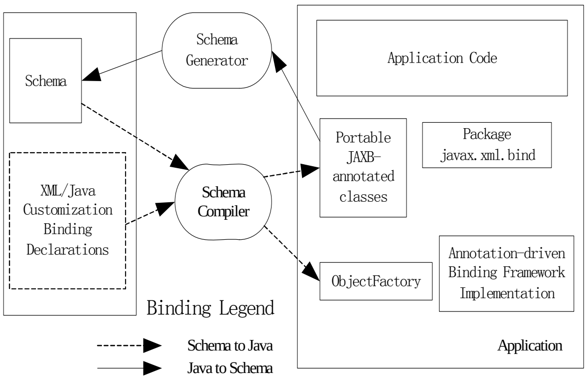


Fig.2. JAXB Architectural Overview

A JAXB implementation consists of the following architectural components:

- schema compiler: binds a source schema to a set of schema derived program elements. The binding is described by an XML-based binding language.
- schema generator: maps a set of existing program elements to a derived schema. The mapping is described by program annotations.
- binding runtime framework: provides unmarshalling (reading) and marshalling (writing) operations for accessing, manipulating and validating XML content using either schema-derived or existing program elements.

JAXB also provides the foundation for the data binding interoperability as part of Project Tango, whose goal is to provide interoperability between Sun's Java product and the other Window operating environment using WCF(Windows Communication Foundation aka "Indigo ").

4. ANNOTATION-DRIVEN PROGRAMMING ARCHITECTURE

The JAXB annotations provide the metadata that describe the Java-to-XML mapping. And all the JAXB annotations have a retention policy of run time, which means that they are captured in the class file. Thus, this enables the schema to be generated from the class files, either at development time or at run time. Fig.3 shows the architecture[6].

From the figure, we can see that when you begin with a schema, you can use the schema compiler to generate the portable annotated source code, which you can then compile using Javac into Java classes. The annotations in the java classes enable the instances of those classes to be marshaled and unmarshaled by the JAXB runtime. The JAXB runtime looks only at the annotations in the class files; it doesn't require the original schema to be present. And, likewise, the schema can be generated from the class files, because the annotations have run time policy and are in the class files.

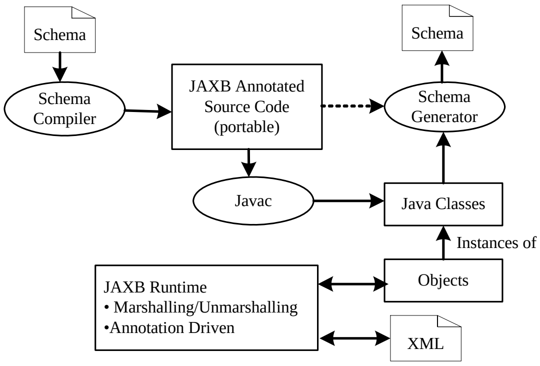


Fig.3. Annotation-Driven programming Architecture

In short, the metadata is actually collocated with a code rather than the external to the code, which makes it easier to deploy and move the code classed around. There's a simplified runtime because there is no special purpose marshallers or unmarshallers. Thus, the runtime is smaller .It enables you to write portable JAXB 2.0 applications on both the Java SE and the Java EE platform. And, the classes are almost like POJOs .It means that there is no artificial constraint on what the objects are, they are really natural classes with the constructor, you can do anything you want there, you don't have to expand a particular base class or implement a particular interface. It's a very freestyle program and that has positive implications for the test of your application. It's much easier now to instantiate one of these objects even outside the container and you need to test it once the classes are generated, you can add code, constrains and work with them just like any other developer-defined classes. And, most importantly, since the annotations actually just determine the mapping from Java to XML, it can also support additional languages.

The annotation-based model with the server side

The first step is writing a POJO and normal class, which implements the service and basically all its public methods will become Web service operations. Then step two, you're going to add the web service annotation to it, now if you need to access to some of the lower lever functionality, you can ask the container to inject WebServiceContext object. Then you deploy the application and as the application is deployed, a WSDL is generated and it will have a URL which ends with this "?"WSDL tag and then you can simply point your client to the WSDL.

Example: Servlet-Based Endpoint

```
@WebService
public class Calculator{
        Public int add (int i , int j ){
                return i+j;
        }
}
```

Here we are operating the Servlet container , it's one method one operation WebService and also it's very readable. Now, all the values for things like the service naming WSDL, the port type or the message name, everything has meaningful defaults which are derived from things like the name of a class, or the name of the method. In general, you can get pretty good contact even with the default use of annotations.

Example: Enterprise JavaBeans 3.0-Based Endpoint

```
@WebService
@Stateless
```

```
public class Calculator {
        @Resource
        WebServiceContext context ;
        public int add (int i ;int j ){
                return i+j ;
        }
}
```

To the EJB container, the only difference is that we use the EJB annotations to specify that component is a Stateless Session Bean. Anything else is the same. In this case, we can ask for a WebServiceContext to be injected by using the ubiquitous @Resource annotation. This is the same resource annotation that you could use to ask for injection of a database connection or message queue or things like that, and it also works for system level object to speak like the WebServiceContext and there is no particular limitation on your EJB as that comes from NetBeans and Web service. You are free to use any of the EJB functionality like transaction and security.

5. DATA BINDING

On the data binding side, we can fully integrate JAXB with JAX-WS. Then we can get 100 percent XML Schema support and attachment support for efficient transmission of binary data as attachments using MTOP/XOP. If you want to use in the past JAX-RPC and JAXB together, you will find that they have different rules for mapping data types to and from XML schema. But now there is only one mapping, one set of annotations. JAXB 2.0 is completely annotation-based, so we can completely follow the same principle, accept the defaults or go in and customize as much as we want every detail of the mapping.

How to write client on SE? The first thing to do is to point a tool like WSimport to the WSDL. This is going to generate the whole bunch of classes and interface. The generated classes are fully portable, so they contain annotation, but they don't depend on any particular implementation, so we can move them across from one tool or implementation to another. Now, one of the generated classes is the service class and this is sort of the entry point to a Web service. Then what you need to do is just simply call "new ". And so after you near the class, you can actually obtain a proxy by calling the "getPort"method and finally you can invoke any methods you want on that proxy.

```
CalculatorService svc =new CalculatorService( );
Calculator proxy = svc.getCalculatorPort ( );
int answer = proxy.add (3,4);
```

On the Java EE, the process is slightly different. You import the WSDL on step one, it's still going to generate classes and interfaces, there is no difference in what generated. The only difference is how you get hold of a proxy. In Java EE, what you need to do is to inject the WebServiceReference type to the proxy type in your component, which could be EJB, a servlet, a JSF managed beans or anything else, and then you can start invoking operations on it.

```
@Stateless
public class MyBean{
        @WebServiceRef (CalculatorService.class)
        Calculator proxy;
        Public int mymethod( ){
                Return proxy.add(3,4);
        }
}
```

From above, we can see that there are no factories and JINI. Two of the most despised things in Java programming are gone from Web services

JAX-WS is also protocol agnostic. The default is usually SOAP /HTTP, we have to mandate a default, so all implementations will do the right thing. But, it's possible when implementing your endpoint to specify that you want a different binding. This is done with the @BindingType annotation and this takes us argument a string which is an identifier, which specifies what binding you want instead of a standard one and this specification defines a number of bindings for SOAP/HTTP and XML/HTTP which we call also the REST binding .Now, bindings are not limited, so implementations can actually add new binding identifiers and new bindings to the runtime .For instance, you can imagine having a SOAP/JMS implementation or you could do a REST like binding .

6. CONCLUSIONS

Web services are Web based applications that use open, XML-based standards and transport protocols to exchange data with clients. Web services can be developed using Java Technology APIs and tools provided by an integrated Web Services Stack. Although Web service is not a Java-specific technology, Java can make it much easier to do.

REFERENCES

[1] Rajiv Mordani, Kohsuke Kawaguchi, Doug Kohlert, *Implementing the New Web Services APIs in the Java™ Platform*, 2005 JavaOne Conference

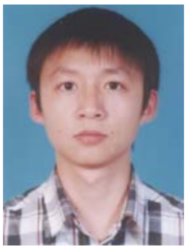
[2] Sekhar Vajjhala, *Java™ Architecture for XML Binding (JAXB) 2.0.* 2005 JavaOne Conference

[3] Roberto Chinnici, Rajiv Mordani, Doug Kohlert. *The Java™ API for XML Web Services (JAX-WS)* 2.0. 2006 JavaOne Conference.

[4] Eric Jendrock, Jennifer Ball, et al. *The Java™ EE 5 Tutorial*, Third Edition, Addison–Wesley,2007

[5] Web Services (JAX-WS) in Java EE 5. http://www.netbeans.org/kb/55/websvc-jax-ws.html

[6] Kohsuke Kawaguchi, Sekhar Vajjhala, Joe Fialli. "The Java™ Architecture for XML Binding (JAXB) 2.1". *Sun Microsystems*, Inc. 2006



Liang Huang, male, born in 1983. He is a master degree candidate of School of Computer Science and Technology, Wuhan University of Technology. His research interests are in e-commence and network security.

Research on Self-Adaptation of Web Component Based on Interface Automata*

Yukui Fei ,Jijun Zhang ,Hongmei Zhu
College of Info. Sci. & Engi. , ShanDong Agri. Univ. TaiAn 271018, P.R. China
Email: fyk@sdau.edu.cn

ABSTRACT

Web component aims at providing support to service-oriented development. We will deal with discuss the self-adaptation of web component based on Interface Automata in detail. We will present a formal description approach for self-adaptation of web component and give three algorithms to infer self-adaptive process of web component.

Keywords: Web Component, Interface Automata, Self-Adaptation, Perceiving Algorithms, self-tuning strategies

1. INTRODUCTION

In service-oriented development technology, web service is a mainframe. However, because of its infancy, it can still not be regarded as an ideal methodology. Current technologies (SOAP, WSDL and UDDI etc.) lack key features to support location, combination and the usage of the functionalities provided by published Web service automatically. In essence, they are origin from remote access and lack flexible. Although enhancing the semantic meaning of service, semantic web[1] can not adjust functions of service in dynamic. Some other methods, for example, Web Service Modeling Framework (WSMF) [2], DAML-S [3], and BPEL4WS [4] have obtained a lots of results in developing service-oriented software systems. Nevertheless, the common problem among them is regarding service as a passive entity rather than a free-will and goal-direction entity. When the environment changes, the service can not makes a pertinence adjustment so as to adapt the having changed environment.

In order to resolve this problem, we imported in the concept of web component in the previous work [5]. As a service-oriented intelligent software unit with explicit goal-directed and contractually specified interfaces based on Internet, web component aims at developing service-oriented software application systems. Comparing with traditional software Component technology, web Component is dynamic and intelligent-adjustable. In [5] we discussed the properties and structure of web Component. In this paper we will deal with the self-adaptation mechanism of web component in detail .By means of interface automata theory [6], we will give a formal description approach to discuss how web component adapt to the environment.

The structure of this paper is as follows. In the following section, there is a consensus on the Adaptability of Web Component. A formal description approach for self-adaptation of web component is discussed in section 3. We conclude with a summary in the last section.

2. SELF-ADAPTIVE MECHANISM OF WEB COMPONENT

Self-adaptation of web component refers to the ability of what web component perceives to environment and how web component adjust itself. Web component is usually made by the third-party. A Web component rarely fits directly into a new reuse context. For a Web component developer it is hard to foresee all possible reuse contexts. Hence, it is also hard for a developer to provide Web component with reasonable configuration options to fit into future reuse contexts. While making use of Web component, the Web component user has to make further adaptation operations on them again. Only so that can the Web component be applied in application system. We will take web component composition as an example to illustrate the self-adaptation of web component.

Web component composition is carried out under composition framework. Composition framework contains a series of rules to realize application. In every step, the composition framework states the functionality a Web component has to fulfill. Then she finds several candidates Web component in a repository, which deliver at least the required functionality. Some cases may occur:

1) If the function the candidate Web component really need from environment is satisfied then it is integrated into framework.
2) If the function the candidate component really need from environment is not satisfied completely then tuning the function what the Web component can offer and select other Web component which can provide the remain function at the given context. Composing these Web components at parallel way to implement the functionality .An other way of implementing the functionality is composing these Web component in sequence.
3) If the function provide by the candidate Web component is redundant then restricting it's provide function and tuning the function the Web component requires from the environment.

That implies two adaptive cases. One is adjusting Web component's provide functions in term of the particular environment, other is adjusting Web component's require functions according to its restricted provide functions. The common point of those cases is tuning Web components in order to adapt the particular environment. How to perceive the change of environment and how to make a corresponding adjustment is premier. In the next section we'll deal with it in detail.

3. FORMAL SPECIFICATION TO SELF-ADAPTATION OF WEB COMPONENT

In his section, we'll give a formal description for the self-adaptive mechanism of Web component. It relates to the represent of Web component, an appreciable method to environment and self-tuning strategies.

3.1 Modeling Web Component

A Web component consists of (1) an external contract made of

* Supported by the SDAU Youth Innovation Foundation under Grant No.23435.

typed input and output ports, and (2) an internal implementation consisting of classes, methods, and data fields. A client can only see the external contract and the internal implementation is completely encapsulated. A component provides services via its output ports, and specifies the services it requires via its input ports.

We use a framework that is similar interface automaton (IA)[6] to model Web component. Abstractly, a Web component automaton (WCA) consists of a set of actions, a set of state, and a set of transitions. The set of actions are classified as either input actions in (A) (corresponding to messages arriving at input ports), output actions out (A) (corresponding to the requirements at output ports), and internal actions int (A) (corresponding to internal calls).

DEFINITION 1. An Web component automaton $WCA=<V,V_0,\ A^I,A^O,A^H,\ \Gamma$ >consists of the following elements:

- V is a set of states.
- $V_0 \subseteq V$ is a set of initial states. We require that V_0 contain at most one state. If $V_0=\varnothing$, then WCA is called empty.
- A^I,A^O,A^H are mutually disjoint sets of input, output, and internal actions. We denote by $A= A^I,A^O,A^H$ the set of all actions.
- $\Gamma \subseteq V \times A \times V$ is a set of all steps
  If $a \in A^I$(resp. $a \in A^O$, $a \in A^H$), then (v,a,v') is called an input (resp. output, internal)step. We denote by Γ^I (resp. Γ^O, Γ^H) the set of input (resp. output, internal) steps. The WCA P is closed if it has only internal actions, that is $A^I=A^O=\varnothing$; otherwise, we say that P is open. An action $a \in A$ is enable at a state $v \in V$ if there is a step $(v,a,v') \in \Gamma$ for some $v' \in V$. We indicate by $A^I(v),A^O(v),A^H(v)$ the subsets of input, output, and internal actions that are enable at the state v, and we let $A(v)=A^I(v) \cup A^O(v) \cup A^H(v)$.

DEFINITION 2. An execution fragment of an WCA P is a finite alternating sequence of states and actions $v_0,a_0,v_1,a_1,\ldots,v_n$ such that $(v_i,a_i,v_{i+1}) \in \Gamma$ for all $o \le i \le n$.

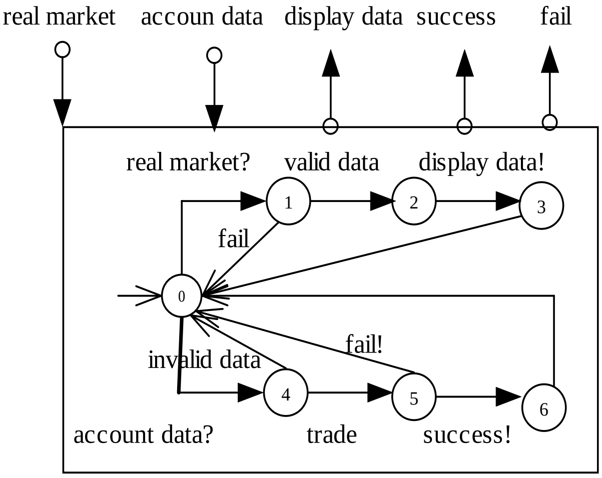


Fig.1. Web component stock trade(complete)

Given two states $v,u \in V$, we say that u is reachable from v if there is an execution fragment whose first state is v, and whose last state is in u. The state u is reachable in P if there exist an initial state $v \in V_0$ such that u is reachable from v.

In the definition of Web component Automaton, it is not required that all states are reachable. However, one is generally not interested in unreachable states, and they can be removed in linear time.

The following example, which is a simplified version of a real-life problem in the area of on-line stock trading(fig. 1).Web component stock-trade has two input actions (real market, account data), three output actions(display data, success and fail) and four internal actions(valid data , fail, invalid data and trade).

3.2 Perceiving Algorithms

In the process of composition, there exists an interaction relation between candidate Web components and the particular environment. The environment needs special functions of Web components and the Web components requires the environment to provide some conditions in order to fulfill its functions. However, that is not always realizable, and collision is inevitable .Web component should perceive the collision and make a corresponding action. When the environment could not provide the necessary conditions, the Web component should adapt itself to the environment. There maybe result in reducing functions what the Web component can offer or out of work. When the environment needs only partial functions provided by the Web component, the Web component should adjust its pre-conditions so as to adapt to the particular environment. The following two algorithms discuss how the Web component perceives the particular environment.

Algorithm 1 deal with the difference between the requirement of the Web component to the environment and the conditions the environment can provide. By means of trace search, the Web component decide weather it can run or what functions it can fulfill under the pre-conditions the environment could offer. If the Web component can run, algorithm 1 makes a record to the trace of the Web component pass, and ascertain the functions what the Web component could fulfill.

Algorithm1: Searching realizable functions

```
Input: WCA, $ACT_{IN} \subset A^I$
Output: $ACT_{OUT}$ // realizable functions of Web component
$Q \subseteq V$ // set of reachable states
    a : A; v,v' : V; st: STACK;
Q:={}
$ACT_{OUT}=\varnothing$
$v=V_0$
$Q=Q \cup \{v\}$
Push (st, v)
While not empty(st) do
{
        v=pop(st)
        if $(v,a,v') \in \Gamma$ && $a \in A^H$&&$v' \notin Q$    //internal step
        {
                $Q=Q \cup \{v'\}$
                Push(st,v')
        }
        if $(v,a,v') \in \Gamma$ && $a \in ACT_{IN}$&&$v' \notin Q$ //specify input
                                        step provided by environment
            {
                $Q=Q \cup \{v'\}$
                Push(st,v')
            }
        if $(v,a,v') \in \Gamma$ && $a \in A^O$&&$v' \notin Q$ //realizable output
                                                        step
            {
                $Q=Q \cup \{v'\}$
            Push(st,v')
            $ACT_{OUT}= ACT_{OUT} \cup \{a\}$
        }
    }
```

As an example (fig.1), when the environment only provides one requirement of the Web component stock-trade (for example real market), algorithm 1 will perceives the change of the context and work out reliable functions provided by stock-trade(display data).

The work of algorithm 2 is to make certain the pre-conditions what the Web component require to the environment. Using backward search, algorithm 2 work out the lowest environment requirement of Web component to fulfill the particular functions.

```
Searching state(aout: $ACT_{OUT}$)    //return correspond state of aout
    a : A; v,v' : V; s: STACK;
    v=$V_0$
    Push (s, v)
    While not empty(st) do
    {
        v=pop(s)
        if (v,a,v') $\in \Gamma$ && a==aout //correspond state of Web
                                              component's functions
            return(v')
        else if (v,a,v') $\in \Gamma$
            Push(s,v')
    }
```

Algorithm2:Searchingthe lowest environment requirement

```
    Input : WCA, $ACT_{OUT} \sqsubset A^O$
    Output: $ACT_{IN}$ // the lowest environment requirement
    Q:V; a : A; v,v' : V; st: STACK;
    Q:={}
    $ACT_{IN}$={}
    For i=1 to # $ACT_{OUT}$
    {
        v= $ACT_{OUT}$[i]
        v= Searching state(v)
        push(st,v)
    While not empty(st) do
    {
        v=pop(st)
        if (v',a,v) $\in \Gamma$ && a $\in A^H$&&v' $\notin$ Q    //internal step
            {
                Q=Q $\cup$ {v'}
                Push(st,v')
            }
        if (v',a,v) $\in \Gamma$ && a $\in ACT_{IN}$&&v' $\notin$ Q //specify input
                                         step required by Web component
            {
                Q=Q $\cup$ {v'}
                Push(st,v')
                $ACT_{IN}= ACT_{IN} \cup \{a\}$
            }
        if (v',a,v) $\in \Gamma$ && a $\in ACT_{OUT}$&&v' $\notin$ Q //valid
                                                       output step
            {
                Q=Q $\cup$ {v'}
                Push(st,v')
            }
        }
    }
```

For example. When the environment only requires web component stock-trade to provide function of display data, algorithm 2 re-computes the requirement of stock-trade and gets an minimal requirement (real market).

3.3 Tuning Policies

When have perceived the particular context (environment), the next step Web component has to do is an adjustment. What the Web component to do is generate a Web component satisfying requirement of the environment. That relates to the reduction of Web component's functions. The following algorithm realizes the process of new Web component generation.

According to the given pre-conditions (or worked out by algorithm 2), algorithm 3 deals with pruning operations and produces a new Web component to meet the particular context.

Algorithm 3: Generating a new WCA

```
Input : WCA, $ACT_{IN} \sqsubset A^I$, $ACT_{OUT} \sqsubset A^O$
Output: $WCA_{NEW}=(V_{NEW},V_{0NEW}, A^I_{NEW}, A^O_{NEW}, A^H_{NEW}, \Gamma_{NEW}>$
    a : A; v,v' : V; st: STACK;
    $V_{0NEW}=V_0$
    $A^I_{NEW}=ACT_{IN}$
    $A^O_{NEW}=ACT_{OUT}$
    $V_{NEW}$={}
    $A^H_{NEW}$={}
    $\Gamma_{NEW}$={}
    v=$V_0$
    $V_{NEW}=V_{NEW} \cup \{v\}$
    Push (st, v)
    While not empty(st) do
    {
        v=pop(st)
        if (v,a,v') $\in \Gamma$ && a $\in A^H$&&v' $\notin V_{NEW}$//internal step
            {
                $V_{NEW}=V_{NEW} \cup \{v'\}$
                Push(st,v')
                $A^H_{NEW}= A^H_{NEW} \cup \{a\}$
            }
        if (v,a,v') $\in \Gamma$ && a $\in ACT_{IN}$&&v' $\notin$ Q //specify input
                                                        step
            {
                $V_{NEW}=V_{NEW} \cup \{v'\}$
                Push(st,v')
                $A^I_{NEW}= A^I_{NEW} \cup \{a\}$
            }
        if (v,a,v') $\in \Gamma$ && a $\in ACT_{OUT}$&&v' $\notin$ Q //specify
                                                        output step
            {
                $V_{NEW}=V_{NEW} \cup \{v'\}$
                Push(st,v')
                $A^O_{NEW}= A^O_{NEW} \cup \{a\}$
            }
    }
```

For example, when the requirement of web component stock-trade is only account data, the result of adjustment is as Fig.2.

4. CONCLUSIONS

In this paper, we have presented a formal approach to deal with the self-adaptation of Web component. Following the function may tuning in term of environment principal, the approach has discussed two kinds of self-tuning strategies in detail and has given correspond algorithms? It is useful to build software system out of components under Internet environment. The future work we have to do is to complete the self-adaptation mechanism of Web component, give a precise semantic description and to construct a practicable framework.

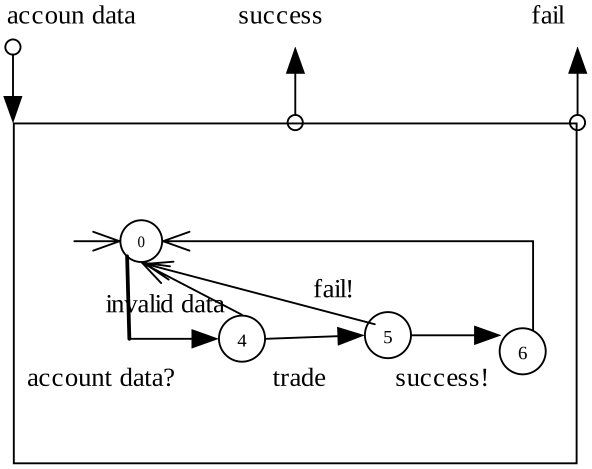


Fig.2. Web component stock trade (self-tunned)

REFERENCES

[1] Berners-Lee, T., Hendler J., and Lassila, O. “The Semantic Web”, *Scientific American*, 2001:284(5), 34-43.

[2] D. Fensel, C. Bussler: “The Web Service Modeling Framework WSMF”, *Electronic Commerce Research and Applications*, 2002:1(2).

[3] The DAML services coalition: DAML-S: Semantic Markup for Web Services (version 0.9), available at http://www.daml.org/services/daml-s/0.9/daml-s.pdf, 2003.

[4] Business Process Execution Language for Web Services, version 1.1, available at ftp://www6.software.ibm.com/software/developer/library/ws -bpel11.pdf, 2003.

[5] Fei Y.K., Wang Z.J. “A Concept Model of Web Components”. In Proc. of *IEEE International Conference on Services Computing*, 2004, 159-164.

[6] Luca de Alfaro and Thomas A. Henzinger. "Interface Automata". *Proceedings of the Ninth Annual Symposium on Foundations of Software Engineering* (FSE), 2001, 109-120.

Research on Method of Learning Web Information Extraction Rule Based on XPATH

Yan Hu, Yanyan Xuan
Dept. Computer Science & Technology, Wuhan University of Technology
Wuhan, 430070, China
Email: huyan168@sina.com, yy.xuan@163.com

ABSTRACT

This paper identifies theme blocks through cleaning website on the basis of the research in HTML documents structure, designs and implements a theme information extraction (IE) method with web based on XPATH, studies the key point of this method-XPATH expression that expresses the IE path, and then constructs an XPATH automatic algorithm. Thereby, IE rules can be learned automatically and generated to implement Web IE.

Keywords: DOM, XPATH, XSLT, Web Information Extraction

1. INTRODUCTION

With the rapid development of Internet, Web has become the main source of information, and the number of website is increasing explosively. How to obtain necessary information effectively from web has become a serious problem. Now, most Web information has been published in the form of HTML. This unstructured or semi-structured Web carrier only gives data but lacks data description. So it is difficult to analyze using procedural means, and also it is unable to provide a structured query language for users to inquire efficiently. As a result, IE technology is becoming more and more necessary. The essence of IE technology is to collect factual information from natural language text and describe it in the form of structure for the application of information inquiries, deep excavation of text, automatic answers to questions, and so on. IE is gradually becoming an important issue in such research areas as information retrieval, data mining, knowledge management, competitive intelligence, etc. The procedure with extracting information from websites is called wrapper, whose main task is constructing extraction rules. As a result, how to prepare flexible and effective rules becomes a focal point in the area of IE.

2. IE PRINCIPLE AND PROCESS DESIGN

Information Extraction (IE) [1] is a process of extracting relevant data from the page pools. Web IE is information collection based on the Web pages, formal description expressed as follows: For a given group of a Web pages S, define a mapping W, W maps the object of S to a more structured and clearer semantic data structure D (e.g. Relational Databases). And the mapping of W has the same functions as S' which is similar to the Web page set S in semantics and structure. So, the key to IE is the definition of mapping W, that's extraction rules.
Overall outlook of Web IE research:

(1) Sample Classification Stage: HTML samples are downloaded from Internet in the field of teaching, and stored into domain knowledge base after classification.
(2) Rules Training Stage: Make the websites from knowledge base as training samples. Use semi-automated method to generate information extraction rules and different rules are stored into the extraction rules training lib.
(3) Rules Optimization Stage: Use inductive learning method for rules optimization with the training samples from Rules Training Stage.
(4) IE Stage: Users extract information from HTML pages with the rules learned when requesting information from websites. If there are no appropriate rules in rules lib, study the structure of such pages produce new rules and store them into extraction rules lib.

In this paper, the main task is generating the extraction rules based on XPATH [2] to complete Rules Training Stage. The work flow is shown in Fig.1.

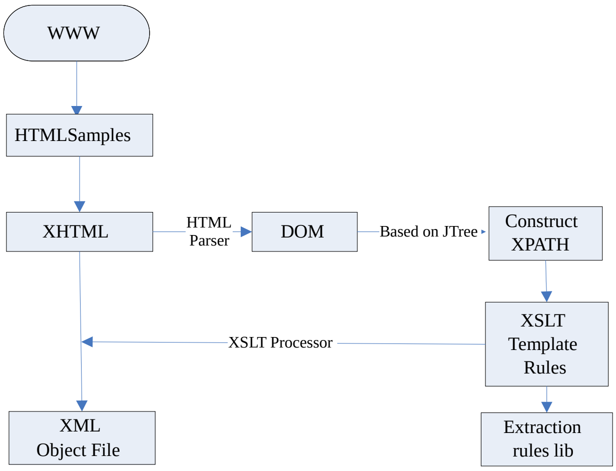


Fig.1. work flow

3. ACHIEVING IE BASED ON XPATH

3.1 DOM [3], XSLT [4] and XPATH

Document Object Model (DOM) is an Application Programming Interface (API) for HTML and XML documents use, which defines the logic structure and the standard method for accessing and manipulating files in various parts of document. Different interface indicates different elements, attributes, analyzed character data, note and processing instructions. All these are sub-interfaces of common Node-Interface. Node-Interface provides the basic processing methods of tree navigation.

XPATH is the fourth-generation statement language, used to posit XML document nodes. XPATH trails designate which points are needed for the document without pointing out algorithm for finding the nodes. As long as a XPATH statement is transferred to method, XPATH engine will be responsible for determining how to find all the nodes satisfying the formula.

The basic syntax of XPATH is constructed by expression, which is similar to finding documents in the file system. If the path begins with the character "/", it shows that this path is an absolute path and it is consistent with the definition of the UNIX System with document on the path. This paper uses XPATH expressions to complete the mapping of data into XML. XPATH expression points the path from the root <html> element to bolt. Assuming an XML sequence of TAG as follows:

tbody/table/tr[2]/td[3]

It means the table No. 1, line 2, para 3 of a HTML documents, and this XPATH is the invoke point of required data.

XSLT (eXtensible Stylesheet Language Transformation) is a language drafted and formulated by the W3C to transform the structure of XML. At the same time, XSLT is a branch of the XSL and is being widely used as a standard and for different data format conversion, such as XML-XML, XML-HTML etc. In the conversion process, XSLT processor reads XML document and XSLT stylesheet, converts document according to directive of XSLT stylesheet, and then outputs the result document after conversion. Because XML is a complete tree structure document, some information points need to be processed in the use of XSLT for transforming XML documents. XPATH is a language used to locate every part of XML, and also it can help XSLT find information points quickly and effectively.

3.2 Cleaning HTML Pages

called "noise" contents. In visual terms, a web page can be divided into a number of regions. A region is known as a content block. Some content blocks contain themes, while others contain noise content. Usually, the content is closely related in a content block, and this means we are able to make our choice of the content from the website according to block content. Based on this analysis, website purification process is to retain the content blocks which contain thematic content in websites and remove the content blocks which contain noise content. Therefore, the process of cleaning website is: First, identify the main website content block, and then identify the content blocks related with subjects on the basis of subject content in the remaining blocks. The rest are noise content. The main cleaning method of HTML documents is:

Using JTidy [5] filter the error of HTML document, constructing HTML tag tree, and then determining the theme content block, the recognition of which is based on the following heuristic rules. A website with a theme is described in text, without including lots of hyperlink intermediately. Instead the text is often accompanied by the Hyperlink. Therefore, in the website with theme, if a content block contains thematic content, then the content of this block is a part of the subject matter of this website. According to this rule, we can get the theme content of the website through depth-first traversing tag tree and recording the path of content block with theme type. Finally, we can cut tag tree according to the theme content block, and delete non-thematic content nodes.

The cleaning process of the entire page is shown in Fig.2:

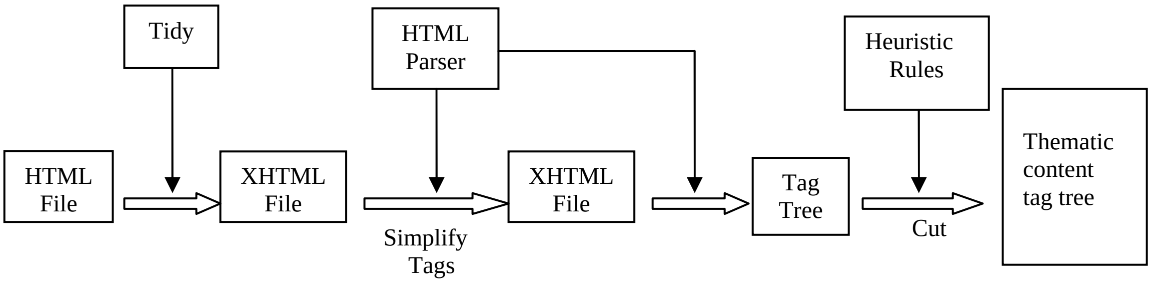


Fig.2. Website Clean Process

3.3 IE Rules

XSLT will treat an XML document as a node tree which is called "source-tree" and it also treat the transform result as a node tree called "result-tree". Source-tree and result-tree are separated, and the structure of the both can be the same or different. When a "result-tree" is being created, elements from which can be filtered and re-scheduled, and also arbitrary structure can be added.

The basic constraint on node tree is the same as XPATH. In XSLT, the transform is called stylesheet, which defines a set of rules for transforming a source tree into a result tree, named Webpage usually contains two aspects. One part reflects the main messages of website. For example, the news part of a news website. The other part is some navigation information, advertising information, copyright information, and several other topics unrelated to the main topic information. That is

template-rule. Stylesheet is defined through the element <xsl:stylesheet> of XSLT document and the content of that is usually more than one template-rule. A template-rule has two parts: Patterns defined with XPATH syntax used to match the element of source-tree and Templates used to instruct how to create part of the result-tree. Template-rule is defined with <xsl:template> element, whose match attribute is a model, which is used to match the element of source-tree, and the content of that is a template. Because the data of an XML document can be extracted and represented as a new XML with XSLT, XSLT file just can be seen as extraction rule in terms of IE. The following is a simple XSLT template generated by using the XPATH of information point after cleaning a HTML sample page and positioning information point to build XPATH expression (in bold).

```
<?xml version="1.0" encoding="gb2312"?>
```

```
<xsl:stylesheet version="2.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output encoding="gb2312" indent="yes"/>
<xsl:template
   match="/html[1]/body[1]/form[1]/table[2]/tr[2]">
<INFORMATION>
<xsl:value-of select="td[5]/font[1]/text()"/>
</INFORMATION>
</xsl:template>
```

JAXP standard method is used to establish a stylesheet template object, generate converter, and transform source XHTML object into the final target XML document object.

4. THE LEARNING METHOD OF IE RULES

From the above XSLT template rules, we can see that building the path of information points, or XPATH expression, is the key to extraction rules. The method provided in this paper completes the theme IE from websites based on the information provided by Web pages themselves, and then according to the position of theme information nodes, IE rules can be created by accessing the XPATH expression automatically.

DOM mode of Xerces-J [6] is used to parse the XHTML document coming from the website which has been cleaned with tidy. First, build a DOM parser and parse the XHTML documents into Document Object. Then find all the Document Object Nodes through traversing the DOM tree. Building XPATH needs TreeNode, therefore, XPATH is created through traversing the source DOM tree to generate the TreeNode based on JTree. Algorithm 1 describes the process of generating JTree from DOM tree. Algorithm 2 describes the process of constructing XPATH automatically.

Algorithm 1: Traverse DOM tree to generate JTree

```
ConvertToJTree (TreeNode, Node)
{
   If (Node<>null) THEN
   NodeType= Node
   If NodeType= DOCUMENT Node
   ConvertToJTree(TreeNode,
   Node.getDocumentElement ())
   End if
   If NodeType= ELEMENT
   NodeList =Node's children nodes list
   If (NodeList<>null) then
   Len = NodeList.getLength () //length of NodeList
   For (I = 0; I < Len; I++)
   // Read all the child nodes in Node List circularly
   Child = NodeList.item (I) //judge the Child's type
   If Child.NodeType = ELEMENT
   If Child.NodeType = ELEMENT
   //Put the node as a parent node (Parent) to JTree
   ConvertToJTree (Parent, Child)
   //structure all the JTree's nodes Recursively
   End If
   If Child.NodeType = TEXT (Leaf node)
   //Add as leaf node
   End If
   End For
   End If
   End if
   End if
   }
```

All the JTree's leaf nodes constructed by the cleaned HTML document are the information points which should be extracted. XPATH is generated through traversing the JTree's TreeNode of information points and the XPATH expression designates the path from <html> (Root Element) to information points. The algorithm of building XPATH is as follows:

Algorithm 2: Generate XPATH Expression

```
BuildXPath(TreeNode)
{
   XPATH=null
   TreeNode=TO-Extract-Node
   ParentNode = TO-Extract-Node's Parent
   IF (ParentNode<>null) Then
   WHILE (TreeNode <>null)
   ParentNode = TreeNode.getParent ()
   XPATH = ParentNode's Tag [index] + XPATH
   // deal with all the path expression between root node and
      leaf nodes circularly
   ParentNode = ParentNode.getParent ()
   TreeNode = TreeNode.getParent ()
   END WHILE
   END IF
}
```

The resulting XPATH expression is what we need from Root to TreeNode. Append the XPATH generated automatically into XSLT template and map the information of theme content block to XML document, then the IE of theme content from website is completed.

5. CONCLUSIONS

This paper presents a Web Information Extraction method based on XPATH, achieving cleaning HTML Pages, parsing XHTML and building XPATH expression. It's easy to compile IE rules using our XPATH generation algorithm. Meanwhile, this algorithm can be applied to parse any standard XML document for extracting data.

REFERENCES

[1] Laender H F, Ribeiro-Neto B A, A S da Silva, et al. "A Brief Survey of Web Data Extraction Tools" [J]. *SIGMOD Record*, 2002, 31(2): 84-93

[2] XML Path Language (XPath), W3C *Recommendatio*n, November 1999. http://www.w3.org/TR/xpath.html.

[3] Document Object Model, W3C *Recommendation* October, 1998. http://www.w3.org/DOM/.

[4] XSL Transformations (XSLT), W3C *Recommendation,* November 1999. http://www.w3.org/TR/xslt.html.

[5] http://sourceforge.net/projects/jtidy，August 2001.

[6] http://xerces.apache.org/xerces-j/index.html

Struts-Based ArcIMS Communication Mechanism and Its Application Development*

Kangshun Li [1], Song Xie [1], Huilan Luo [1], Yuanxiang Li [2]
[1] **School of Information Engineering, Jiangxi University of Science & Technology**
Ganzhou, jiangxi 341000, China
[2] **State Key Laboratory of Software Engineering, Wuhan University,**
Wuhan, Hubei 430072, China
Email: lks@public1.gz.jx.cn

ABSTRACT

ArcIMS is a frame through which Internet can issue GIS information. What's more, it is very easy to create a map service, develop a website, manage a web station and organize spatial data with ArcIMS. In this paper, the parts composing ArcIMS, as well as the ArcXML document communicated between various parts are discussed, and in this foundation the technical structures of ArcIMS re-development based on Struts construction has constructed.

Keywords: ArcIMS,ArcXML,WebGIS,Struts

1. INTRODUCTION

With the rapid development of network, Internet has become a new system platform through which the information of GIS can be issued. And it is an inevitable trend of GIS to use technology of Internet to issue spatial data on web for user to browse and utilize [1].

WebGIS has following characteristics. Firstly, it is based on Internet/Intranet standard. Secondly, it has a distributional service system structure. Thirdly, high effective use spatial data resources. Fourthly, the speed of data issue is very quick and the scope it covers is very large. Fifthly, it has a timely data renewal, a rich development kit, a friendly contact surface with a relatively low investment of system construction [2].

Aiming at managing and issuing data more effectively, the Open GIS alliance has formulated the unified data standard—The Geography Markup Language (GML). It is a coding standard which is based on transmit and storage of geography information of XML (including geometry and attributes of geography characteristic). And the GML1.0 edition is released by GIS alliance in April, 2000. Later, the GML2.0 is also released in February, 2001. GML, based on the text, is a simple code standard of geography characteristic. It uses the features of geography to describe the world. And it has the capacity to code geography entities with high complex [3]. In this paper, a WebGIS platform (ArcIMS) which is widely used by many persons is introduced. And ArcXML is a data standard that used in communication between different parts of ArcIMS. Meanwhile, it is the expansion of XML, and conforms to the standard of GML2.0.

In this paper, section one introduces the components of the ArcIMS, section two introduces the characteristics and types of ArcXML document, section three discusses the Communication process of ArcIMS, then Cited a example of re-development ArcIMS.

2. BRIEF INTRODUCTION OF COMPONENTS OF ARCIMS

ArcIMS is a frame which is proposed by ESRI Corporation and its main purpose is to issue GIS functions on Internet. Meanwhile, it is also a distributional system composed of client side module, ArcIMS Connector and the server end module. As is shown in Fig.1.

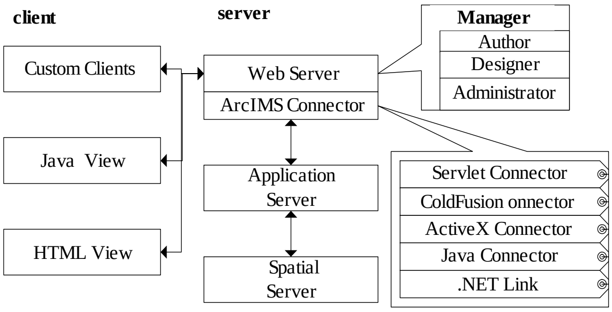


Fig.1. ArcIMS structure drawing

The client side module is composed of a variety of browsers such as IE, Netscape and special browser for java.

Web server and ArcIMS application server are connected together by ArcIMS Connector. Servlet Connector is the default connector of ArcIMS. Of course, there are many other connectors, including ColdFusion Connector, ActiveX Connector, Java Connector and .NET Link.

The server end module includes three parts:

(1). Application server: It in charge of the requests, transmits the requests to a suitable spatial server, and tracks the services running in the spatial server.

(2). Spatial server: It is the essential part of ArcIMS. It manages the services, processes requests from maps, factor data, geographic coding and data capture. The main task the spatial server performs is dynamic romances the maps, and then displays it on the HTML page; it also produces some dynamic information such as some failure logs of requests and responses.

(3). Manager: contains three independent modules (Author, Designer and administrator), these modules can register the graphic file, design homepage, issues the map service and manage space server.

3. ARCXML DOCUMENT

ArcXML (the Arc extensible markup Language, called AXL

* This work is supported by the National Natural Science Key Foundation of China with the Grant No.60133010, the Research Project of Science and Technology of Education Department of Jiangxi Province with the Grant No.Gan-Jiao-Ji-Zi [2007]205.

for short) is a developed markup language, and it is always used to descript content rather than the performances of data structure, all the ArcXML sentence is composed of elements and attributes which are organized together by the rank structure.

Under the rank structure, elements are divided into to two classes (the parent elements and the child elements). And the child elements embed in the parent elements. Most elements are consisted of attributes including name and value. These elements and attributes do not save the real data chart level. Instead, they restore geography information concerned such as service information, chart level information and so on.

The client side transmits to server a variety of ArcXML requests; and then the ArcIMS space server will analyze the ArcXML requests and return to the client the responses in the term of ArcXML. There are several types of requests and response in ArcXML. They are all shown in Table 1.

Table 1. Types of requests and responses

requests	responses
GET_IMAGE	IMAGE
GET_FEATURE	FEATURES
GET_SERVICE_INFO	SERVICEINFO
GET_LAYOUT	LAYOUT
GET_METADATA	METADATA
GET_GEOCODE	GEOCODE

4. COMMUNICATION PROCESS

Because there exist a various kinds of connecting ways in ArcIMS, we can choose different languages and modes to help us develop based-on ArcIMS WebGIS. In this paper, we just take the project which I am now developing for example:

The development environment is as bellows: The operating system is Windows 2000; Web server is jakarta-tomcat-5.5.9; WebGIS platform is ArcIMS 9.0; Java Connector; Data source are Orcale9i and ArcSDE; Data transmission way is Grid data; Development language are JSP and Java.

In the development process we introduced the technology of struts which is based on the MVC pattern, and look it as the center of web serves. The inter relationships among Struts, ArcIMS space server, graph configuration files, graph service, request and response are shown in Fig. 2.

All request coming from the client side are all received by the ActionServlet which reads ActionMapping information in configuration files struts-config.xml, and then calls correspondent Action. And then Action calls JavaBean to process the requests (example: JavaBean accepts the request of gain image, reads the GET_IMAGE.xml document, fills in minx, miny, maxx, maxy as well as the id value), and finally JavaBean sends the processed request to ArcIMS application server.

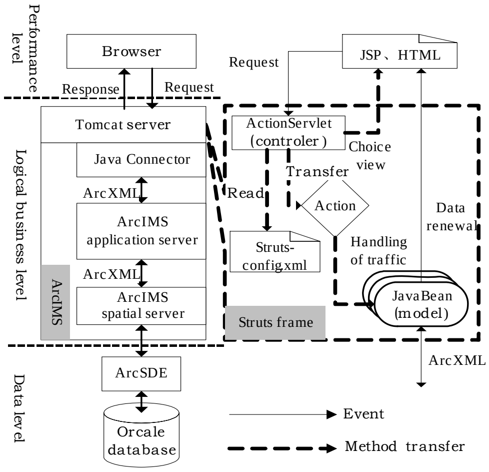


Fig.2. Development technology diagram

Here, we proposed an example to describe the process that how the ArcXML to ask for image from spatial server. The contest of the transmitted ArcXML are as bellows:

```
<?xml version="1.0" encoding="UTF-8"?> // edition and code
form of XML
<ARCXML version="1.1"> // the outset part of ArcXML
  <REQUEST> // this is the request indicated
    <GET_IMAGE> //this is the request type for obtains the
image
      <PROPERTIES> // the attribute of request image
        /*the image scope */
        <ENVELOPE minx="114.899" miny="025.827"
maxx="114.981" maxy="025.891" />
          /*the picture size which will display on the monitor */
         <IMAGESIZE width="800" height="567" />
          /*there are two chart levels, the chart level which id
number is 1 will display*/
         <LAYERLIST>
           <LAYERDEF id="2" visible="false" />
           <LAYERDEF id="1" visible="true" />
           </LAYERLIST>
          </REQUEST>
      </PROPERTIES>
  </GET_IMAGE>
</ARCXML>
```

ArcIMS application server accepts the requests and then transmits them to the ArcIMS space server. The space server analyzes the ArcXML requests, connects the database to obtain the data, and produces the corresponding images and response ArcXML which will be feed back to the application server. The image information included in response is as bellows:

```
<ARCXML version="1.1">
  <RESPONSE> //this is the response, corresponded to
<REQUEST>
      <IMAGE> // this is the response type for returns the image,
corresponded to <GET_IMAGE>
      /* the image scope, is consistent with ENVELOPE in
REQUEST */
      <ENVELOPE minx="114.894850088183" miny="25.827"
maxx="114.985149911817" maxy="25.891" />
      /* the depositing position of spatial server produces map */
      <OUTPUT
                    file="d:\ArcIMS\Output\gzmapservice_XIESO
```

```
                  NG3192363251.jpg"
                  url="http://xiesong/output/gzmapservice_XIES
                  ONG3192363251.jpg" />
   </IMAGE>
  </RESPONSE>
</ARCXML>
```

JavaBean accepts the response ArcXML, withdraws the image URL and renew the parameters of image object in the JSP page. And then ActionServlet informs the JSP page to refresh the view to display the requested image in the JSP page. The results are as shown in Fig.3.

5. CONCLUTIONS

With the development of Internet technology, comes an opened, standardized, and platform –crossed network time. And the commercial product and the development of applications of WebGIS will become increasingly vigorous. It will bring us a more convenient and very efficient life.

REFERENCES

[1] XIAO Xin-zhi,SU Fen-zhen,DU Yun-yan,ZHOU Jun-qi,"Performance Analysis and Optmization of WebGIS," *Platform[J].GEOMATICS & SPATIAL INFORMATION TECHONLOGY*,Aug,2005,Vol 28,No.4,pp.1-3.

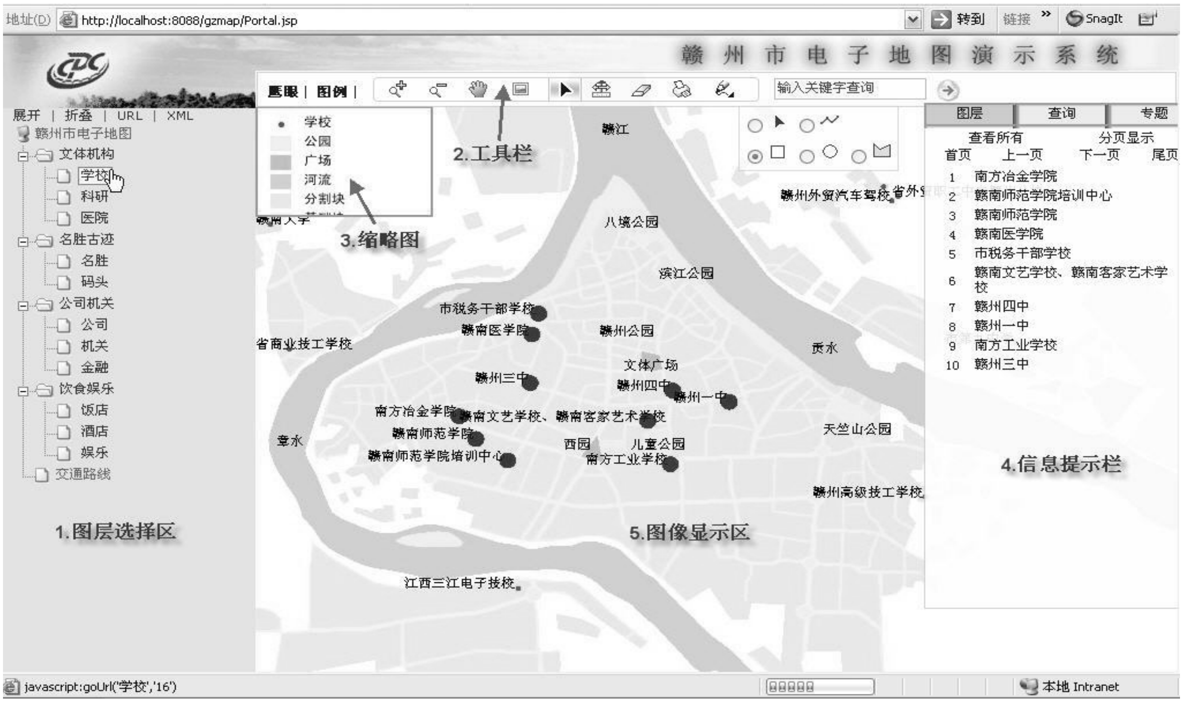
[2] Kang Zhiyu,Wang Mingsheng,"Research on Developments and Application of GIS," *JOURNAL OF SHIJIAZHUANG RAILWAY INSTITUTE[J]*,Mar,2005, Vol.18,No.1,pp.62-66

[3] Li Xu Zhuoqun, Ma Jian, Wang Xiaolin, Luo Yingwei,"Study on GML-based WebGIS [J],"*COMPUTER ENGINEERING*, July,2002(7), pp.23-26.

[4] *ArcGIS9:Installing ArcIMS 9 on Windows[Z]*,ESRI Corporation,2004.

[5] HUANG Kang,SHI Zhou,"The Principle Analysis and Application Development of ArcIMS[J]," *GEO-INFORMATION SCIENCE*. Sept,2005,Vol.7,No.3, pp.61-66.

Kangshun Li is currently a senior member of IEEE, senior member of China Computer Federation, member of High Performance Computation Committee of China Computer Federation. He is an associate professor in School of Information Engineering of Jiangxi University of Science and Technology. He received the B.S. degree in Computational Mathematics from Jiangxi University (now is Nanchang University), and Ph.D. degree in Computer Software and Theory from Wuhan University, in 1983 and 2006, respectively. From 1983 to 2002, he worked in Computer Center of Statistical Department of Ganzhou as director. His current research interests are Evolutionary Computation, Inverse Problems, Genetic Programming, Gene Expression Programming, Multi-Objective Optimization, Evolvable Software of Embedded System, Evolvable Hardware, Evolutionary Modeling, Intelligent Computation, Parallel Computation, and Neural Network. He has 23 papers published on international journals, international conference proceedings, and Chinese core journals, he also join many international activities. He is Advisor of graduate students, Reviewer for several Journals, e.g. the Journal of System Simulation, the proceedings of the First International Conference on Natural Computation(ICNC'05) and the Second International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'05).



1 南方冶金学院
2 赣南师范学院培训中心
3 赣南师范学院
4 赣南医学院
5 市税务干部学校
6 赣南文艺学校、赣南客家艺术学校
7 赣州四中
8 赣州一中
9 南方工业学校
10 赣州三中

Fig.3. Map of contact surface and request image

Research on Effective Web Information Retrieval Based on Semantic Web

Min Xiao [1], Qianxing Xiong [1], Chunhua Wang[2], Qiumei Pu [1]
[1]School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei 430063, China
[2]School of Computer Science and Technology, Huanghuai University
Zhumadian, Henan 463000, China
Email: xiaomin@whut.edu.cn

ABSTRACT

The use of ontology to overcome the limitations of keyword-based search has been put forward as one of the motivations of the Semantic Web since its emergence in the late 90's. In the paper, an improved information retrieval model based on ontology is proposed. Exploiting ontology, the information retrieval system can filter irrelevant information to gain high precision and find latent relevant information to retain high recall. The analysis result shows that the improved information retrieval model can effectively enhance retrieval performance.

Keywords: Information Retrieval, Semantic Web, Ontology, Resource Description Framework, Web Ontology Language

1. INTRODUCTION

As Internet information has been exploding, information retrieval has facilitated people with getting knowledge. Keyword-based search engines, such as Alta Vista, Yahoo, and Google, are the main tools of using today's Web. It is clear that the Web would not have been the huge success as it was, were not for search engines. However, there are some serious problems associated with their use [1]:

- High recall, low precision. Even if the main relevant pages are retrieved, they are of little use if far more mildly relevant or irrelevant documents were also retrieved. Too much can easily become as bad as too little.
- Low or no recall. Often it happens that there is not any answer to our request, or that important and relevant pages are not retrieved.
- Results are highly sensitive to vocabulary. Usually the search engines based on keywords do not get the appropriate results. In these cases the relevant documents use different terminology from the origin query keywords. This is unsatisfactory because semantically similar queries should return similar results.
- Results are single Web pages. If information is needed to spread over various document, several queries are necessary in order to collect the relevant documents, and then the partial information is extracted and put together manually.

One of the most important reasons for this unsatisfactory state of affairs is that existing Web resources are usually only human-understandable: the mark-up (HTML, Hypertext Markup Language) only provides rendering information for textual and graphical information intended for human consumption [2]. At present, the meaning of the Web is not machine-accessible. Of course, there are tools that can retrieve texts, split them into parts, check the spelling, and count their words. But when it comes to interpreting sentences and extracting useful information for users, the capabilities of current software are still very limited [3] [4]. For example, if we want to retrieve "experts familiar with XML" on the Web, Google, the most predominant search engine, will find 1,010,000 results, but none of them is relevant. The seemingly simple retrieval actually involves very complex logic concepts, semantics, and syntax relations.

It is believed that the Semantic Web as propagated from Tim Berners-Lee will bring some relief in this unsatisfying situation. The idea behind the "Semantic Web" is to make the information more "machine-process" and thus enable tools to assist people effectively [5]. The Semantic Web supplies a new technological approach to the bottleneck problems above. At the same time, it shows a new way to the research of the semantic information retrieval.

2. THE SEMANTIC WEB AND OWL

According to Lee's view, the Semantic Web, aims at machine-understandable Web resources, whose information can then be shared and processed by automated tools, such as search engines. It is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [1][6][7]. It is obvious that the information representation form is suitable for effective retrieval and process and can enhance recall and precision greatly.

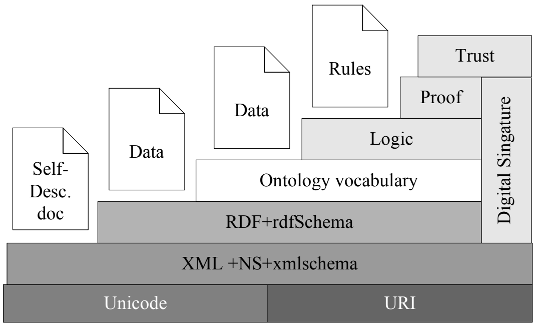


Fig.1. a layered architecture to the semantic web

The architecture of the Semantic Web looks like a layered pyramid (Fig1). It includes seven layers. Unicode and URI layers make sure that international characters sets can be used to identify the objects in Semantic Web. XML (eXtensible Markup Language) is a language that lets one write structured Web documents with a user-defined vocabulary. RDF is a basic data model, and it is particularly suitable for sending documents across the Web. RDF is a basic data model for writing simple statements about Web objects (resources). RDF Schema provides modeling primitives for organizing Web objects into hierarchies. OWL (Web Ontology Language) is on top of RDF, and it allows the representations of more complex relationships between Web objects. The Logic layer is used to enhance the ontology languages further and to allow the writing of application-specific declarative knowledge. The Proof layer involves the actual deductive process as well as the representation of proofs in Web languages and proof validation. Finally, the Trust layer will emerge through the use of digital signatures and other kinds of knowledge. Among the seven layers, XML, RDF and Ontology are three core layers, which are used to represent machine-understandable resources.

The XML standard can be used to declare simple data structure. However, since XML is only defined at the syntactic level, it does not provide any means of talking about the semantics (meaning) of the data. In a given XML documents, machines cannot unambiguously determine the correct meaning of XML tags. For example, there is no intended meaning associated with the nesting of tags; it is up to each application to interpret the nesting. This point is illustrated through an example below. Suppose we want to express the following fact:

Tim Berners-Lee is inventor of Semantic Web.

There are various ways of representing this sentence in XML.

Three possibilities are below:

```
<invention name=”Semantic Web”>
   <inventor> Tim Berners-Lee</inventor>
</invention>

<inventor name=”Tim Berners-Lee”>
   <invents>Semantic Web</invents>
</inventor>

<inventionsomthing>
   <inventor>Tim Berners-Lee</inventor>
   <invention>Semantic Web</invention>
</inventionsometing>
```

It is obvious that the first two formalizations include essentially an opposite nesting although they represent the same information. So there is no standard way to assign meaning to tag nesting.

Considering the deficiency of semantics in XML, RDF presents the semantics in a simple data model in which the resources are described in 3-triple statements. RDFS is a vocabulary description language for describing properties and classes of RDF resources, with a semantics for generalization hierarchies of such properties and classes.

Although RDF and RDFS are more expressive than XML, they are far from satisfactory: RDF is limited to binary ground predicates, and RDF Schema is limited to a subclass hierarchy and a property hierarchy, with domain and range definitions of these properties. Considering the drawbacks of RDF and RDFS, OWL is aimed to be the standardized and broadly accepted ontology language of the Semantic Web.

The goal of OWL is to enable the transformation of the currently human-oriented web into a Semantic Web. OWL provides some rules for describing further constraints and relationships among resources, including cardinality, domain and range constraints, as well as union, disjunction, inverse and transitivity rules. OWL will enable the development of intelligent agents and applications that can retrieve and manipulate information in today’s Web and the Semantic Web of tomorrow.

3. INFORMATION RETRIEVAL BASED ON SEMANTIC WEB

The key problem in achieving efficient and user-friendly retrieval is the development of a search mechanism to guarantee delivery of minimal irrelevant information (high precision) while insuring relevant information is not overlooked (high recall). The notion of ontology is introduced to information retrieval systems to solve the key problem above. The architecture of information retrieval system based on Semantic Web is composed of five modules as shown in Fig .2:

（1） Ontology design and maintenance
（2） User query interface
（3） Preprocessing
（4） Semantic retrieval
（5） Retrieval results optimization

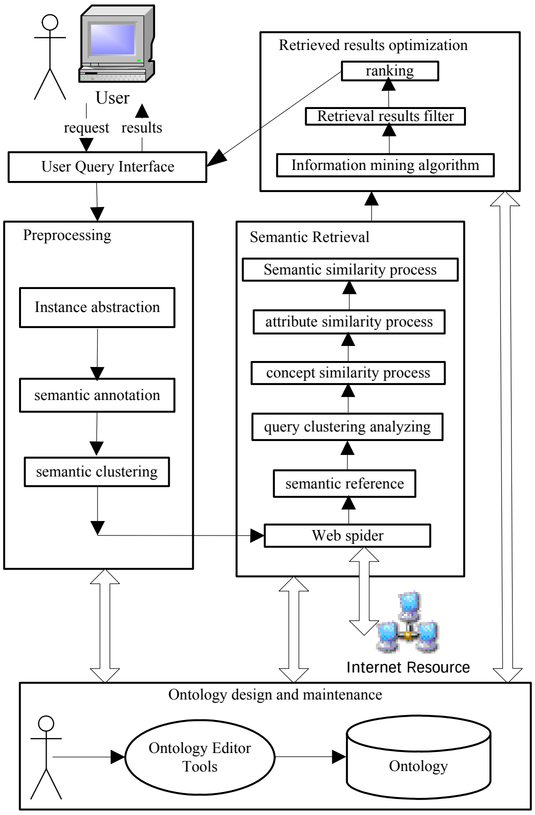


Fig.2. Architecture of Information Retrieval System based on Semantic Web

Ontology design and maintenance is the foundation of the whole architecture. OWL is used to precisely annotate syntax and the semantics of concepts, their attributes and relations among concepts, and it provides *equivalence between classes and properties* to express synonymous relationship. Support Tools will certainly be required for ontology design and maintenance, e.g. protégé.

The user query interface receives a user’s query requests, commits the query requests to preprocessing module, and returns optimized results to the user.

Preprocessing module is composed of three parts: Instance abstraction, Semantic annotation, and semantic clustering. Instance abstraction is the key step of the automated annotation. In order to analyze and compute the content of the Web page, it is necessary to preprocess the Web page. Resource preprocessing includes a sequence of operations:

- Remove HTML tags and acquire full free text.
- Analyze syntactic functions and identify words in documents.
- Tag part-of-speech in resources and annotate identified words.
- Remove stop-words and filter words with lower discrimination.
- Abstract stem and remove prefixes and suffixes of words.

Preprocessing can insure the treatment of documents more accurate; moreover it can enhance efficiency. The preprocessed Web pages can require necessary conceptual instances by classifying, clustering, and syntax schema learning. The abstracted instances will be stored in the base of instances. After abstracting conceptual instances, the next step is to automatically annotate the Web pages according to certain rules and domain ontology. Automatic annotation can be achieved by means of RDF and RDFS. The aim of semantic clustering is to cluster the documents that have been annotated and to abstract the center of semantic clusters. It is helpful to locate query intention in relevant classification to enhance retrieval efficiency.

Semantic retrieval module computes the similarity of semantic vector and commits the results to optimization module for further analysis and process. To increase search efficiency, which is worthy of notice, it is necessary to compare user query vector with the semantic center of document to identify the kind of user query object quickly, and then retrieve it from semantic clustering.

Retrieved results optimization module mainly retrieve valuable data using data mining algorithm, filters retrieved results, and ranks all the retrieved results according to its corresponding relevance to the retrieval items, then presents the ranked results to users.

4. PERFORMANCE ANALYZING

The fundamental retrieval performance evaluation measures of information retrieval are precision Þ and recall γ:

$$Þ=R_a/A,\ \gamma= R_a/R$$

In these two equations, A is the number of retrieved resources, R is the number of relevant resources, and R_a is the number of retrieved relevant resources.

The essential difference between information retrieval model based on the Semantic Web and keyword-based model is illustrated by a simple example.

Suppose there are 4 resources: r_1, r_2, r_3, and r_4; String x is in r_1, terminology A defined in ontology is used to markup x; String y is in r_2, terminology B defined in ontology is used to markup y; String z is in r_3, terminology C defined in ontology is used to markup z (A, B, and C belongs to the same equivalent class); String u is in r_4, and terminology D defined in ontology is used to markup x; Terminology A defined in ontology is used to markup user information need x.

From the statements above, it is known that relevant resources to user information need are r_1, r_2, and r_3. According to the keyword-based information retrieval model, information retrieval in "syntactic" level is conducted, r_1 and r_4 are retrieved, the precision Þ is 50% and the recall γ is 33.3%; According to ontology-based information retrieval model, "semantic" information retrieval can be conducted, r_1, r_2, and r_3 are all retrieved, so the precision Þ is 100% and the recall γ is 100%. It shows that information retrieval based on Semantic Web has higher precision and recall than that is based on keyword.

5. CONCLUSIONS AND EXPECTATION

Based on the existing semantic retrieval methods, the paper makes a deep research on information retrieval based on Semantic Web and ontology, proposes an improved semantic retrieval model with an overall architecture, and then gives an emphasis description on the key modules: Ontology design and maintenance, User query interface, Preprocessing, Semantic retrieval, and retrieval results optimization. In the end, a performance analysis for the improved semantic information retrieval model is presented. The analysis result shows that the improved semantic retrieval model can effectively improve the precision and recall of the information retrieval.

To implement ontology-based information retrieval on the Semantic Web, it is in urgent need to markup the existing web's content with terms defined in ontology. As ontology is very large, how can the suitable terms in ontology are effectively found and used to markup the Web's content? This needs further research.

REFERENCES

[1] Grigoris Antoniou and Frank van Hermelen, *A Semantic Web Primer*. Cambridge, Massachusetts, London, England. The MIT Press. 2004.

[2] Tim Berners-Lee, J.Hendler, and O.Lassila, *The semantic Web*. Scientific American. May, 2001.

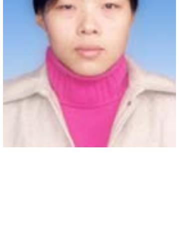
[3] John Davies and Richard Weeks," QuizRDF: Search Technology for the Semantic Web." in *Proceedings of the 37th Annual International Conference on System Science*. Hawaii, 2004,pp.112~119.

[4] José Saias and Paulo Quaresma," a Methodology to Create Ontology-Based Information Retrieval Systems,"in *Artificial Intelligence*,2003,pp.424~434.

[5] Takashi Hattori, Kaoru Hiramatsu, Takeshi Okadome, Bijan Parsia, and Evren Sirin, Ichigen-San, *An Ontology-Based Information Retrieval System*, APWeb, 2006: pp.1197~1200.

[6] Urvi Shah, Tim Finin, Anupam Joshi, R.Scott Cost, and James Mayfield,"Information Retrieval on the Semantic Web,"in *Proceedings of the 7th international conference on Information and knowledge management*, McLean, Virginia,USA, 2004, pp.461~468.

[7] Wenjie Li, Zhiyong Feng,Yong Li, and Zhoujun Xu, "Ontology-based intelligent information retrieval system," in *Electrical and Computer Engineering*, 2004, pp. 373~376.



Min Xiao is a PH.D student in the School of Computer Science and Technology, Wuhan University of Technology. She major in computer application, her research interests are Semantic Web and Data Mining.

Near-Duplicates of Web Pages Detection Algorithm Based on Single Fingerprint of Textual Chunk*

Dazhen Wang, Yuhui Chen
School of Computer Science, Hubei University of Technology
Wuhan, HuBei 430068, P.R.C
Email: chenyuhui7@163.com

ABSTRACT

Finding near-duplicates of Web pages is an important part in the field of search engine. As there are many noises in Web pages which are difficult to distinguish, detection algorithms in common use are all unable to get rid of the influence effectively at present. So the fundamental defect exists in these algorithms and the recall rate is comparatively low. This paper has proposed a near-duplicates' detection algorithm based on single fingerprint of textual chunks (SFTCA). It can solve the problem effectively. SFTCA chooses biggest several textual chunks of a Web page, links these textual chunks together, and produces a MD5 fingerprint which represents this Web page. If two Web pages have the same fingerprint we think they are similar. Finally, we compare SFTCA with two algorithms which extensively used, and prove the advantages of SFTCA by the experiments.

Keywords: Search Engine, Near-Duplicate of Web Pages, MD5 Fingerprint

1. INTRODUCTION

With the constant development of Internet, information on the Internet grows in the type of exploding. An increasing number of people are joining the "Information Age" via the Internet. Due to the quantity of information being very huge, people can only seek help from the search engine. In the result, however, they will find a large number of repeated Web pages which have the same core content while people using the search engine (such as Google [4]). Among these Repeated Web pages, some are mirror files of others without any change, some make slightly change while being edited and typeset.

How to get rid of near-duplicates has already become one key technology which improves the service quality of the search engine system.

First of all, by finding out these repeated Web pages and deleting them from the database, we can save memory space, and make use of the space to preserve more effective Web page content.

Second, by removing redundant copies the search engine can provide a neat and useful result to user.

Third, we can avoid downloading these pages again in the course of collection in the future. Thus the collection efficiency of the spider program will be improved apparently.

Fourth, if a certain Web page is repeated many times on the web, this Web page is more important in general. We should give this page a higher priority than ordinary page. The search engine system should give it higher weight value while responding users' requests and arranging display list of output result. It can also conduce to improve the service quality of the search engine system.

2. PRESENT CONDITION

Detection algorithms of near-duplicates of Web pages at present application system are mostly based on content detection.

There are three types of algorithm which utilize this detection logic.

(1) Near-duplicates detecting algorithm based on whole-text partition matching. SCAM (Stanford Copy Analysis Mechanism) [3] is an experimental prototype for finding intellectual property violations. Narayanan [1] and his colleagues proposed a betterment algorithm based on SCAM (We will call this algorithm Algorithm1). The new algorithm developed by Stanford has been applied by Google in their search engine system and has received favorable comments.

(2) Near-duplicates of Web pages detection algorithm based on keywords picked up from web pages. Wang Jianyong [2] proposes an algorithm (We will call this algorithm Algorithm2) of this kind is adopted by WebGather Search Engine System [6].

(3) The third method is based on eliminating noises originating from templates in Web pages.

There are two other kinds of algorithms used singularly in practical application: near-duplicate of Web pages detection algorithm based on anchors, near-duplicate of Web pages detection algorithm based on link information.

3. DEFECTS OF THE EXISTING NEAR-REPLICAS DETECTION ALGORITHMS

3.1 Algorithm1

Algorithm1 adopted a kind of algorithm based on whole-text partition matching. This algorithm divides a Web page into T sections according to a certain principle (such as being every N lines as one section), then signs (calculate fingerprint) to every section. Thus every Web page can be expressed with N signatures. As to two Web pages, when the number of same fingerprint is more than M in N signatures of them (M is the threshold value defined systematically), the author regard they as a pair of near-duplicates. This algorithm uses triplets of the form <archive identification (DocID), section identification (ChunkID), fingerprint (Fingerprint)>. This method reduces the algorithm complexity to some extent.

But this algorithm has two pieces of critical defect.

First, because the MD5 algorithm is very sensitive, and in the Web page there are many noises such as much template

* Hubei Provincial Department of Education Funding Project (ID: Q20074006)

information or advertisements, etc., it is very difficult to judge near-duplicates with threshold value M defined in unison.

Second, time complexity and space complexity still remain sizable. If it applies to the search engine system of magnanimity (there usually include billions of web pages on the Internet), it is still difficult to make the ideal result.

3.2 Algorithm2

Near-duplicates of web pages detection algorithm used by the WebGather Search Engine System is a single MD5 fingerprint elimination near-duplicates algorithm based on Web page purifying and Vector Space Model of text [5].

Each Web page is only represented by one MD5 fingerprint in Algorithm2, so its time complexity is relatively low and the precision is relatively high. Unfortunately this algorithm is based on key words. So this algorithm needs the process of purifying Web page and building the VSM of text. During these operations errors are inevitable. The recall rate is low because it is difficult to eliminate lots of noises existing in the Web page. More over this algorithm needs the process of purifying Web page and building the VSM of text.

Algorithm1 and Algorithm2 are based on content detection. They are interfered with noises in two ways. If, on one hand, the same core content in different templates, these algorithms may regard them as different pages. On the other hand, if the different text placed in the identical templates, these algorithms may regard them as similar pages.

4. SFTCA

Aiming at the advantages and faults of Algorithm1 and Algorithm2, in this paper we propose a near-duplicates of Web pages detection algorithm based on single fingerprint of textual chunks (SFTCA).

4.1 Description of SFTCA

SFTCA utilizes the tag tree structure to choose biggest several textual chunks, linking these textual chunks together, and producing a MD5 fingerprint which represents this Web page. If two Web pages have the same fingerprint we think they are similar.

If P_i and P_j are Web pages, function C() acquires the textual chunk consist of the largest several textual chunks of a page, and function Md5() obtains the Hash code of a textual chunk, $S(P_i, P_j)$ represents that P_i and P_j .are similar, we can express SFTCA as or reduce to a formula:

$$(MD5(C(P_i)) = MD5(C(P_j))) \Rightarrow S(P_i, P_j)$$

The details of the algorithm are as follows:

(1) Extract all textual chunks in a Web page utilizing the tag tree structure in HTML file. There are many HTML tags in Web pages that we can’t see in the explorer. The textual chunks we draw are the core content without tag information.

(2) Arrange these textual chunks in a queue in descending order according to the size.

(3) Choose the largest T textual chunks from the queue above, and consolidate them into a big textual chunk named "chunk".

(4) Remove the special characters that may be produced in edit or typesetting, such as the blank character and Enter character etc., then acquire the result string named "result".

(5) Hash down the string "result" into a 32-bit fingerprint utilizing the MD5 algorithm. The fingerprint represents this Web page.

(6) Compare MD5 fingerprints of two Web pages. If they have the same fingerprint, then we can conclude that the two Web pages are similar.

4.2 Analysis of SFTCA

The length of the template information (the website's common information in frameworks, hyperlinks, etc.) and advertisements are generally short and small. However, there are some longer textual chunks made of sentences in the core content of a web page. SFTCA plays the role of purifying the Web page by selecting the biggest T textual chunks of the Web page to a great extent. In this way it solves the difficult problem in other algorithms in this way: Get rid of the interference of noises and improve the recall rate. At the same time we can omit the purification course of Web pages and the work of building the VSM of text. So it reduces the workload.

In SFTCA, we consolidate T largest textual chunks into a textual chunk, and produce a MD5 fingerprint for every Web page, i.e. M =1. We can put the value into the time complexity and space complexity formula of near-duplicates web pages detection algorithm: $O(M^2N^2)$ and $O(MN)$, then work out the time complexity and space complexity of SFTCA as follow: $O(N^2)$ and $O(N)$. We can compare the results of different algorithms and get a result: the time complexity and space complexity of SFTCA are lower than Algorithm1 and Algorithm2.

Three types of algorithms are compared in terms of their time complexity in Table 1:

Table 1. Time and Space Complexity

Algorithm	Time Complexity	Space Complexity
SFTCA	$O(N^2)$	$O(N)$
Algorithm1	$O(\mu M^2 N^2)$	$O(MN)$
Algorithm2	$O(N^2)$	$O(N)$

Of course SFTCA algorithm is flawed as well. If the contents of two Web pages are similar as a whole, but one of them has revised several characters of the core content in a small quantity, SFTCA algorithm may regard these two Web pages as dissimilar. This defect is determined by the sensitivity of MD5 algorithm. However, the experiments prove that the defect has not brought serious problems. This also shows that a lot of Web pages are not generally revised the core text when the Web pages are spread on the Internet.

5. EXPERIMENTAL RESULTS

We have fetched 1,000,000 Web pages at random from the Internet as the data set of our experiments. We run the experiments on an ordinary computer, furnished with 3.2GHz of CPU, 1GBs of memory, 200GBs of hard disk, and running Linux OS.

Following, we prescribe that SFTCA represents the new algorithm proposed in this paper and the threshold value T=3, Algorithm1 represents Narayanan’s algorithm [1], and

Algorithm2 represents the algorithm used by WebGather [2].

5.1 Time Complexity

We have tested three kinds of algorithms separately in a test set consisting of 1,000,000 Web pages. The experimental result is as follows:

Table 2. Time

Algorithm	Time (Second)
Algorithm 1	7021
Algorithm 2	642
SFTCA	583

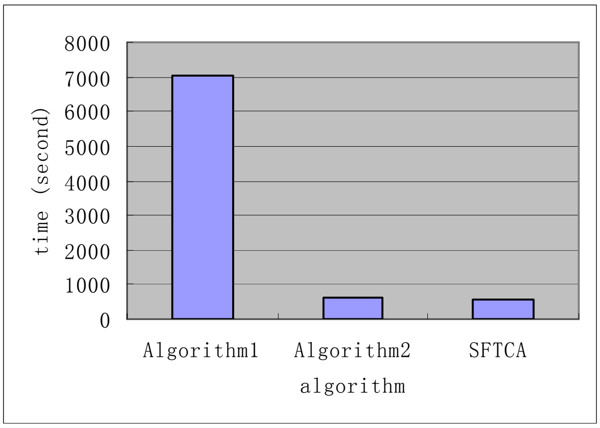


Fig.1. Time

Seen from table 2 and Fig.1, in a lager test set consisting of 1,000,000 Web pages, the time disparity between these three kinds of algorithms is very large. The result of Algorithm1 illustrates that it does not suit for operating on magnanimity data. The time of TW algorithm and SMFA algorithm nearly increase linearly. The time complexity of SMFA algorithm is obviously superior to TW algorithm.

5.2 Precision and Recall

In this paper, the term Precision indicates the ratio of real duplicates among all the web pages in the result set obtained by SFTCA, and the term Recall is a percentage between near-duplicates and all the pages in the data set. Test data of experiments is as follows:

Table 3. Precision and Recall

Algorithm	Precision (%)	Recall (%)
Algorithm 1	97.4	22.3
Algorithm 2	96.5	22.7
SFTCA	98.7	29.6

From Table 3 and Fig.2 we can see that the recall rate of SMFA is obviously higher than Algorithm1 and Algorithm2. As to precision, three algorithms only have a little difference, and SFTCA is slightly higher.

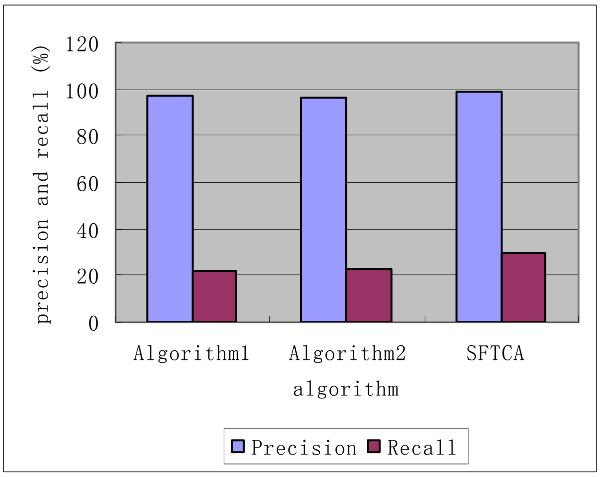


Fig.2. Precision and Recall

6. CONCLUSIONS AND FUTURE WORK

After analyzing the time complexity and space complexity of above three algorithms and testing in the experiments, we can find out the advantage of SFTCA. It can reduce the time complexity and space complexity efficiently of near-duplicates web pages detection. What's more, it can apparently improve the recall rate.

We can find out that the number of near-duplicates on the Internet is very large. The recall rate obtained in our experiments is higher than we had estimated.

Containing relationship between two Web pages can not be manipulated further in this paper. The case occurs in forum in general. So the search engine needs classify all the Web pages firstly. Then the content of a Web page needs to be divided into several sections properly. If we adopted Algorithm1, the time complexity is too high. How to resolving the problem efficiently is our next task.

REFERENCES

[1] Narayanan Shivakumar and Hector Garcia-Molina, “Finding near-replicas of documents on the web,” *WebDB* 1998. pp: 204-212.

[2] Wang Jianyong, Xie Zhengmao, Lei Ming, Li Xiaoming, “Research and Evaluation of Near-replicas of Web Pages Detection Algorithms,” *Acta Electronica Sinica*, vol. 28, pp. 130-132, Nov 2000.

[3] Narayanan Shivakumar and Hector Garcia-Molina, “SCAM: A Copy Detection Mechanism for Digital Documents.” in *Proceedings of the 2nd International Conference on the Theory and Practice of Digital Libraries*(DL’95), June 1995.

[4] Google Search Engine, http://google.stanford.edu.

[5] Bao-Yi W, Shao-Min Z, “A Chinese text classification model based on vector space and semantic meaning,” *Machine Learning and Cybernetics*.

[6] WebGather Search Engine, http://e.pku.edu.cn.

E-Commence Techniques and Applications

Application of EJBCA on Special Transportation Mobile Commerce

Liyi Zhang, Qihua Liu，Min Xu and Guo Chen
Center for Studies of Information Resources, Wuhan University
Wuhan,430072, P.R.China
Email: lyzhang@whu.edu.cn

ABSTRACT

The existence of wireless certificate authority is the basis for the existence of mobile commerce. This paper carries on the exhaustive analysis and the research of the opened source system EJBCA based on the J2EE, furthermore conducts the distribution and deployment in accordance with the required software on the Linux platform, on this basis, introduces the process of using EJBCA to build China special transport wireless certificate authority, hope to have the important significance of model to the independent own research and development of present domestic WPKI technology and product.

Keywords: EJBCA, Mobile Commerce, WPKI, Special Transportation

1. INTRODUCTION

With the development of communications, the safety problem of wireless transmission draws the people’s attention. On the one hand, despite the GSM use of advanced encryption technology, the mouth of the air between mobile phones and base stations is open. This provides opportunities of deciphering the codes of network communications for some people. And as soon as information leaves the network of operators, it will lose the encryption protection of this operator. So, the entire communication process all has possibility of interception of third-party, which include the establishment of communication links, information transmission, etc. On the other hand, in mobile communications systems, it doesn’t have the fixed physical connectivity between mobile users and Internet, how to identify the legitimate status of users and prevent users from denying the business behavior are some urgent security problems. In summary, the existence of wireless certificate authority is the basis for the existence of mobile commerce.

EJBCA is an enterprise class Certificate Authority using J2EE technology. EJBCA builds on the J2EE platform to create a robust, high performance, platform independent, flexible, and component based CA to be used standalone or integrated in any J2EE application[1]. It has all key components in WPKI environment and plays an important role in the realization of WPKI environment.

This paper carries on the exhaustive analysis and the research of the opened source system EJBCA based on the J2EE, furthermore conducts the distribution and deployment in accordance with the required software on the Linux platform, on this basis, and introduces the application of EJBCA on special transportation mobile commerce.

2. THE BASIC FRAMEWORK OF EJBCA

2.1 The Main Components of EJBCA

EJBCA system is mostly composed of Web component, RA component, CA component, LDAP server and database server. The framework of EJBCA is shown in Fig.1.

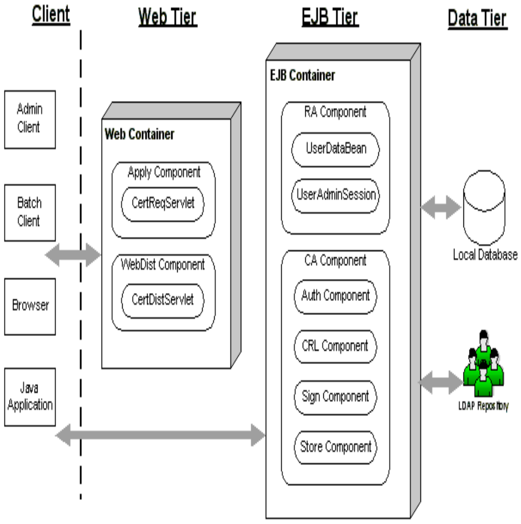


Fig.1. System framework of EJBCA

(1) Web Component
It is mainly faced to ordinary users, and used to provide some request and services between the application server (RA component and CA component) and client browser, such as CertReqServlet, CertDistServlet, CertBroServlet, etc. At the first, users receive certificates of Web component through application server. In the second, all communication between users and Web components, including some information of users and public key of browser, are encryption transmitted through encryption key of Web component. So, it is very safe to apply and transmit certificates, the process is shown in Fig. 2.

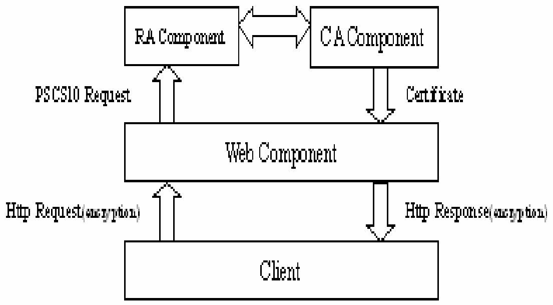


Fig.2. The process of information transmission in EJBCA

(2) RA Component
It is also named registered authority, mostly provides some functions of user registration and auditing. RA component plays a bridging role in EJBCA. On the one hand, it transmits services of Web component’s CertReqServlet and CertDistServlet to CA component; on the other hand, it transmits services of CRL and certificates which are given by CA component to LDAP and Web component.

(3) CA Component

It is the core component in EJBCA, can provide some functions such as CertDistServlet, certificate signature, certificate storage, CRL, SubCA foundation and etc. First, CA component have their own private key and public key, and then transmit certificates which are given by CA to RA component through Web component. CA also has responsibility to generate some digital certificate for all levels of administrations, such as Web components, subCA and RA. In EJBCA, types of certificate are optional. There are three types in the initialization time, ENDUSER (FIXED), ROOTCA (FIXED) and SUBCA (FIXED). In addition, users also can define their own types of certificates.

(4) LDAP Server

It provides service of catalog browsing, and charges for adding users' information and digital certificates which are transmitted by RA to servers. So, other users can receive their digital certificates through visiting LDAP server. In EJBCA, configurations of LDAP serve are optional; we can match certificates and their list to relevant LDAP servers through amending configuration files of LDAP [4].

(5) Database Server

It is a very important component in EJBCA, used to storage and manages users' information, digital certificates, diary document, statistical information and etc.

2.2 The Construction of EJBCA

(1) Software installation and configuration environment variable

Download and install the relevant software Jdk 1.4.2, JBoss4.0.1 SP1, ejbca3.0.7, jce_policy-1_4_2 and apache-ant-1.6.3. System database can choose from among SqlServer, Mysql, Oracle, and etc, we use Mysql in this article. It is necessary to configure environment variables after configuration relevant software. There are several environmental variables must be allocated: Jdk, Ant and Jboss. First, use sentences of "export" to assignment categories of Jboss, Jdk and Ant to variables of JBOSS_HOME, JAVA_HOME and ANT_HOME in the operation system of .Linux. Second, add these sentences to the tail of the document of "/etc/profile", which locate in the installation directory of EJBCA.

(2) The deployment of EJBCA

Implement the command of "ant" to compile EJBCA source code in the directory of EJBCA. The internal business logic and deployment descriptor of CA will be packaged into an enterprise application file of "ejbca-ca.ear" after running the command of "#ant deploy". Copy this file to the deployment directory of Jboss. So far, the entire CA system of EJBCA has been deployed to sever of Jboss. Use sentences of "CREATE" to grant the database of Mysql. But, it is necessary to establish an own certificate authority before running the EJBCA, this is root CA. And, it must be established on the J2EE sever. First, start the server of Jboss. Second, implement the command of "#install.sh" in the directory of "EJBCA". In the installation process, CA will create three types of certificates: client administration certificate, sever certificate and certificate which is signed by root CA. Client administration certificate and sever certificate locate in the directory of "P12", which is subdirectory of EJBCA directory, but the certificate of root CA locate in the letter. On the one hand, the certificate of root CA is automatically imported into the private key file of "carcerts" which locate in the Jdk security directory of "JAVA_HOME\lib\security", "JAVA_HOME" is installation directory of Jdk. On the other hand, the sever certificate is imported into directory of "JBOSS_HOME\bin"; "JBOSS_HOME" is installation of Jboss. Ii is necessary for EJBCA to manage CA in SSL layer, so client need to import the client certificate of "P12/superadmin.P12" into browsers, and then can manage CA through browsers.

(3) The configuration of Two-way SSL

The SSL protocol is a standard protocol to ensure the secure communication between the Web browser and Web server, which is developed by Netscape. It is located in the transport layer. It is seen as a standard security measure of server and web browser. The methods of configuration of EJBCA are as follows:

First, open the browser; input the address of http://localhost:8443/ejbca/adminweb; obtain the certificate of root CA. Second, add users of client and server in home page; designate the method of building private key. Third, input user names and password of client and server preserve the file of ".req" and install these certificates through the installation guide. A point worth noting is that we must award certificates to every client who can visit server and server through using root CA in the process of configuration of Two-way SSl.

3. THE CORE TECHNOLOGY OF WIRELESS CERTIFICATE AUTHORITY——WPKI

3.1 The Introduction of WPKI

WPKI is the initials of Wireless Public Key Infrastructure. It is not a new standard about PKI, but it is the wireless standard which extends to the traditional standard of PKI. WPKI is the core component of some security program on mobile commerce, such as mobile WAP (Wireless Application Protocol), WLAN (Wireless Local Area Network) and WVPN (Wireless Virtual Private Network).

3.2 The Communication Security Models Of WPKI [3]

WPKI defines three different communication security models:
(1) WTLS Class2, it uses the server certificates;
(2) WTLS Class3, it uses the client certificates;
(3) SignText, it uses client certificate and WMLScript.

We adopt the three communication security model at here. The mode of SignText provides a mechanism of Sending certificates of WTLS format to mobile devices, and can provide some functions such ad digital signatures and etc. Concrete steps are shown in Fig.3

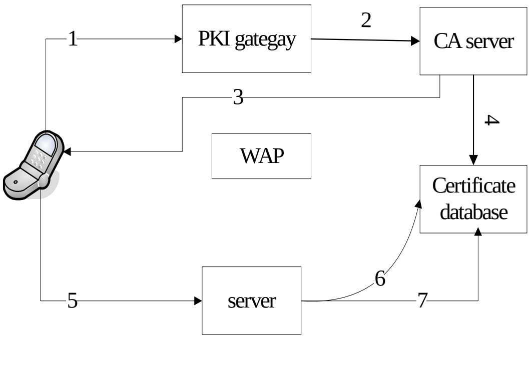


Fig.3. The concrete steps of SignText

In this mode, the root CA must provide certificates to mobile devices and servers. The concrete steps are as follows:

(1) The mobile user sends the request of certificate to PKI gateway;

(2) PKI gateway transmits the request to CA server after recognizes the ID of mobile user;
(3) CA server generates certificate of mobile use, and send the URL of the certificate to mobile user;
(4) If necessary, CA server send certificate of the use to certificate database;
(5) The user signs the transaction in the client and sends its content, digital signature and the URL of the user to server;
(6) Server uses the URL of certificate to receive the certificate of the user from certificate database
(7) If necessary, certificate database sends user certificate to server.

There is a two-way certification mechanism through using commercial servers and mobile devices in this model. So, it provides the convenience for the application of EJBCA on WPKI environment.

4. THE APPLICATION EXAMPLE OF EJBCA ON SPECIAL TRANSPORTATION MOBILE COMMERCE

This section focuses on the process of using EJBCA to build China special transport wireless certificate authority (CSTWCA). Through description of the example, a detailed analysis of how to use the famous system EJBCA for practical application is conducted.

4.1 The Introduction of CSTWCA

China special transport network is a third-party agency, which is independent from the enterprises of special logistics and transport. It provides all-way solution for some special goods such as big pieces, cold storage, refrigeration, constant temperature goods and dangerous materials. Its users are mainly drivers and maintenance workers. They have little opportunity to access to computers on working times. To meet the need of these users through mobile devices to certificate vehicles, drives and maintenance, we use EJBCA to build CSTWCA.

4.2 The Building Of CSTWCA

4.2.1 The design ideas of CSTWCA

There are two methods to build wireless certificate authority.
The first method is to create a new certificate authority to provide services only for wireless application protocol. This method is similar with building traditional certificate authority.

The second method is to expand traditional certificate authority to wireless area. The advantages of this method are: less hardware investment; the procedure of users only need small changes to support wireless certificate authority. And this method also has the following characteristics:
(1) RA server only provides services only for mobile terminal such as certificate establishing/compiling/canceling/deleting and CRL.
(2) Use wireless application protocol gateway to create a connection between wireless transport layer CA servers and secure socket layer CA server. It plays a bridging role.

We use the second method at here. It is to build traditional certificate authority——CHINA IC, and expand it to wireless area.

4.2.2 The building of China IC

China IC's main objective is to certificate entities of china special transport network, such as vehicles, drives, maintenance units.

The basic framework of the application example is shown in Fig.4.

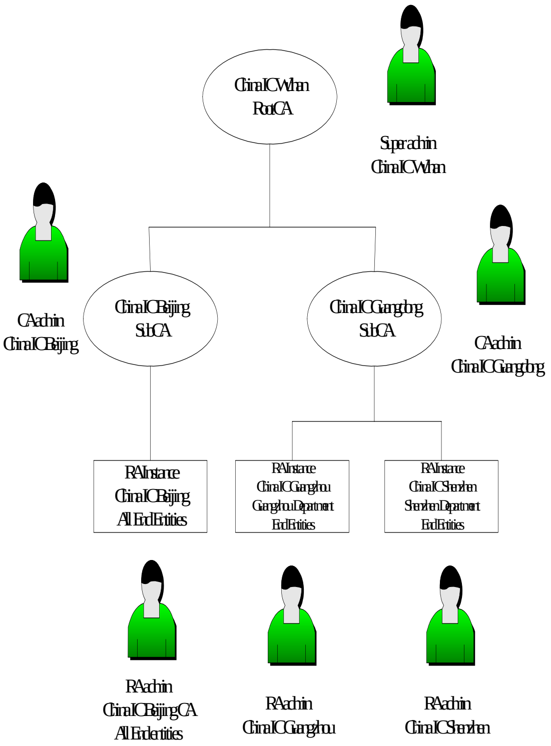


Fig.4. The basic framework of the application

Note: In this figure, circular used to show examples of CA, rectangular used to show examples of RA

Can be seen from the Fig.4, the application example has the following characteristics:
(1) China IC RootCA locates in Wuhan. It has two branches in Beijing and Guangdong. Therefore, each branch has a subCA; in particular, the branch of GuangDong all has businesses in the two cities of Guangzhou and Shenzhen.
(2) China IC needs three examples of RA to manage.

We will build the system of PKI/CA from the perspective of roles. The system is divided into four roles, and is shown in Fig.5.

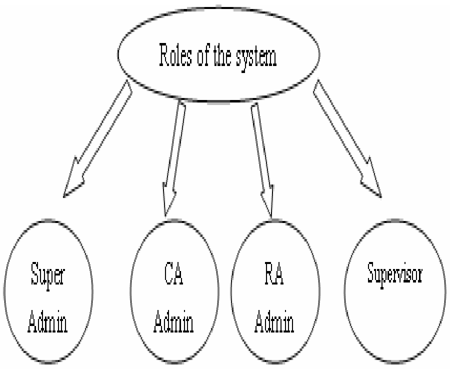


Fig.5. Roles of the application

(1) Super Admin

He has the authority to manage the entire system. As the role of super admin, he can do some things, such as editing system configuration, managing CA, building CA admin and etc. Detailed below:

① System configuration. Set up the title, slogans and language of the top and tail of pages, the theme and the number of data of each page. Choose the item of "Enable End Entity Profile Limitations" to manage RA, and set up two items of "Enable Key Recovery" and "Issue Hardware Tokens" to "unchecked".

②Manage publisher. The publisher connects some form of certificate storage system, whose certificates will be sent to the entity. A publisher is an LDAP directory or Active directory or publisher connector of definition established. We will build the publisher:

China IC LDAP

```
suffix "O=China IC,C=CN"
rootdn "CN=Manager,O=China IC,C=CN"
```

③Manage CAs. Now we need to build the structure of CA. Can be seen from Fig.4, root CA of China IC will have two subCAs. One is China IC Beijing, another is China IC Guangdong. We designate that every CA all has a private key of RSA. Its length is 2,048 spaces, and it also have valid for 10 years.

④Establish CA admin. In the PKI system, we allow the companies to manage their own certificates and RAs. We hope to have a major admin in each place. "China IC Beijing CA Admin" and "China IC Guangdong CA Admin" are administrations of China IC in Beijing and Guangdong. It is crucial that the administrators should not see each other's data, such as users, log, and etc, especially when two agencies are competitors.

(2) CA Admin

His responsibilities include managing certificate files and terminal entity files, configuration log and establishing RA admin.

(3) RA Admin

He is responsible for establishing/compiling/canceling/deleting terminal entity and seeing existent entities and their historical record. An RA only can manage the terminal entity of their own purview, so each other are transparency, as shown in Fig.6.

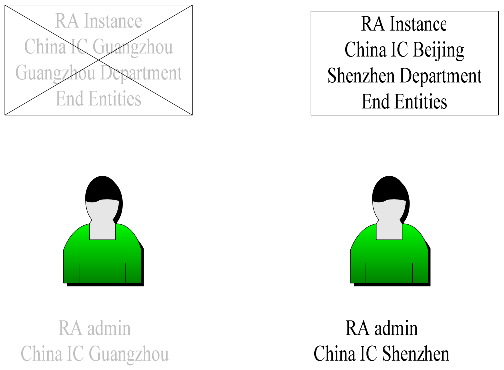


Fig.6. The relation of RAs

(4) Supervisor.

His responsibility is seeing entities and visiting logs.

4.2.3 The expansion of China IC on wireless area

We must do the following modifications to expand China IC to wireless area.

(1) The certificate storage modes need to change.

We use X.509 certificate in China IC. But CSTWCA has the objective limitations of transmission bandwidth and storage capacity on wireless environment, so we must reduce the storage of certificate. Here, we use mobile identification certificates (URL). A URL is used to show a standard X.509 certificate. Users only need to send own URL and signatures data to the other entities. And the other entities can retrieval corresponding digital certificates according to URL. URL only has a few bytes, and don't need to do any changes on X.509 certificates.

(2) WAP Gateway certification.

Wireless users don't retrieval relevant revocation of WAP Gateway certificates on wireless environment. In other words, OCSP and CRL don't achieve on wireless environment, but which can achieve on traditional PKI. Here, we use the short-cycle certificates. So, wireless devices don't need to consider the validity of the WAP Gateway certificates.

So far, we have completed the work of the building of CSTWCA. From the user's point of view, wireless equipment can complete certificate establishing/compiling /canceling/deleting only through WAP Gateway and user's communication interface of CSTWCA.

5. CONCLUSIONS

In summary, EJBCA is assembly simple, flexible, easy to manage. It can be applied to the security framework of mobile commerce through transplantation and appropriate allocation. EJBCA is a valuable opened source system, has the important significance of model to the independent own research and development of present WPKI technology and product.

REFERENCES

[1] EJBCA: readm.txt [EB/OL]. http://ejbca.sourceforge.net/do-cs/frame.htm

[2] Chen Qing, Ling Qinsheng: "The example of security CA ——researchment of EJBCA", *Computer Engineering and Design*, (2005)

[3] Zhou Bishui, Zhang Lei: "Research of EJBCA on WPKI environment." *Computer Engineering and Design*, (2005)

[4] PKI Tutorial. http://www.cs.auckland.ac.nz/pgut001/pubs/pkitutorial.pdf

[5] EJBCA-Architecture. http://sourceforge.net/project/showfiles.php?group_id=3971

[6] Chen Xiong: *The design and development of EJBCA*: [degree thesis], Wuhan University, (2006)

[7] Zhang Hua: "Research on the Security Authentication of Electronic Commerce and the Design and Implementation of the CA Model," *Computer technology,* (2006)

[8] Xiao Tianwei, Zhang Shiyong and Zhong Yeping: "Design and Implementation of PKI/CA-based Middleware System", *Computer Engineering,* (2006)

[9] Li Jicai, Liu Hailin: "Research on Secure Payment Protocols of Electronic Commerce", *Value Engineering* (2006)

[10] LAN Lina, LIU Xinyue: "Research on Security Architecture in E-Commerce System", *China Information Security* (2007)

[11] Shuai Qinghong: “The Analysis of safe Certificate in E-Commerce”, *Net Security Technologies and Application* (2007)

[12] Enterprise Text Message Platform. http://www.jrsoft.com.cn/Product/Aviation/sms.asp 2007/03/20

Liyi Zhang is a professor and dean of Department of Information & E_commerce in School of Information Management, Wuhan University. He graduated from Wuhan University of Hydraulic & Electric Engineering in 1988; from Xi’an jiaotong University in 1991 with specialty of Pattern Recognition & AI; from Wuhan University in 1999 with specialty of System Engineering. He is a member of E_commerce Major Guiding Committee of China, is Secretary-general of Association of Hubei Electronic Commerce, and a member of AIS(Association of Information System). He has published five books, over 40 Journal papers. His research interests include information system, e-commerce and information retrieval.

Qihua Liu is a postgraduate of information science of Wuhan University. He has participated in a number of enterprise application projects. From 2006, he started to study CA and made a lot of achievements. His research interests include information security, mobile commerce and information retrieval.

Min Xu is an Undergraduate of information management and information system of Wuhan University. From 2006, she started to study CA. Her research interests include information security, mobile commerce and information retrieval.

A Study on Application of Wireless Technology in E-Tax Management

Pinglu Chen
School of Management, Huazhong University of Science and Technology
Wuhan, 430074, China
Email: chen.pinglu@hust.edu.cn

ABSTRACT

The mobile tax transcends traditional E-tax through the integration of wireless technology. Mobile tax platform serves taxpayers any time and anywhere, which has the advantage of geographical convenience beyond the former one. The paper evaluates different stage of the mobile tax management development, analyses application of wireless technology in revenue sector. Finally, the paper discusses possible problems during the transformation to mobile tax, and gives advices for the future development.

Keywords: Wireless Technology, E-tax Management, Mobile Government, Mobile tax.

1. INTRODUCTION

E-tax is not new issue any more. As the leading E-government key areas, as well as human services, revenue, postal, education, justice & public safety and democracy [1], E-tax management is quite popular in developed countries and even some developing countries with mature IT infrastructure. Filing tax returns on line has been daily life for many citizens and cooperates. In China, metropolises such as Beijing, Shanghai, are proud of their leading edge E-tax technology.

Tax agencies have historically been among the first in the public sectors to deploy new technology, due to the relative ease of establishing an administration platform for faster revenue collection and increased compliance. Mature online service delivery gives tax agencies the opportunity to deliver a highly effective personalized service and reduce the costs of compliance.

Now we see the trend to mobile tax management as part of mobile government. Mobile government means extending power to every telephone and television set; they will enable governments to deliver services to almost every household and business. This will be particularly valuable in those countries where few homes have personal computer access. Wireless technology provides the capability to enable tax agents to reach out more effectively to taxpayers. According to Accenture, by providing services through, for example, interactive digital television, governments can not only reduce the cost of delivering services, but also encourage the broader take up of interactive television. For those nations, a further step will transmit the ongoing E-tax into the mobile tax management. It is believed that mobile tax will provide new and exciting opportunities to build the relationship between tax agencies and taxpayers [2].

2. REASONS FOR THE WIRELESS TREND OF E-TAX MANAGEMENT

Tax agencies worldwide are facing increasing pressure to accelerate both revenue collection and compliance levels. In order to achieve these goals, the leading tax agencies are articulating and implementing sophisticated online strategies. Online filing of tax returns is just one element of these programs, and covers a broad range of functions from simple, one-directional transfer of forms to sophisticated, interactive and transactional facilities. The research on mobility is expanding rapidly with increased attention given to the rapid diffusion of mobile technologies such as GSM telephones, PDAs, microwave Local Area Networks (for example Bluetooth), and Wireless Application Protocol (WAP). The number of mobile users is increasing and in many countries, through wide spread use of advanced mobile phones, has already surpassed the number of households with internet access[3].

2.1 Benefit of mobile tax management

The benefit of E-tax is extraordinary yet, tax authorities are usually organized in departments each with own responsibilities, tasks, structure and taxpayers. IT infrastructure and equipment as well as corresponding technical background knowledge combine different working process from door to door. E-tax is seen as a promising approach to harmonize the quality of public services and to overcome related segmentation phenomena. Mobile tax management goes even further; the function of it is many-fold and encompasses the following advantage and benefit.

Economical sustainability. The participation in the mobile tax platform is thought to be open to all interested stakeholders, and some have limited financial capabilities to purchase enough PCs on an individual basis. The wireless framework for tax administration takes into account the diverse needs and interests of public, which will cut the cost during the taxing compliance procedures.

Enlarged access to public information services. Mobile tax management ensures broad access by a significant part of the population; the platform provides openness and interoperability with regard to the interconnection with different networks, the integration of external content providers and public authorities providing their services and must further consider diverging mobile device characteristics and capabilities. Computers generally do not travel with taxpayers, mobile tax provides for instant availability of services and information, helping frequent travelers and people on the move to access tax authorities.

Intuitive and user-friendly mobile interfaces. Levying services are designed taking into account heterogeneous user characteristics, addressing the common needs of the taxpayers with different educational or even cultural background, age and interests, allowing for easy access to and search of information considering location, context and user interests.

2.2 Three stages of mobile tax developmen

There are lots of standards identifying the development phases of mobile tax management; one salient feature is the heterogeneity of communication mechanism between agencies and taxpayers. Dealing with heterogeneity means that active (user-initiated) and passive (system-initiated) adaptation mechanisms should be considered for filtering and presenting relevant information to users. However, intelligent systems raise specific usability issues related to the need for users to control their system. Research must explore acceptable guidance strategies that will help users decide how they to set their preferences. It is also important to identify the procedures by which an intelligent system could maintain user's awareness of its current setting, allowing them to easily understand how and why it behaves differently in different circumstances.

Accenture defined three level of service maturity in traditional E-government, which could be applied in mobile tax management. As shown in Figure 1, the first level is defined as Publish stage, also called Passive/Passive Relationship. The taxpayers do not communicate electronically with the tax agency and the agency does not communicate (other than through what is published on the WAP website) with the taxpayers. The second level is defined as Interact stage, also called Active/Passive Interaction. The taxpayers must be able to communicate electronically with the tax agency, but the tax agency does not necessarily communicate with the taxpayers. The last level is Transact, also called Active/Active Interaction. The taxpayers must be able to communicate electronically with the tax agency, and the tax agency must be able to respond electronically to the taxpayers.

Right now, some nations are at the Publish or even Interact stage, the third stage still remains in labs and experiments. In Europe, USE-ME GOV is the leading project in the fields; in United States, some metropolises are trying some advanced Transact platforms, which are applied in public sectors beyond the tax authorities.

The core technologies of USE-ME GOV program are the open standards and Service semantics as following. Openness is ensured by the utilization of open standards, namely with regard to W3C that is regarded as the leading organization for open standards provisioning.

1) HTTP, XML for transmission and message encoding

2) WSDL, SOAP and XML for message representation, transport and resources representation

3) RDF, OWL and OWL-S for service description

4) UAProf for representation of terminal capabilities and preferences

5) OWL DL for simple inference rules

6) XACML for provisioning of authorization rules

As service semantics concerned, for services to be properly discovered and provisioned, an appropriate unambiguous and coherent service description must be provided. Platform design therefore included the development of an OWL-S model to which several extensions were added in an iterative process. The final model serves as a backbone for advances service discovery protocols and is believed to have general applicability.

3. APPLICATIONS OF THE MOBILE TAX

Following the Accenture strategy, supporting from government is important for mobile tax administration[4]. Accenture believed that the leadership in the tax sector is closely correlated with E-government leadership. Countries lagging in the E-government leadership stakes would be well advised to focus their limited resources on their Tax agency's online strategy. It is clear that starting with Revenue has proven to be a smart strategy given that results can be driven quickly through E-Filing programs. In addition, they are highly visible given that almost every citizen and business is a customer of the Tax agency, and the accelerated collection and increased compliance can drive development of more mobile government initiatives. Furthermore, providing Revenue services online to business builds business confidence in E-government and in eCommerce generally. These services can act as a proof of concept for E-government as a whole.

Table 1 shows the five possible applications in mobile tax management classified by the interact direction between tax sector, business and citizens. Mobile tax is not a replacement for e-tax, rather it complements it. While mobile devices are excellent access devices, most of them, particularly mobile phones, are not suitable for the transmission of complex and voluminous information. Despite the emergence of more sophisticated handsets, mobile phones do not have the same amount of features and services as PC-based internet applications.

Table 1. Five Types of Mobile Tax Applications.

Type	Definition	Application Examples
G2G	Government to Government	Official document transformation
G2B	Government to Business	Corporate income tax return delivery
B2G	Business to Government	Business balance sheet upload
P2G	Person to Government	Individual income tax return delivery
G2P	Government to Person	Personal private information verification

In the G2B and G2P modes, from a taxpayer's perspective, mobile tax services stands for new front-end access to public services that have been made available specifically for mobile devices or adapted from existing e-tax applications. G2G means that tax officers are using handheld and wireless devices to provide more access to public data, which enable employees to communicate with each other, and give them another tool to do their jobs more efficiently. In the B2G and P2G modes, mobile devices can also be used to make payments and other transactional services. For example, Norway has introduced a mobile tax-collecting system. Taxpayers who have no changes to make to the tax form they receive, can now simply send a text message with a code word, their identity number and a pin code instead of returning the form by mail.

The mobile tax-collecting system, together with the grid management and process reengineering has enabled the tax officers to better manage its mobile work with both efficiency and effectiveness. Through the split of the enforcement and supervision, the process is changed, and stimulated the resolution of the problem. The reinforcement of the coordination functionality has facilitated the information flow between the fragmented departments. In mobile collection implementation, the most important issue is the alignment of organizational change with organizational strategic goals, followed by information flow integration and then technology issues. Mobile technology thus must go together with other management measures. The involvement of the top leadership of the district in the initiative and their full support has certainly been a key factor to success.

4. OBSTICLES IN THE REALIZATION OF MOBILE TAX MANAGEMENT

No one doubts the utilities of mobile tax management, however, to realize it is not always with ease. For example, typical mobile terminals feature a small screen and limited, potentially error-prone input devices, which leads to troubles of identifying appropriate alert mechanisms, new navigation functionality and input modes, and adaptive information presentation. What is more, tax authorities should consider the obstacles in security, platform performance and other fields.

4.1 Security concerns

Perhaps the most important issue in mobile tax management is security concerns. Considerations of data security and protection of individual privacy are assuming increasing pressing. Mobile tax programs typically incorporate approaches to issues such as uniform privacy practices, digital signature standards, and encryption standards for sensitive information. Tax authorities recognize that they must be able to assure taxpayers that the information they are maintaining is secure and will be protected from unauthorized use, and that taxpayer’s privacy will be protected. The level of security must be appropriate to the service concerned, as not all transactions require the same level of security.

Security is also common issues in public affairs[5]. Consider the plethora of numbers that an individual requires to interact with their government; passport number, driver’s license number, health insurance number, social security number, identity card and so on. The efficiencies possible in integrating this proliferation of identification points into one single record are considerable. Equally outstanding is the risk if this information is abused. Identity theft is a very real concern under a scenario where each citizen has a single record. A single record however is not necessarily the only solution to this issue. Governments are looking at alternative ways to bring this information together without the need for a single piece of identification, an approach that should alleviate citizen concerns about the risks associated with a single identifier.

Tax agencies must be able to guarantee the security of personal information. The technology to enable this security is now commonplace. The challenge tax agencies now face is putting the right processes in place to ensure public confidence in government information management. Authorities must convince its citizens that it can manage the risks appropriately.

4.2 Performance concerns

Another technical concern is the performance of the platform. The network platform is always WLAN. A fairly common issue with WLAN is that coverage areas with low signal strength often exist. This occurs because of poor placement of access points and the dynamics of the environment. After the initial installation of access points, for example, tax authorities might add walls to create new offices or move large machinery. This affects the propagations of radio waves, which leads to lower and sometimes inadequate signal strength in parts of the building.

Some applications, such as e-mail and Web browsing, hold up pretty well as users roam through coverage holes. At least a user can read e-mails or view Web pages in cache on the mobile device while on route to a covered area. However, a remittance auditing application commonly requires a constant connection between the terminal and the application server host. If the wireless connection is temporarily lost, the user usually must log back on to the system. Sometimes this can even cause errors on the server if the loss of connection occurs in the middle of a transaction.

For the taxpayers, the integration of multi-channel service may cause performance problem at client side. Services rich in interactivity require fairly advanced mobile phones. The co-existence of phones from several technological generations on the market could still limit or at least retard the success of deployment[6].

4.3 Other concerns

Beyond the technological issues, there are other concerns of mobile tax service. In dependence of the particular characteristics of the service, the integration of content from the authority side requires structural changes to organization, the administrative work process and IT infrastructure. For example, the automatic delivery of personalized notifications (e.g. confirmation that a tax certificate is ready) impacts quite significantly on the usually implemented work-flow. Social factors related to digital segmentation phenomena, by default diverging levels of user interest and acceptance for mobile services depending on social background, age and educational level.

So far, the management of tax collecting in China is highly geographically constrained by boundary and government organization is highly hierarchical. Through the integration of mobile system and GIS, GPS enabled grid management. The fluid interaction enabled by mobile handset in local or regional mobility has fitted well in to a hierarchical tax authority’s structure. We can observe a more fluid information flows from supervisor at field to command and supervision center and vice versa. What is more interesting from this case is that the supervisors at mobile work have closer contact with the citizen and build up trust with them – something much easier to do in the local context.

5. CONCLUSIONS

Mobile tax goes far beyond public service delivery and encompasses re-organization and structuring of public administrative processes. For decades, the driving forces behind the introduction and adoption of information technology in revenue organizations have been efficiency and productivity, better knowledge management, simply more and better information to support the business process as a whole, and finally – increased competitiveness. These processes have invariantly brought along significant impacts on the organizational structure and on all kinds of work-flows, processes and activities.

Wireless technology will make tax authorities quickly mobilize its broad resources to accelerate results for taxpayers; there will be a bright future for the integration of the technology and government agency. Tax authorities ought to learn that: platform sharing explored on the basis of attractive business models would also provide the conditions for cost-efficient mobile services in geographical areas with low internet penetration. The software application will be designed in order to meet the criteria for openness, interoperability, scalability and security. Following a suitable strategy, the public sectors will pay particular attention to advanced solutions for discovery and binding of e-tax services that are associated with the physical environment of mobile users.

REFERENCES

[1] Accenture. (2002). E-government Leadership: Realizing the Vision. The Government Executive Series.

[2] Accenture. (2003). E-government Leadership: Engaging the Customer. The Government Executive Series.

[3] Roggenkamp, K. (2004). “Development Modules to Unleash the Potential of Mobile Government: Developing mobile government applications from a user perspective,” *Proceedings of the 4th European Conference on e-Government*, Dublin, Ireland.

[4] Accenture. (2004). E-government Leadership; High Performance, Maximum Value. The Government Executive Series.

[5] Song, G. (2005). “Mobile Technology Application in City Management: An Illumination of Project Nomad in UK.,” *Municipal Administration and Technology*, Vol.7(3) 103-106.

[6] CNNIC. (2006). “Statistical Report on the Conditions of China’s Internet Development,” *China Internet Network Information Center,* Beijing.

Pinglu Chen is an Associate Professor and vice chair of public finance department in School of Management, Huazhong University of Science and Technology. He graduated from Huazhong University of Science and Technology in 1993; with specialty of Management Science and Engineering. He was a visiting scholar of Prague University of Economics (2004~2005). He has published over 30 Journal papers. His research interests are in E-government, public finance, and e-commence.

The Technology for Bill Data in Commerce Environment*

Dawei Jin[1,2] , Yongyue Chen[1], rui Wan[3]
[1] Research Center of Information Resources, Wuhan University, Wuhan, China 430072
[2] Information school, Zhongnan University of Economics and Law, Wuhan, China 430060
[3] China Constructive Bank Research Center, Wuhan,China, 430000
Email:jindaweiok@yahoo.com.cn

ABSTRACT

In the actual application, the multimedia data is more and more widespread. The electronic bill the now research key. In this paper we discuss the technology for these bill data. And we provide a successful applied example for bank.

Keywords: Content Management, Non-Relational Data, OCR Recognition, RGB and HIS Model, Sobel Edge Examination Technology, Image Retrieval

1. INTRODUCTION

Under the electronic commerce environment the data form is more complex, therefore content management method is introduced. We determine non- relational data for the content management object. But under the EC environment, non-relational data which we most common face is picture bill. In this article we mainly discuss bill's application which is the extremely universal in EC environment.

2. TRADITIONAL INFORMATION STRUCTURE

2.1 Data Was Preserved in Relational Database

We extract the data from the bill by using the artificial method, and store the relational database after the standardization . Its shortcoming is: (1) According the different form bill, we must design the different transformation method. For example past we frequently developed the report form system, which actually was the bill electronic form. When the new demand requested, the system must often be redesigned. (2) The stamp information is unable to process. In daily commercial activity, The seal expresses information which is very important. For example bank check seal or each kind of contract official seal are important information in EC. If we merely depend upon the bank staff's skill, which is very difficult to guarantee the higher recognition rate. Therefore the EC system treating processes set the request to this kind of information.

2.2 Picture Data Processing

In order to solve the basic memory problem, we transform the picture data into the binary flow and deposit in database big field. But efficiency is obviously is lower when we save the multimedia data than text data. It is especially obvious by the data quantity increasing. For example during one day bill data quantity possibly achieves G the magnitude in the bank.

2.3 Therefore We May Summarize These Problems Which We Will Face with When Deal with Non-relational Data.

Firstly, we should transform the picture for the electronic data. Secondly, we will how to store picture data. Finally we should withdraw the useful information from the picture data.

3. UNDER EC ENVIRONMENT INFORMATION STRUCTURE

3.1 In the Practical Work, We Discover the Most Ideal Way Which Unifies Two Merits

Firstly, we use scanner to electrify the bill. Secondly, using the text information recognition technology we withdraw the text information and stores the relational database. To graph data in the bill, we can withdraw useful information to examine through the special algorithm. Its logical value is deposited in the database. The kind of graph data is withdrew from bill which may is stored in distributional file form. But we must create index on these picture document. In this process, the data storage, the text recognition and the picture recognition technology is the essential key.

3.1.1 The Scanning System Uses The Opening Interface

Therefore it can support each kind of low speed, the medium speed and even the high speed documents scanning. In scanning module，when scanning is called, we use international standard's TWAIN interface. To scans outside support equipment it has not special demand.

3.1.2 The Picture Amendment

In order to cause the picture clearer and take the space smaller, some operation must be carried on include: rotation, elimination spot, removing black side and so on.

3.1.3 OCR Recognition

OCR software is inserted into the documents scanning module. Through calling the scanning template which has custom-made according to the demand, we can carry on the partial picture data recognition and finally these recognition essential data are putted into the database. The work of the template custom-made should be completed in the parameter platform and the operation easily makes every effort, to be easy to grasp.

3.1.4 The Pict Ession Storage

Using image the undamaged compression technology, it will reduce image document to storage space and efficiency influence in greatest degree. In order to undamaged restore image, the appropriate image decompression technology is used. The image compression uses the CCITT IV compression protocol and standard TIFFlevel6 document depositing form. The 200DPI black and white image is approximately compressed into 50K and the colored image compression approximately is 150K. The database which handling ability is strong to big object should be adopted. Oracle 9i is recommended. We use three levels storage strategy. NAS memory is taken as the on-line memory and the library of tapes is taken as the near-line memory and the compact disc of engraved records is taken as the off-line memory. The data during half year is preserved in the disk array, the data during two year is preserved in the library of tapes and data surpassing two years is engraved records into the compact disc the off-line preservation.

*[*] Supported by the NSFC project (No 70473068).

3.2 Key Technology for Bill Recognition

Using the equipment of scanning ,the system firstly transfer the bill which is filled by customer into gradation image, then remove the bottom color of bill and change into black and white image. After black and white image got ride of noise, the writing is divided which is subdivided several parts including rectifying, the essential factor localization, two values, cutting, and the writing merging. The primitive bill is separated into several essential factors by this process including amount, date, bar code, stamp and so on. Each kind of essential factor we have the different processing way. For example the single Chinese character separately is stored into the Chinese character lattice, then these Chinese character lattice are carried on character recognition process. Because in certain types bills, even if for the identical bill, its various essential factors background noise is not all same, therefore the different two values method are used to each different essential factor region. After dealing with, various essential factors already became the independent character lattice. The handwritten capital Chinese character, the block capital Chinese character as well as the block small letter number are stored into 64×64 the lattice. But the handwritten small number are stored into 96×80 the lattice. About this step many mature technology are regarded, also in reality they are widespread applied.

Now we lay special stress on analyzing recognition technology for stamp.

3.2.1 The Color Space Transformatio

In order to gather the data from image, we must describe the color through the quota method. In a word, color model is established. Our commonly used model has three kinds, namely computation color model, industry color model and visual color model. But the industry RGB model has obvious the shortcomings. Firstly, the RGB space is the color demonstration space, it does not suit person's visual characteristic. Secondly, separately process to three color components can bring the color information the loss and confused. Therefore we need to use the chromatic information to carry on the picture division. Because the different region has the different Hue, degree of Saturation and Intensity. According to the above, the background, the seal and the noise area may be separated. Therefore, the picture maps from the RGB space to the HSI space and obtains H, S, I three components histogram.

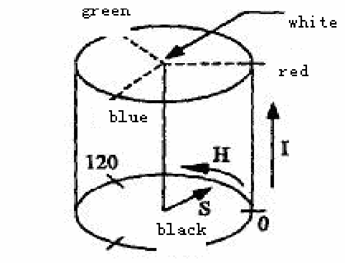


Fig.1. RGB space to the HSI space

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{pmatrix} \sqrt{3}/3 & 0 & 2/\sqrt{6} \\ \sqrt{3}/3 & \sqrt{2}/2 & -1/\sqrt{6} \\ \sqrt{3}/3 & -\sqrt{2}/2 & -1/\sqrt{6} \end{pmatrix} \begin{pmatrix} I \\ S\cos H \\ S\sin H \end{pmatrix}$$

Fig.2. The Transformed formula of RGB space to the HSI space

3.2.2 Sobel Edge Examination Technology

It is one kind of non-linear edge examination algorithm. The efficiency is very high. The essential method is in x, the y direction weighting summation. It will obtains the weighting picture element value of some one picture element which is separately x, y. This picture element boundary intensity q and the direction p are calculated by serviceable formula:

$$q = \sqrt{X^2 + Y^2}$$

$$p = \arctan(y/x)$$

Fig.3. Intensity and Direction

This method computation quantity is extremely big in the actual application, therefore frequently the approximate method is used. the current examination the region is:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

Fig. 4. Research region

Passing central point e straight line have four, including a—e—i, c—e—g, b—e—h, d—e—f. Each straight line all divides the region two parts one of which is called sub region and has three picture elements. For example, straight line c - e - g divides the region into a, b, d and f, h, i two sub regions. By calculating difference two sub regions which is divided by some straight line have absolute value of mean value. The biggest in four values is assigned for middle picture element e.

3.2.3 Example Explanation

In automatic appraisal system for the seal, the system tolerance is good, namely it can correctly distinguish and does not allow receiving the wrong seal. Because seal above the bill does is influenced by the ink pad quality, how many ink pad and so on. And bill itself also has the base map, the horizontal written line and the different gradation. It can cause some phenomenon including the color depth uneven, the picture fuzzy and so on. The scanned seal picture may be divided into three parts: white or gray background region, red seal representative's goal region as well as disturbance noise, black or blue color handwritten signature, block letter character and so on. Using the above technology one seal picture will be divided into these three parts.

- The seal withdrawing

In HSI space seal's Hue change generally between 0° and 20° and between 340°and 360°. It present coordinates 0 and 360 two side interrupted distributions situation in the histogram. When angle in above scope changing, the hue cosine value all is bigger than 0.94, then the feature can be utilized. Like this we not only avoid the histogram the coordinate transformation, but also avoid the trigonometrical function inverse operation and reduce computation complex and the computation quantity. After hue division, but disturbance also is caused light red and black background. At the same time, seal red shade is uneven because an effort is not consistent when sealing. The degree of saturation decides the colored shade degree. In the actual application, the person eye may accept red degree of saturation which at least is bigger than 20%. Therefore we select 20% red degree of saturation as division threshold value which may very good remove the background disturbance. After hue and degree of saturation division, the last disturbance is the black. Because the HSI space brightness is most depended on by the illumination effect, when the light changes strongly or weakly all cannot well remove the black the disturbance. According to Adobe Corporation Photoshop the colored model, its Brightness is taken as the brightness decision variable. Its definition is: I=Max (R, G, B) /255. Under the limited hue and the degree of saturation, by observing Photoshop the color plate, red I is at least bigger than 60%, the black and the gray I all is smaller than 50%. The threshold of 60% brightness not only avoids illumination influence but also eliminates black disturbance. Through above analysis, the seal to withdraw the model is:

$$\begin{cases} 0.94 < \dfrac{[(R-G)+(R-B)]/2}{\sqrt{(R-G)^2+(R-B)(G-B)}} \leqslant 1 \\ 0.2 < S \leqslant 1 \\ 0.6 < \dfrac{\mathrm{Max}(R,G,B)}{255} \leqslant 1 \end{cases}$$

Fig. 5. The model of the seal withdrawing

- After withdrawing the red information , table line and the bottom grain disturbance must also be removed.

This function weaken picture low frequency part by strengthening edge. After picture dealt with, image brightness maintenances unchanged and the picture element changing slowly region changes black. Thus the picture element changing fiercely region prominently came. As a result of frame line itself together, we firstly use the method based on the Sobel examining edge whose essence is the union of direction difference operation and the partial even homogeneous. After the Sobel algorithm dealing with each picture element in image, the output picture makes threshold value processing. When the central point picture element value is bigger than the threshold value, the output picture element sets white, otherwise sets black. Finally obtains the picture is only contains the edge the black and white two values chart.

- Rebuilding the real

In the 300DPI resolution, the stroke of the gathering 500×500 picture width approximately is 5 picture elements. Firstly, symmetrically distribution packing algorithm is used. The thinking of Algorithm is that in the random white spot 3×3 the neighborhood, we count the symmetrical distribution number of black spot. If number is more than 2, this spot is set black. The iterated action doesn't stop until all satisfied white spot is filled. Then Algorithm quit circulation and complete the packing. This packing algorithm can very good fill the small hole.

Then the evenly distribution the packing algorithm is used. The thinking of Algorithm is that in the random white spot 5×5 the neighborhood, if the coordinate center of mass is equal to the central point coordinates, the central point is set black.

Finally, the even value filter method is used. The value in assigned region is arranged by size. The middle picture element value was assign to the central spot. Because the picture element value in the region can occur random changes suddenly which is the end or the bottom on the array. Therefore the middle picture element value is the normal picture element value. By this step some random noise is got rid of, and it provide the guarantee for further the seal recognition.

Through the above step we complete the seal recovery. Then the primitive seal is downloaded from the database performs to judges the seal false or truth.

4. THE EXAMPLE OF PROCESS AND METHOD

The above technology we all already realized in China Contracture Bank accounting system. Under is this module flow chart and the class function.

The first model is primitive template of kinds of bill. Its process is:

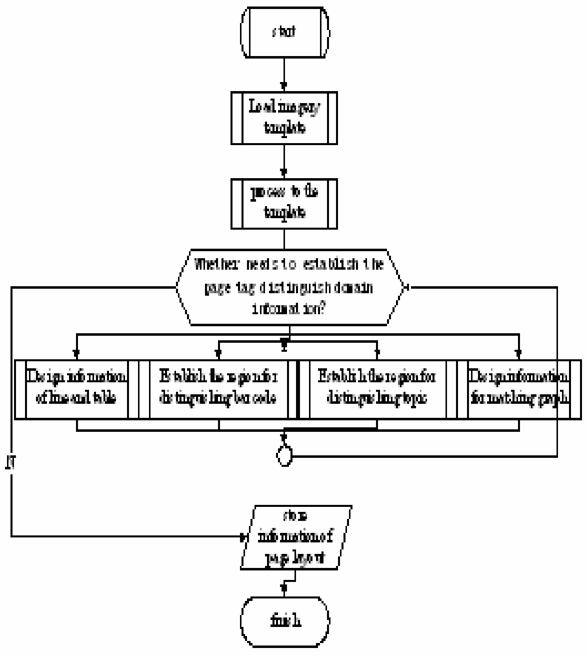

start
Load imagery template
process to the template
Whether needs to establish the page tag distinguish domain information?
Design information of line and table
Establish the region for distinguishing bar code
Establish the region for distinguishing topic
Design information for matching graph
N
store information of page layout
finish

Fig.6. Process of designing template

The second model is reorganization system.

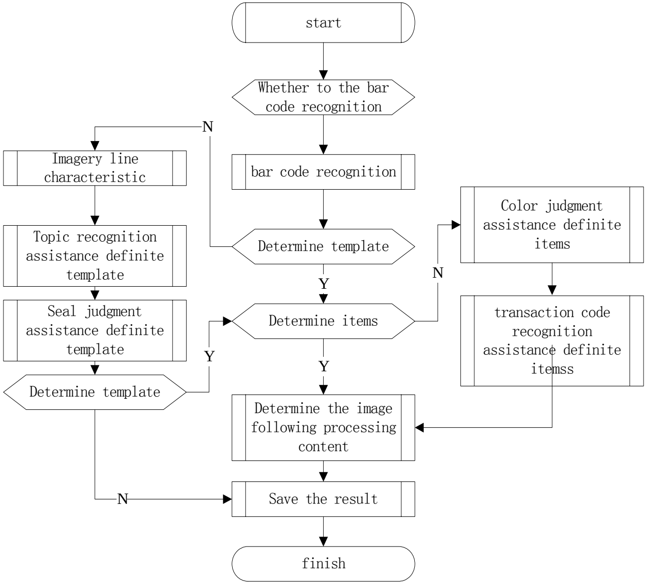

start
Whether to the bar code recognition
N
Imagery line characteristic
bar code recognition
Topic recognition assistance definite template
Determine template
Y
N
Color judgment assistance definite items
Seal judgment assistance definite template
Determine items
Y
Y
transaction code recognition assistance definite itemss
Determine template
Determine the image following processing content
N
Save the result
finish

Fig.7. The recognition for the template and items

5. SUMMARY

This module has already entered the implementation stage, at present the situation run to be good. Its main technical risk mainly manifests in the OCR technology. After many years development, the OCR technology has been mature gradually. But the recognition rate was still low to Chinese and the handwritten form. But because application object in certain scope, the recognition rate may enhance through the threshold value and the manual intervention.

REFERENCES

[1] Zhou Ning, Yang Feng, Liu Wei, "A study of Methods of Visual Interface for Digital Libraries", *Journal of library science in China*, No.4 2004 Serial No.152 62-66

[2] Gonzalez R, C., Woods R, E. *Digital image processing, Beijing*,The sciense publishing house. 1981

[3] M.J.Swain and D.H.ballard, "Color indexing". *International Journal of Computer Vision*, 1991,7(1):11-32

[4] Liu Zhongwei, Zhang YUjin, "The comparitive and analysis study of ten color feature-based image retrieval algorithms", *Singal processing*, 2000(1) (in chinese)

[5] M.Stricker and M.Orengo,"Similarity of color images", SPIE, V.2420,1995: 381－392

Dawei Jin is studying in Wu Han University for PHD and working in Zhongnan University of Economics & Law. The Main research direction is Knowledga Mangement and imforation Visualition.

Research on Vendor Selection Based on Entropy Weight and TOPSIS in Supply Chain

Rong Chen [1,2] ,Peide Liu[2], Shukun Tang [1]
[1]Management School,The University of Science and Technology of China, He'fei Anhui, 232000, China
[2]Information Management School, Shandong Economic University, Ji'nan Shandong, 250014, China
Email:Heron121@163.com

ABSTRACT

Under the background of globalization, the market competition is not means the competition among the enterprises anymore but the supply chain. The vendor selection is not only the function of supply chain cooperation but also the key factor for improving competitive strengths of supply chain. Firstly, this paper analyzed the worldwide study actuality of the copartner selection for supply chain. Based on some related literature this paper constructed an indicator system for copartner selection in supply chain and adopted the information entropy approach in order to confirm each attribute objective weight. Then counted the distance and relative closeness between each vendor with positive ideal solution (PIS) and negative ideal solution (NIS) by using TOPSIS method. Then according to the size of relative closeness the rank of vendor can be confirmed. Finally, this paper certificated the validity of the evaluating method and system via real example analysis.

Keywords: Entropy Weight, TOPSIS, Relative Closeness

1. INTRODUCTION

Under the background of globalization, the market competition is not means the competition among the enterprises anymore but the supply chain. The vendor selection in supply chain is the key for improving competitive strengths of the whole supply chain. It has been proved that rational vendor selection can realize the decrease of enterprise cost, increase of flexibility and improvement of competitive strengths directly [1]. At present, the vendor selection study is very popular in the world and it mainly includes two parts: the study of attribute system for vendor selection and the study of approaches for vendor evaluation. Dickson (1966) [2] induced 50 factors from the literatures which studied the purchasing problem. After his study on this question he finally presented 23 evaluating rules on vendor selection. He emphasized that quality, delivery and past performance should become the main considerable factors; Weber (1991) [3] summarized the study achievements on vendor selection criteria which started at the publication of Dickson's thesis. After the analysis on 74 thesis about vendor selection, he found that most articles mentioned about price, delivery period, quality and capability standard; Johnson(1995)[4] believed that time (T), quality(Q), cost (C) and service(S) are the key factors for getting success in the process of choosing vendor. In this study he adopted the enterprise strengths and weakness evaluating method. In side of China, based on survey about the auto components vendor, Kan Shuyong and Chen Rongqiu (1988) [5] claimed the vendor evaluation should rely on the following attributes: quality, delivery period, batch flexibility, the balance between delivery period and price, the balance between price and batch, variety etc. In the book of "Supply Chain Management " (2000), Ma Shihua, Lin Yong and Chen Zhixiang published a integrated evaluating attribute system for copartners selection under the supply chain management circumstance and generalized the 4 main factors which can affect the copartner selection: enterprise outstanding achievement, operation structure and throughput, quality system and enterprise environment[1]; Qian Bibo et al (2000)[6] pointed out that time (T), quality (Q), cost (C), service(S) are key factors for getting success when the enterprise evaluate the vendors; Ma Lijuan (2002) [7] proposed that the vendor selection criteria is composed by 9 evaluating attribute: product quality, product price, after service, distance, technological level, supply capability, economic revenue, delivery and market influence.

About the evaluating method, American scholar T.L.Saaty(1973)[8] presented Analytical Hierarchical Process(AHP); Gaballa(1974) [9] applied linear programming in vendor selection problem for the first time; Schinal (1980) [10] put forward Data Envelopment Analysis model (DEA) to deal with the vendor selection problem; Fillip Roodhooft and JozefJoings(1996)[11] brought forward Activity Based Costing(ABC) to select and evaluate the vendors; Ghodsypour and Brien(1998) [12] studied the decision-making for vendor selection, with an integrated analytic hierarchy process(AHP) and linear programming(LP); Isao Shiromaru et al (2000) [13] used Fuzzy Programming Approach for dealing with fuzzy goals problems in the process of Vendor Selection and used Inheritance Arithmetic to request the solution; JoeZhu (2004)[14] simplified the Data Envelopment Analysis approach (DEA) via swapper twain stages game model and conducted an efficiency interior to evaluate the vendors; Manoj Kumar (2004) [15] utilized fuzzy optimization theory on vendor selection. In side of China, Wang Ying et al (2002) [16] put forward vendor evaluating methods based on Euclid Norm; Ma Shihua et al (2002) [17] published a Grey Relating Model to settle the evaluation on the weight of attribute; Wang Jiashun and Wang Tianmiao et al(2001)[18] presented a vendor evaluating model which is based on a fuzzy analytical hierarchical process(AHP) method. Bai Rong (2006) [19] claimed the vendor selection evaluating approach based on TOPSIS.

Based on the study above, the author claim that information entropy approach can be used to confirm the objective weight of each attribute then try to evaluate the vendors by using TOPSIS method.

2. THE ATTRIBUTE SYSTEM FOR VENDOR EVALUATION

Based on the study above and the real situation of supply chain management, the author claimed that technologic level, service level, performance capability and enterprise circumstance the four aspects should become the criteria for supplier selection. In this way, the actual capability and development potential of the evaluated enterprise can be reflected objectively and truly; the copartner relationship in supply chain can be maintained and the two cooperators can get the extended strategic benefit. The detailed attribute system shows table-1.

Table 1. Vendor Evaluating Attribute System

Content	Attribute
A1 Technologic Level	B1 product research capability B2 product quality B3 product reliability B4 production system quality authentication level
A2 Service Level	B5 price B6 ahead of schedule B7 credit standing B8 after service satisfaction
A3 Performance Capability	B9 financial state B10 supply capability B11 cooperation capability B12 performance management capability B13 development capability
A4 Enterprise Circumstance	B14 political and law circumstance B15 economic and technological circumstance B16 natural and geographic circumstance B17 social and cultural circumstance B18 compatibility of enterprise culture B19 compatibility of management system

3. TOPSIS APPROACH BASED ON ENTROPY WEIGHT

TOPSIS is a useful technology for order performance on evaluated object by using PIS and NIS of multi-attribute problem. The ideal solution is an assumed best solution (V+). It's each index value reach the best point among the selection. The negative ideal solution is another assumed worst solution (V-). It's each index value reach the worst point among the selection. Compare each selection with the V+ and V- in the original selection group X, then use the distance information as the criteria to rank the part selections (m) in the selection group (X).

3.1 Confirm the Weight of Evaluating Attribute

Assume that there are m evaluated objects (vendors) and n evaluating attributes. All attribute values of each vendor evaluation can be used to construct a matrix X, x_{ij} indicates the j th attribute value of the i th vendor.

1) Normalize the data. There are many methods in deal with the normalization on the data, we choose the following way:

$$r_{ij} = x_{ij} / \sqrt{\sum_{i=1}^{m} x_{ij}^2} \quad (1 \le i \le m, 1 \le j \le n) \qquad (1)$$

2) Count the decision-making information entropy value There are many methods to confirm the weight of each attribute such as Delphi Approach and Analytical Hierarchical Process (AHP) et al. All these methods can not avoid subjectivity in some degree when using them to confirm the weight of evaluated attributes. In this paper, the author adopted the concept of information entropy to confirm the weight of evaluating attribute which can avoid the influence of subjective factors efficiently. Entropy is a measurement used to appraise the uncertainty based on probability theory. It indicates that the data are more dispersive the data are more uncertain. The decision-making information of each attribute can be indicated by their entropy value Ej.

$$E_j = -K \sum_{i=1}^{m} r_{ij} \ln r_{ij} \quad (1 \le i \le m, 1 \le j \le n) \qquad (2)$$

$$0\,1n0 \equiv 0$$

In the Eq.(2), m is the amount of evaluated object. $K = 1 / \ln m$.

3) Count the differentia degree of attributes:

$$G_j = 1 - E_j (1 \le j \le n) \qquad (3)$$

4) Count the entropy weight A

$$a_j = G_j / \sum_{j=1}^{n} G_j (1 \le j \le n) \qquad (4)$$

3.2 The Evaluating System of TOPSIS

1) Construct weighted normalized matrix. Weighted the normalized data based on entropy weight in order to construct weighted normalized matrix.

$$V = (v_{ij})_{m \times n} = \begin{bmatrix} a_1 r_{11} & a_2 r_{12} & \cdots & a_n r_{1n} \\ a_1 r_{21} & a_2 r_{22} & \cdots & a_n r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_1 r_{m1} & a_2 r_{m2} & \cdots & a_n r_{mn} \end{bmatrix} \qquad (5)$$

2) Confirming the PIS and NIS of the evaluated object, they are

$$V^+ = \left\{ (\max_i v_{ij} \,|\, j \in J_1), (\min_i v_{ij} \,|\, j \in J_2) \,|\, i = 1, 2, \cdots, m \right\} \qquad (6)$$

$$V^- = \left\{ (\min_i v_{ij} \,|\, j \in J_1), (\max_i v_{ij} \,|\, j \in J_2) \,|\, i = 1, 2, \cdots, m \right\} \qquad (7)$$

In the Eqs.(6) and (7), J_1 is the beneficial attribute group; J_2 is cost attribute group.

3) Counting distance. The distance of the evaluated object with PIS and NIS are

$$d_i^+ = \left[\sum_{j=1}^{n} (v_{ij} - v_j^+)^2 \right]^{1/2},$$
$$d_i^- = \left[\sum_{j=1}^{n} (v_{ij} - v_j^-)^2 \right]^{1/2}. \qquad (8)$$
$$(i = 1, 2, \cdots, m)$$

4) Confirming the relative closeness. Evaluating the relativecloseness between the object and ideal solution is

$$C_i = \frac{d_i^-}{d_i^+ + d_i^-} . (i = 1, 2, \cdots, m) \qquad (9)$$

Based on the size of the relative closeness the strengths and weakness of the evaluated objects can be ranked. Then the most suitable vendor can be found.

4. EMPIRICAL APPLICATION

The data and materials of this example come from literature [7]. The author, in this example, used 9 attributes of the attribute system of the literature. In this case, a core enterprise needs to choose one partner among 6 element suppliers. It can use product quality, product price, after service, distance, technological level, supply capability, economic revenue, delivery and market influence as the evaluating standards. Among the 9 attributes, the 6 attributes like product quality, technological level, supply capability, economic revenue, delivery and market influence are beneficial attribute, the bigger the better; product price, after service and distance are cost attribute, the smaller the better. According the attribute system, all evaluating values are in the table 2.

Table 2. Attribute value of vendor evaluation for a enterprise

Vendor	Product Quality	Product Price (￥:Yuan)	After Service (Hour)	Distance (km)	Technological Level	Supply Capability (Piece)	Economic Revenue	Delivery	Market Influence
1	0.79	335	3. 2	15	0.12	230	0.12	0.83	0.13
2	0.91	268	1. 4	37	0.25	130	0.08	0.96	0.15
3	0.99	304	1. 9	22	0.09	200	0.14	0.99	0.20
4	0.97	270	2. 0	19	0.33	180	0.09	0.87	0.21
5	0.86	310	0. 8	26	0.20	150	0.15	0.80	0.12
6	0.95	303	2. 7	8	0.19	170	0.17	0.91	0.19

Based on table 2, the original evaluating matrix is

$$X = \begin{bmatrix} 0.79 & 335 & 3.2 & 15 & 0.12 & 230 & 0.12 & 0.83 & 0.13 \\ 0.91 & 268 & 1.4 & 37 & 0.25 & 130 & 0.08 & 0.96 & 0.15 \\ 0.99 & 304 & 1.9 & 22 & 0.09 & 200 & 0.14 & 0.99 & 0.20 \\ 0.97 & 270 & 2.0 & 19 & 0.33 & 180 & 0.09 & 0.87 & 0.21 \\ 0.86 & 310 & 0.8 & 26 & 0.20 & 150 & 0.15 & 0.80 & 0.12 \\ 0.95 & 303 & 2.7 & 8 & 0.19 & 170 & 0.17 & 0.91 & 0.19 \end{bmatrix}$$

Normalize the original matrix based on Eq. (1), get a normalized matrix

$$R = \begin{bmatrix} 0.144 & 0.187 & 0.267 & 0.118 & 0.102 & 0.217 & 0.160 & 0.155 & 0.130 \\ 0.166 & 0.150 & 0.117 & 0.291 & 0.212 & 0.123 & 0.107 & 0.179 & 0.150 \\ 0.181 & 0.170 & 0.158 & 0.173 & 0.076 & 0.189 & 0.187 & 0.185 & 0.200 \\ 0.177 & 0.151 & 0.167 & 0.150 & 0.280 & 0.170 & 0.120 & 0.162 & 0.210 \\ 0.157 & 0.173 & 0.067 & 0.205 & 0.170 & 0.142 & 0.200 & 0.149 & 0.120 \\ 0.174 & 0.169 & 0.225 & 0.063 & 0.161 & 0.160 & 0.227 & 0.170 & 0.190 \end{bmatrix}$$

Based on the Eqs. (2), (3) and (4), the entropy weights of each attribute are
A={0.1368, 0.1367, 0.0785, 0.0707, 0.0781, 0.1263, 0.1137, 0.1369, 0.1222}
Based on Eq. (5), construct weighted normalized matrix.

$$V = \begin{bmatrix} 0.0483 & 0.0625 & 0.0477 & 0.0188 & 0.0180 & 0.0660 & 0.0432 & 0.0518 & 0.0381 \\ 0.0556 & 0.0500 & 0.0209 & 0.0464 & 0.0376 & 0.0373 & 0.0288 & 0.0599 & 0.0439 \\ 0.0605 & 0.0567 & 0.0283 & 0.0276 & 0.0135 & 0.0574 & 0.0504 & 0.0618 & 0.0586 \\ 0.0593 & 0.0504 & 0.0298 & 0.0238 & 0.0496 & 0.0517 & 0.0324 & 0.0543 & 0.0615 \\ 0.0525 & 0.0578 & 0.0119 & 0.0376 & 0.0300 & 0.0431 & 0.0540 & 0.0499 & 0.0352 \\ 0.0580 & 0.0565 & 0.0403 & 0.0100 & 0.0285 & 0.0488 & 0.0612 & 0.0568 & 0.0557 \end{bmatrix}$$

Confirm the ideal and negative ideal solutions V+ (PIS) and V − (NIS) for evaluated objects.

$$V^{+} = \{0.0605, 0.0500, 0.0119, 0.0100, 0.0496, 0.0660, 0.0612, 0.0618, 0.0615\}$$

$$V^{-} = \{0.0483, 0.0625, 0.0477, 0.0464, 0.0135, 0.0373, 0.0288, 0.0499, 0.0352\}$$

Based on Eq. (8), the distance with PIS and NIS are

$$d^{+} = \{0.0603, 0.0613, 0.0461, 0.0400, 0.0493, 0.0406\}$$

$$d^{-} = \{0.0427, 0.0410, 0.0497, 0.0577, 0.0495, 0.0582\}$$

Based on Eq. (9), the relative closeness will be

$C = \{0.4148\ ,0.4011\ ,0.5191\ ,0.5904\ ,0.5011\ ,0.5888\ \}$

Hereby, we can get the ranking of the 6 suppliers is: 4, 6, 3, 5, 1, 2.

5. CONCLUSIONS

Vendor selection is the foundation of the supply chain cooperation. This paper constructed the vendor selection model for supply chain by using TOPSIS method which is based on the entropy weight. To get the objective weight of evaluated attributes by using information entropy. TOPSIS method in this way can avoid the subjectivity which comes from the influence of the weight of low level multi-attribute and confirms the objectivity and practice of the evaluation. Then the ranking of vendors can be confirmed based on the relative closeness. The author carried out a model calculation by using the data of literature [7] and got the similar evaluating conclusion with the original literature. It can be certificated via the application in this real case that TOPSIS method is very convenient for operation and easy to generalize in practice.

REFERENCES

[1] Shihua Ma, Yong Lin and Zhixiang Chen. *Supply Chain Management,* Mechanical Industry Press, Beijing. 2000.

[2] Dickson G.W,"An analysis of vendor selection systems and decisionsm," in J*ournal of Purchasing*, 1966. 2(1):5-17

[3] C.A. Weber, J.R. Current, W.C. Benton,"Vendor selection criteria and methods,"in *European Journal of Operational Research*, 1991:50.

[4] Johnson M. Partner,"Selection in the Agile Environment , Creating the Agile Organization: Models, Metrics and Pilot[C],"in *4th Annual Conference Proceedings*,1995.

[5] Shuyong Kan and Rongqiu Chen,"The Relationship of Manufacturers and Suppliers Under JIT,"in *Journal of Management Engineering*,Mar.1998,pp46-51 .

[6] Bibo Qian,"Evaluating System Study for Prompt and Suppositional Enterprise Partner Selection,"in Chinese Mechanical Engineering,Apr.2002,pp397-401.

[7] Lijuanm Ma,"The Primary Study of Supplier Selection Based on Supply Chain Management,"in Industrial Engineering Management,Jun.2002,pp23-25.

[8] T. L. Saaty & J.M. Alexander,Thinking With Models: Mathematical Models in the physical Biological and Social Sciences, Chapter8, Pergamon Press, London, 1981

[9] A.A Gaballa. Minimum cost allocation of tenders, Operational Research Quarterly,1974: 25

[10] Timmerman. E,"An approach to vendor performance evaluation,"in Purchasing and Supply Management, 1986 (1).

[11] Filip.Roodhooft, & Jozef Konings,"Vendor selection and evaluation: an activity Based Costing Approach," in European Journal of Operation Research,1996,pp96.

[12] S. H. Ghodsypour and C. O'Brien,"A decision support system or supplier selection using an integrated analytic hierarchy process and linear programming," in International journal of Production Economics, 1998, pp56.

[13] Isao Shiromaru,"A fuzzy satisfying method for electric power plant coal purchase using genetic algorithms," in Operational Research,2000,pp126.

[14] Joe Zhu,"A Buyer-Seller Game Model for Selection and Negotiation of Purchasing Bids: Extension and New Models,"in European Journal of Operational Research, 2004, pp154.

[15] Manoj Kumar., Prem Vrat & Shankar,R,"A Fuzzy Programming Approach for Vendor Selection Problem in Supply Chain,"in Computer & Industrial Engineering, 2004 (46), pp69-85.

[16] Ying Wang, Sun Linyan and Zhao Yimeng,"Vendor Evaluating Methods Based on Euclid Norm," in Systematical Engineering,Jan.2002,pp46-50.

[17] Shihua Ma and Wang Xubin,"A Method of Confriming the weight of attributes for Supplier Evaluation," in Industrial Engineering Management,Jun.2002,pp5-8.

[18] Jiashun Wang, Wang Tianmiao et al,"A study of Vendor Evaluation Model Based on Fuzzy Analytic Hierarchy Process,"in Microelectronics and Computer, Feb.2001, pp59-64.

[19] Rong Bai, Cui Bingmou,"An Application of TOPSIS for Suppliers Selection," in Railway Transport and Economy,Sep.2006.

[20] Chaoyuan Yu. Theories and Methods for Decision Making,Science Press, Bejing. 200.

Networked RFID and Its Impact on the Future Logistics

Qian Huang, Lisheng Qiu
Economics and Management School, Wuhan University,
Wuhan, Hubei, 430072, China
Email: Huangqian_hq@sina.com,rgb-qiu@126.com

ABSTRACT

Radio frequency identification (RFID) used in logistics information service management has attracted considerable media attention in recent years. This paper explores a variety of issues concerning RFID technology. It includes an overview of research and development, technology and current and future applications. The framework of networked RFID used in logistics is our concerns and a systematic architecture and communication protocol are proposed in our paper. Furthermore, the service opportunities and potential challenge on future logistics which brought by the networked RFID are discussed extensively in this paper.

Keywords: Radio Frequency Identification Networked RFID, Logistics, Impact

1. INTRODUCTION

Technology continues to drive logistics [1]. Barcode technology was the first step toward reaching the ultimate goal of item-level, real-time supply chain visibility [2]. Obviously, barcode scanning improved the accuracy of the data that was collected, i.e., it eliminated human error and provided greater automation. It was a revolution for the logistics world. Barcode scanning is now a very reliable technology that continues to demonstrate a lot of potential with the new 2-D barcodes [3].

Radio Frequency Identification (RFID) is the latest technology that supplies chain professionals are using to reach the goal of item-level, real-time supply chain visibility[4]. RFID is a technology which allows remote interrogation of objects using radio waves to read data from RFID tags which are distant from an RFID reader. RFID has several advantages over manual scanning using optical barcodes, since many tagged items (or embedded sub-components of composite product) could be simultaneously identified in an automatic manner, very quickly and without the need of line-of-sight to each item [5].

The adoption of RFID throughout the supply chain can provide two main advantages to the logistics. Firstly, it improves the visibility of inventory and demand level predictions throughout the supply-chain. Secondly, via track and trace processes, it provides additional security through shrinkage avoidance, grey market diversion and black market counterfeit cloning [6]. Improved visibility and security of the supply chain will benefit the overall business, directly by more optimized and efficient production and distribution flows, and indirectly through improved brand protection and certainty of produce [7]. RFID has the potential to create a truly adaptive supply chain, enabling all aspects of the business cycle (production, storage, distribution, retail and returns) to be monitored in real time, optimizing for present conditions and making predictive changes based on expected demands.

Today RFID is a generic term for technologies that use radio waves to automatically identify people or objects. There are several methods of identification, the most common of which is to associate the RFID tag unique identifier with an object or a person. An RFID system (Fig.1) will typically comprise the following:

(1) an RFID device (tag);
(2) a tag reader with an antenna and transceiver; and
(3) a host system or connection to an enterprise system.

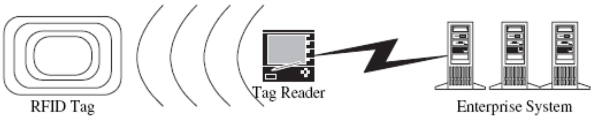


Fig.1. RFID System Structure

Although cheap, tiny, plentiful radio-frequency identification tags will make it possible to tag almost everything, RFID itself is merely an input device for connecting physical objects with computer systems. For many years, the major stumbling block to widespread deployment of RFID was the reliable networked technology, as perhaps the most important implication of RFID technology today relates to its use within bigger information systems connected to the Internet [8]. The identifiers retrieved from a tag can be used to query or update online databases that hold information about objects and people alike. As networked RFID used in logistics is our concerns, the service opportunities and potential challenge on future logistics which brought by the networked RFID will be discussed extensively in this paper.

The remaining sections of this paper are organized as follows. Section 2 describes the framework of networked RFID-based system. Section 3 discusses the service opportunities of future logistics using networked RFID. Section 4 presents the potential challenge and the conclusion of this paper is in Section 5.

2. FRAMEWORK OF NETWORKED RFID

2.1 Systemic Architecture

In logistics flow, according to the transportation of materials, a large amount of data is transferred and shared. It is important to integrate and control large amount of logistics information according to the standard information management framework. For the effective management of a large amount of logistics information such as product descriptions, transports of goods, and packing of products, networked RFID system are required. A systemic architecture for networked RFID used in logistics management is showed in Fig.1.

At the lower (passive) end of RFID technology the system simply provide a tag that can remotely identify an object by returning an ID when interrogated over short ranges. As RFID systems are introduced and find acceptance in business and other environments (situations), the functionality provided by these low cost tags will be increasingly seen as insufficient as new applications are developed. There is likely to be a natural progression for RFID that includes the widespread incorporation of sensor functionality. Such devices will be able to make measurements concerning their surroundings and physical location about such variables as pressure, temperature, flow rate, speed, vibrations etc. They can be networked either through RFID technologies or through other wireless communications systems and these developments are often referred to as a localized wireless sensor network (WSN). In these types of networked RFID, RFID-enabled objects can be precisely located in space and time and they will become protagonists of a documented process.

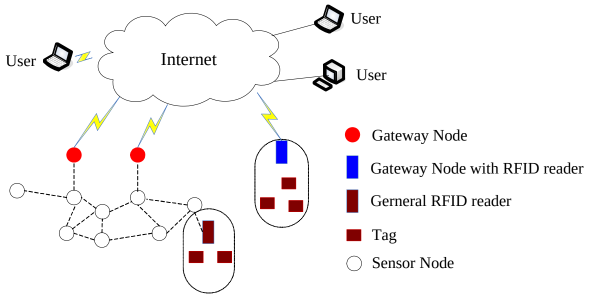


Fig.2. Systemic Architecture for Networked RFID

To implement our networked RFID architecture, we adopted a loose hierarchy of three distinct computing platform types (shown in Fig.2), which match the computation resources of readily available computing and sensing components. These are broadly classified (from top to bottom) as user terminals, gateway nodes and sensor nodes. Inter-nodal messaging is performed using a diversity of wired and wireless communication methods.

Senor nodes and tag perform the sensing, initial processing and communicate using ad-hoc networks to provide flexibility of deployment and nodal mobility. Such platforms include miniature wireless sensor nodes, for example the Intel/Berkeley Motes and the HP Labs developed Locus system of location aware sensor nodes. These are primarily low power, resource constrained and battery powered devices. Gate nodes are primarily designed to operate at fixed locations as they require wired connections, either for power and/or bandwidth connectivity. These may include part RFID readers. Such powered resources have superior computational capabilities and so often function as gateways– collecting and aggregating data from adjacent senor nodes before passing this to higher level services. Computing Nodes host the software components that work alongside of other IT services. The computing nodes do not perform any physical sensing and thus their location dependence is significantly reduced. This allows their virtualization, resource re-allocation and implementation using industry standard IT building blocks, for example, industry standard servers and network routers.

2.2 Communication Protocol

These RFID-based sensors will need to communicate in order to participate in the network of things. Some protocols currently proposed or developed include ZigBee, Near Field Communication Technologies (NFC), Bluetooth and WIFI – all systems that offer local and personal area networks.

Network communications follow a similar hierarchical divide. The upper layer communication use generic standardized wired IP based solutions and wireless LAN (e.g. 802.11b/g) where appropriate. The lower layer communications are predominantly wireless, however standards (e.g. ZigBee, wireless USB) in the sensor networking space are still evolving.

ZigBee is a complex new standard that continues to evolve at a rapid pace, which originated in 1998 when Motorola began to work on this type of meshing networking. This standard specifies a low-power, low-cost, and low-complexity wireless technology for personal-area and device-to-device wireless networking such as home and building automation, consumer electronics, industrial control, and medical sensor applications.

Application Layer	
Network Layer	
IEEE802.2 LLC	IEEE802.154 LLC
IEEE802.15.4MAC	
2.4GHz PHY	868/915MHz PHY

Fig.3. ZigBee Protocol Stack

As depicted in Fig.3, ZigBee stack architecture is composed of the application (APL) layer, the network (NWK) layer, the medium access control (MAC) layer and the physical (PHY) sub layer, which is based on the standard Open System Interconnection (OSI) model. The IEEE802.15.4 standard defines the lower layers including the medium access control (MAC) layer, and the physical (PHY) layer while the ZigBee Alliance establish the Network layer and the framework for the application layer.

The APL layer consists of the application support (APS) sub-layer, the application framework (AF) layer, the ZigBee device object (ZDO) and the manufacturer-defined Application objects. The APS layer is responsible for the maintenance of binding tables that match two devices according to their services and needs. The function of the ZDO layer is to define the role of devices within the network, start/respond to binding requests and build a secure relationship between devices.

The responsibilities of the NWK layer involve adding or deleting a device to/from the network and starting of a new network (e.g. assigning addresses to new devices).In addition, the discovery of one-hop neighbors and the storing of pertinent neighbor information are done. The NWK layer is also responsible for the discovery and maintenance of routes between devices.

3. SERVICE OPPTUNITIES

RFID is a new space, yet it embodies many of the attributes of our existing business. While the markets are still developing, we have an opportunity to innovate technologies and business processes. In 1984 Wal-Mart mandated the use of barcodes and in the process transformed the industry. The same company is behind the push for a similar transformation through their adoption of RFID.

The adoption of RFID based supply chain solutions in the short and medium term, and throughout different business sectors, creates a suite of service areas. Among these offerings opportunities exist for:

Infrastructure Deployment: Installation and testing of RFID readers, sensing infrastructure and IT services within ports, factories and distribution centers.

Data Management Services: Implementation and management of EPC-IS information services for end-product and information sharing between trading partners and multiple execution environments of complex supply chain systems.

Brand Protection Services: Aggregation of RFID data from track-and-trace process. Such services monitor the pedigree of a product, certifying its authenticity and alerting the company to product diversions or shrinkage.

Enterprise Integration: The transformation of a traditional supply-chain into a real-time adaptive supply-chain. To enable this vision, the entire cycle of data capture, data management, analysis and response needs to be integrated into the business processes of the enterprise.

4. FUTURE CHALLENGE

Despite the promising applications of RFID in SCM, a number of challenges have hampered the adoption of RFID. This section will address the attitudes needed to face these challenges which are listed below:

Management Commitment - The most significant challenge to implementation is the commitment of management to adopt new technology and have appropriate expectations of RFID capabilities. Without executive sponsorship implementation will be unlikely to succeed. “Early Adopter” and “Fast Follower” corporate cultures are much more likely to adopt this new technology into their business environment than “watch-and-wait”.

Customer Schedules - Compliance mandates put in place by Wal-Mart and the US Department of Defense have provided a strong incentive to implement RFID. Even those companies not currently required by their customers to actively implement the technology will likely lose customers if they do not start to actively assess the technology. Customer schedules have also led to recent significant increasing familiarity and experience with the technology and the business case for RFID integration.

International Standards - A key challenge is the continually evolving standards in technology, application, data, conformance, firmware changes, and tracking methods. In addition, different companies often use different standards making cooperation between suppliers and manufacturers difficult. There are three major advantages of developing international standards for RFID systems. First of all, a common RFID standard will ensure interoperability among tags and readers manufactured by different vendors and allow seamless interoperation across national boundaries. Secondly, due to compatibility and exchangeability, there will be high demand on RFID components and equipment, as the result of which the cost will be cut down. Finally, an internationally accepted RFID standard will facilitate the growth of the worldwide RFID market.

Currently, there are two major organizations working to develop international standards for RFID technologies in the UHF spectrum. These two organizations are EPCglobal and International Standards Organization (ISO). EPCglobal released its EPC class 1 G2 protocol (EPCglobal Inc., 2005) for the UHF band at the end of 2004, and the ISO released its 18000-6 in August of 2004. Both standards are still evolving and are not completely compatible with each other (EPCglobal Press Release, 2005). A unified, globally interoperable RFID standard is ideal to realize the full benefits of RFID applications. The lack of a complete and unified RFID standard has caused many companies to hesitate in adopting the RFID systems; these companies were afraid of making a commitment that might render their entire RFID system investment worthless in the future.

In addition, regulations on radio spectrum allocation for RFID use are not unified among nations. A large portion of the UHF spectrum has already been auctioned to cellular phone service providers for high license fees by a few countries. It would be difficult, if not impossible, to buy that portion of spectrum back for RFID use. This adds complexity to the adoption of RFID for global supply chain management applications where tagged goods must often travel across borders.

Technology - Continuously and rapidly evolving technology presents unique implementation challenges to integrating hardware, software and infrastructure due to upgrade management requirements. Further, read ranges of current tags are still short, and read-logic ability to distinguish between different pallets is still an issue, and there are operational environment limitations on read accuracy such as liquids, metals and electro-static devices which can distort, absorb, scatter or reflect signals.

Availability of Resources - Resource availability is also limited due to the lack of sufficiently trained, skilled personnel which is complicated by the aforementioned rapidly evolving standards and technologies. There is also a lack of public domain reference case studies that comprehensively document failures and lessons learned. The shortage of existing skilled resources and lack of comprehensive, accessible information has cost implications for training and presents potential implementation problems.

Security - For certain implementations, illicit tracking of RFID tags presents problems. This is particularly relevant to military installations but security challenges are relevant also to corporations and individuals. For example, scanning and cloning of RFID tags can potentially provide undesired access to important facilities or be used for payment in commercial transactions.

Change Management - RFID implementation poses challenges of managing change associated with integrating RFID and of reengineering work process. This requires strong management commitment and support. For example “Slap and ship” often does not provide a return on investment as the real benefit typically comes from back-end integration. Because most back-end systems are not designed for the level of detail that RFID provides, enterprise applications currently often cannot specifically retain meaningful information.

Cost –There are two major cost elements constituting the tag cost; one is the chip cost and the other is the assembly cost. Chip cost is related to the die size and fabrication yield. For example, one six-inch wafer can produce more than 15–20 thousand chips, but the wafer, mask, and chip preparation costs are shared among only the good chips. An RFID chip consists of analog logic, digital logic, and memory circuitry; thus, it is a challenge to keep error rates low in the chip fabrication process.

The chip cost can decrease if the chip order volume becomes large.

Because RFID chips are very small in size (0.4–1.0mm2), and the antenna inlay material, such as PET, is very soft, it is also a challenge to perform RFID chip assembly at a very high through put with a high accuracy.

As a result, the goal of a five-cent tag will require efforts from every participant in the value chain before it can be realized. Unless customers have an urgent need that can be solved by RFID, they tend to wait until RFID tag prices drop and standards are sorted out before making any investments. The high cost of the RFID tag is a major reason why the market penetration of RFID remains stagnant.

5. CONCLUSIONS

Networked RFID can be a viable technology to use in tracking assets in real time and automatically access an inventory database. It is flexible enough to be applied in many application domains for real-time tracking and automation. For the application of logistics tracking and management, it saves time, money, and the hassles of finding and tracking inventory. Networked RFID is a very promising and rapidly developing technology with a number of significant advantages over the traditional optically scanned barcode systems and has the potential to replace them in the near future. Through our study, we can make conclusion that there are opportunities for networked RFID technology to provide significant benefits in logistic management, which well beyond the automation oriented advantages such as labor savings. However, our results do show that networked RFID technology is still becoming mature and the industry is still young. Its full impact is not yet foreseeable and there is still much challenge for the application.

REFERENCES

[1] H. Petri, S. Bulcsu, “Logistics information systems: An analysis of software solutions for supply chain co-ordination”, *Industrial Management and Data Systems*, Vol. 105, No.1, 2005, pp. 5~18.

[2] N. C. Wu, M. A. Nystrom, T. R. Lin et al, “Challenges to global RFID adoption”, *Journal of Elsvier Technovation*, Vol. 26, No.1, 2006, pp.1317~1323.

[3] D. Smith, “Exploring radio frequency identification technology and its impact business systems”, *Journal of Information Management and Computer Security*, Vol.13, No.1, 2005, pp.16~28.

[4] S. Kumar, M. Thomas, S. Pauly et al, “Impact of radio frequency identification technology on manufacturing and logistics: Challenges and issues”, *International Journal of Manufacturing Technology and Management*, Vol.10, No.1, 2007, pp.57~70.

[5] K. M. Penttila, D. W. Daniel et al, “Radio frequency identification systems in supply chain management”, *International Journal of Robotics and Automation*, Vol.19, No.3, 2004, pp.143~151.

[6] W. Falinski, “An overview of Radio Frequency Identification (RFID) tags technology”, *Proceedings of SPIE - The International Society for Optical Engineering*, V 6347, *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments* 2006, Oct.12, 2006, pp.634720.

[7] J. Kim, D. Choi, I. Kim, et al, “Product Authentication Service of Consumer's mobile RFID Device”, *2006 IEEE Tenth International Symposium on Consumer Electronics*, 2006., ISCE '06, June 2006, pp.1~6.

[8] W. Zhu, D. Wang, H. Sheng, “Mobile RFID technology for improving m-commerce”, *IEEE International Conference on e-Business Engineering, 2005*., ICEBE 2005, 2005, pp. 118~125.

Hierarchical Task Analysis of Individual Online Shopping Method: Laptop-Shopping As a Case Study

Yue Guo
Ph.D Candidate in School of Management, Wuhan University of Technology, P.R. China 430070
Jiangxi Vocational College of Financial & Economics, P.R. China 332000
Email: gy@jxvc.jx.cn

ABSTRACT

Internet shopping (or e-shopping) is emerging as a new method of shopping mode with its requirement of computer access and use, it is interesting to know whether consumers associate e-shopping with factors like value, convenience, performance, trustworthiness, overall satisfaction, and ease of use. The aim of this analysis is produce models of e-shopping, compare those models to existing shopping websites, and design an improved hierarchial task model for evaluation.

Keywords: Hierarchical Task analysis, Online Shopping, Website Interface.

1. INTRODUCTION

Traditional businesses ‘shop’ through a leasing contract with a third party, a bulk purchase arrangement with a manufacturer, or by outsourcing the task to a third party. As it would be impractical to try and align these methods of shopping with those available to the individual this study will focus on the individual’s task of shopping ruther than a business’. As a case sduty, here will examine the individual’s experience of online shopping for a laptop computer.

This research takes a pragmatic approach that builds upon accepted definitions of knowledge, e.g. (Nonaka, 1994) and has been used elsewhere in examining the deployment of knowledge via Internet information systems (Saward, Shah and Hall, 2005). The often quoted key “benefits” of e-commerce, e.g. (Alba *et al*, 1997; Timmers, 1999), are: price; availability of service; availability of a broad range of goods and services; and the ability to create personalized services and products. Documented user surveys (GVU, 1999) highlight similar motivators and corresponding inhibitor.

A key insight is that these “benefits” are really “features”, such as low price, availability of service, availability of products, and personalization (Saward, Hall and Barker, 2004). They are characteristics of the Internet as a shopping medium rather than “benefits” for the customer. More importantly, since these features are shared by other types of e-commerce, it is necessary to translate the features of e-commerce into competitive advantages that deliver consumer benefits (Saward and Hall, 2005). The focus of this model is on the core decision-making process and assumes that problems have already been recognized or needs established(Kotler, 2000). It also excludes post-purchase behavior in evaluating the shopping experience, or adjustments to needs and perceptions on the basis of how the product delivers the anticipated benefits or value.

The goal analysis Kotler (2000) uses to generate the accessibility requirements is restricted to the “functional” aspects of shopping and focuses on the knowledge required to achieve a ‘rational’ purchase. It does not consider all the shopping tasks, or other high-level goals such as shopping for leisure or emotional validation. In the present study, Human computer Interface(HCI) considerations such as shopping for recreational activities, are taken into account since these features can be seen in good computer interface design, e.g. Garden Escape (Fleming, 1998) and HCI (Wilson and Johnson, 1997).

2. TASK ANALYSIS MODEL

2.1 Task Knowledge Structures (TKS)

In this analysis used the TKS framework (Johnson *et al*, 1995, Wild and Johnson, 2004) to provide a detailed analysis of laptop shopping, as it covers actions and objects along with task goals. The resultant model (see Fig.1., table1) has been drawn using all of the data collected, and again shows the particular importance of knowledge in this task. The complex nature of laptops means that if a shopper does not have a proper understanding of laptops, and their own needs, they may buy an inappropriate laptop thus failing in their primary goal. TKS allows us to see the critical nature of knowledge by defining it as a ‘central task feature’ (Wild, Johnson & Johnson, 2003). This is necessary as the knowledge gained by research is key in making all of the decisions that lead to the final purchase of the laptop.

Reassurance is also an important goal since laptop shoppers are spending a large sum of money on a complicated piece of equipment; they will look for reassurance at almost every opportunity to ensure that they are making the right decisions. It is for precisely this reason that the reassurance goal appears separate to the other goals on the goal structure, as it can be a goal for a shopper at any point in the structure. Reassurance can also be given by furthering a shopper’s own knowledge, be it the details of a particular model’s specification advice gleaned from a laptop magazine. It is clear from this model that the goal of knowledge and it is associated tasks will need further analysis, as it is key not only to the primary goal but also to the universal goal of reassurance.

2.2 Hierarchical Task Analysis (HTA)

Drawing upon the original data and the completed TSK model a general HTA model (Fig.2.) was drawn up using the methods described by Faulkner (2000) and Preece *et al* (2002). HTA provides more detail about the individual tasks and sub-tasks involved in fulfilling the goals outlined in the TSK model. The model also gives an indication of the order in which these tasks are likely to be carried out. Reassurance and knowledge are still key to the shopping task, as they must be carried out before either a decision on which laptop to buy or the actual purchase can take place. From this we can see that further analysis of the research task needs to be undertaken, particularly since it is the most complex task in the model.

3. RESEARCH METHOD

In order to carry out a task analysis of shopping for a laptop, I selected semi-structured interviews as the primary research method. Observation was used first as it allows the researcher to study a task quickly and informally if desired (Preece *et al*, 2002). This was especially important to this study, as time was a constraint on the amount and depth of research that could be carried out and analyzed. Although it is readily acknowledged that observation of a task can alter the way in which it is carried out and documented, the so-called *Hawthorne Effect* (Faulkner, 2000), observation often reveals details that would be missed if the observer had to rely solely on second hand accounts of the task rather than witnessing the task first hand.

In addition to observation I also elected to use semi-structured interviewing. This technique enabled me to compare individual shoppers, as well as groups of shoppers. Semi-structured interviews also enabled the documentation of any unexpected data, something that structured interview questionnaires cannot allow for (Faulkner, 2000; Payne, 1995; Preece *et al*, 2002). A list of interview goals was drawn up, followed by a draft set of questions, which was piloted and improved upon to make the final interview questions. I also used online-buying guides as a secondary source of data.

✧ During observation, I noticed the following customer actions on both the DELL (www.Dell.com.cn) and eBAY(www.eBay.com.cn) Web sites. Review of the categories and details

✧ Menu selection of a shopping category

✧ Description of each of the features and behavior of the system and how the customer interacted with it

✧ Customer habits on Web site.

In order to carry out the evaluations it was necessary to:

✧ Review the above two sites to establish sector and purpose of each site

✧ Apply the task-based criteria to a single site to evaluate their feasability

✧ Define a customer purchase task for each sector based on customer needs

✧ Review each web site by simulating the execution of the customer purchase

✧ Document the customer purchase process using the task-based criteria

✧ Analyze the results of the user study group.

Through the interview, observations and user studies I developed the following sequence for an online catalogue system:

✧ The user arrives at the home page, sees a description of the shopping system, and views the initial graphic of the shopping area

✧ The user clicks the “Go To Catalogues” button

✧ A graphic display of all the catalogues appears

✧ The user clicks on a catalogue and the catalogue home page appears with a graphic and table of contents for the catalogue

✧ The user clicks on the “Open Catalogue” button and see a catalogue page with the picture of the item display form.

The analysis of customer purchase task requirements defined above identifies the types of shopping procedure required in purchase decision-making. Taking an observation approach, administered to six University students, these above criteria were then be applied to the two Web sites (Dell and eBay) in order to see if their success was related to the accessibility of their task.

4. RESULTS

4.1 Summary Scores for Methodologies

There were two main groups of shoppers in the laptop market: businesses and individuals. Following preliminary enquiries into the possibility of interviewing business shoppers I ascertained that businesses shop in a dramatically different manner to individuals. From 20 questions applied to a total of five actions, 17 received affirmatively responds. The remaining three responses held mixed meanings and, overall, none of the questions asked about the actions was responded to completely negatively. Similarly from a total of 8 heuristics assessed, the site scored 42 out of a possible 56.

4.2 Result from Interviewing

I conducted forty interviews, in addition to six hours of observation and an analysis of buying guides (see table3). Considering the quality of data collected and time constraints, several groups emerged.

After analyzing each group independently it became apparent that all shoppers undertake some form of research, no matter which group they are in. The expert buyers keep their knowledge current through research, and first time buyers will need to build an initial knowledge base. The form of research can also vary dramatically between and, indeed, within shopper groups. It is worth noting that the research stage is the most time consuming in the majority of cases and that it is also the stage that a large number of shoppers felt they could have improved upon.

4.3 The Result of Observation

It is worthy of note that all participants were familiar with the Internet, so that none of them really had significant problems accessing or navigating the sites.

The result shows the satisfaction scores across the two sites (max satisfaction possible=5). Results from the comparison reveals the participants are more satisfied with Dell.com.cn(mean=4.00) than eBay.com.cn (mean=3.58).

There were no significant differences in aspects of layout and design, navigation efficiency, content quality, product and service pricing, search facility, and purchase tracking between the two sites. However, participants did report preferring Dell.com.cn over eBay.com.cn. One user pointed out the slow page download speed at eBay.com.cn, on which also existed problems in these stages: security and privacy statement, ability to customize, and customer service support.

According to the observation, the following five tasks were fundamentally important to users of the system: acquiring knowledge about laptops, making up a shopping list and browsing the e-store inventory, placing the order for selected items, purchasing a laptop, and finding reassurance. The search result listing on the two sites provided detailed laptop information, which was usually classified by the product features, such as appearance, specification and other particular parts. It was this information on product features that attracted potential customers to purchase the laptop online.

5. ENVISIONED TASK MODEL

5.1 Overview of Areas for Improvement

While researching laptops, shoppers might need to search in many different media for information. This was felt necessary due to the quality, and the quantity of the relevant information they could obtain in different media. The drawback was that it demanded more time from the shoppers, both physically (traveling to shop, visiting all the vendors’ websites, etc.) and mentally(reading and gathering relevant information from reviews, memorizing all available options, etc.).

Shoppers might not get the best options available to them due to limited resources (e.g. no access to friends with relevant expertise), and the amount of time and effort they were prepared to devote to the process of searching, understanding and choosing the relevant information for short-listing purposes. It is possible for a shopper to be unaware that a better option exists even after he has bought the laptop.

It was also noticed that shoppers frequently seek reassurances in choosing suitable laptop models or making decisions on which laptop to purchase. However, these reassurances might not be constructive since they could be coming from people with no expertise.

The research indicated that improvement was needed in the following areas.

To improve the efficiency of information gathering by:

- **Bringing together** as much up-to-date, accurate and reliable information from different media as possible so that it is accessible through the method chosen by the shoppers.
- Make this collection of information **accessible** via different media so that the shoppers can choose the most suitable medium depending on the situation they are in (e.g. a shopper might have an unexpected free time slot, and would like to visit the nearest retailer that sells the laptop models he is interested in, therefore wants both the location of retailer, and his shortlist of laptops via his mobile phone).
- Assisting the **search** of relevant information (e.g. search tools and other facilities for asking an expert to guide the shopper).
- **Presenting** information to the shoppers in an easily understandable format.

To aid decision making by:

- Presenting information to the shoppers in a consistent format that can be easily **comparable**.
- Providing unbiased and **reliable** second opinions that are relevant to the shoppers.
- Offering suggestions of **suitable** laptops that are tailored to their requirements.

5.2 The Envisioned Task Model

In order to get a better understanding of the different types of knowledge shoppers can obtain from different channels, we can group similar type of information together in order to create an overview. In the envisioned TKS, in which I outlined the proposed goal structure, the reasons behind these changes can best be illustrated in conjunction with the HTA model.

This model improves on how shoppers obtain information by delegating the job of gathering and organizing information to computers. In the original model, first time buyers spend considerable time searching for the right laptop and tend look at as many models as they can before forming a general specification they want from a laptop. This reduces the amount of time and effort shoppers have to spend in the information gathering stage.

The search feature helps shoppers narrow down their options quickly. However, it might be argued that this might not be a sufficient way of short-listing, because the narrowed down results might still be too many for the shopper to cope with. This is due to the amount of information that is supposed to be encapsulated in this feature. Also the shopper might not know enough requirements in advance in order to make the feature useful. Therefore, I have included an option for asking experts to recommend models for the shopper. The experts will find out the requirements by asking the shoppers questions such as their knowledge levels, why they need a laptop, and other questions that are relevant to their knowledge levels. Note the ‘expert’ can be a human expert or in the form of intelligent knowledge base that relate the answers from the shoppers to a suitable model.

Enabling shoppers to compare any group of laptops by setting the elements they want to compare and prioritizing the results, user can be greatly aided in making a decision. Once again by delegating the job of organization to the computer, it is important to note that the shopper can also get a second opinion at this stage.

In order to provide reliable, sound, and up-to-date information gained through either product/reviews or expert guides, there must be ways for the shopper to verify the source of the information. The model should also provide a way for identifying the shopper, so that it is possible to recognize the shopper’s preferences (e.g. A shopper who looks for laptops during his last visit and frequently visits the news bulletin section, can be given suggestions on relevant news about laptops which he might find useful).

The capability of identifying individual shoppers would also allow for the creation of a community structure where people can come together and share opinions. This gives a sense of belonging and provides the shoppers a channel for debating different issues related to laptops.

This model is based on the assumption that the shoppers have no disabilities. Data collected did not include shoppers with disabilities due to time limitations. It may be possible to extend and modify this model to accommodate this category (Takemura & Kiyokawa, 2001).

6. CONCLUSIONS

The second HTA model (as Appendix) focuses purely on the research task and reveals its true complexity. By looking at this task in detail we can see that the sub-tasks are highly iterative and that most of the sub-tasks can run concurrently. That is to say, for example, that researching manufacturers has an effect on researching vendors and short listing, while occurring at the same time as researching specification, which in turn has an effect on the research being carried out on manufacturers. Although not every shopper is likely to carry out all of these sub-tasks a number can be said to be universal, for instance it is highly unlikely that a shopper would not research price.

The HTA of the research task also reveals that the shopper spends a lot of time repeating procedures in order to fulfill each sub-task. It is likely that a shopper may revisit a website or re-read a magazine in order to find information on the ideal specification of laptops, whereas earlier the shopper was gathering opinion on a particular manufacturer. Not only that, but for each of the four researching sub-tasks it is possible for a shopper to look at up to four different sources. It also gives rise to a number of problems, particularly the difficulty inherent in comparing different laptop models for which you have differing levels of information from disparate sources. Other problems include: finding relevant information, bringing disparate information together in a cohesive fashion, ensuring information is accurate, ensuring information is up-to-date, knowing when to stop researching, and so on. These problems and the others detailed in the model will need to be looked at further in this study.

In light of the results obtained, it appears clear that conceptually the sites function well. There are however certain areas of redesign required but these are largely aesthetic considerations that will increase the usability of the site rather than how the site itself works. The most pertinent criticisms were related to feedback, which in turn is linked with error prevention. Redesign at this level for the next iteration of the design process should yield significant improvements in functionality. It is worthy of note that these issues were taken into consideration in the previous section and that the site reviewed is simply a prototype at the first iteration of the design process.

REFERENCES

[1] Alba, J. Lynch, J. Weitz, B. Janiszewski, C. Lutz, R. Sawyer, A. Wood, S. (1997). Interactive home shopping: consumer, retailer and manufacturer incentives to participate in electronic market places, in *Journal of Marketing*, July 1997.

[2] Dell website: www.dell.com.cn (March 29, 2006).

[3] eBay website: www.ebay.com.cn (March 29, 2006).

[4] Faulkner, X. (2000). *Usability Engineering*, Palgrave, New York.

[5] Fleming, J. (1998). *Web navigation*. Sebastopol: O'Reilly & Associates, Inc.

[6] Franzke, M. (1995), Turning research into practice: Characteristics of display-based interaction, of Proceedings CHI'95, 421-428.

[7] Georgia Institute of Technology. Graphics, Visualization and Usability Center. "GVU's WWW User Surveys," October 1999 (http://www.gvu.gatech.edu/user_surveys/; current April 26, 2006).

[8] Johnson, P., Johnson, H. and Wilson, S. (1995). Rapid Prototyping of User Interfaces Driven by Task Models. In: Carroll, J, M (ed.) *Scenario-Based Design: Envisioning Work and Technology in System Development*, John Wiley & Sons, New York.

[9] Kotler P., (2000). *Marketing Management*, Prentice Hall.

[10] Nielsen, J. and Mack, R.L. (1994), Usability Inspection Methods, John Wiley & Sons.

[11] Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. Organization Science, 5 (1), 14-37.

[12] Payne, S. (1995). Interview in Qualitative Research. In: Smith, J.A., Harré, R. and Lagenhowe, L.V. (ed.s) *Rethinking Methods in Psychology.*, Sage, London.

[13] Preece, J, Rogers, Y and Sharp, H. (2002). *Interaction Design: Beyond Human Computer Interaction*, John Wiley & Sons, New York.

[14] Rieman, J., Franzke, M., and Redmiles, D. (1995), Usability evaluation with the cognitive walkthrough, Proceedings of CHI'95, 387-388.

[15] Rosson, M.B. and Carroll, J.M. (2001). Usability Engineering. Morgan Kaufman.

[16] Saward, G., Shah, M.H. & Hall, T. (2005) "The role of navigation in intranet acceptance", in *Proceedings of IEEE sponsored, 3rd Atlantic Web Intelligence Conference*, Lodz, Poland, 6-8 June, 2005.

[17] Saward G, Hall T, Barker T (2004) "Information scent as a measure of usability" IEEE Metrics Conference, September, Chicago.

[18] Saward G, Hall T (2005) Using existing website ontologies to assist navigation and Exploration, AWIC2005, June, Lodz, Poland, LNAI Proceedings, Lecture Notes in Computer Science, Springer-Verlag.

[19] Takemura H. and Kiyokawa K. (2001, Eds.): *Proceedings of IEEE Virtual Reality 2001*, Yokohama, Japan, March 13-17,, 317.

[20] Wild, P.J., and Johnson, P. (2004) Deepening Consideration of Temporal Factors in Task Knowledge Structures. *Workshop on 'Temporal Aspects of Work for HCI' CHI'2004,Vienna, ACM Press*.

[21] Wild, P.J., Johnson, H., & Johnson, P. (2003) Understanding Task Grouping Strategies. *People and Computers XVII - Proceedings of HCI 2003: Designing for Society*, P. Palanque, P. Johnson, & E. O'Neill (Eds.), Springer-Verlag, 8th - 12th September.

[22] Wilson, S., Bekker, M., Johnson, P. and Johnson, H. (1997). Help and Hindering User Involvement – A Tale of Everyday Design, Human Factors in Computing Systems, *Proceedings of CHI97*, ACM Press.

FIGURES

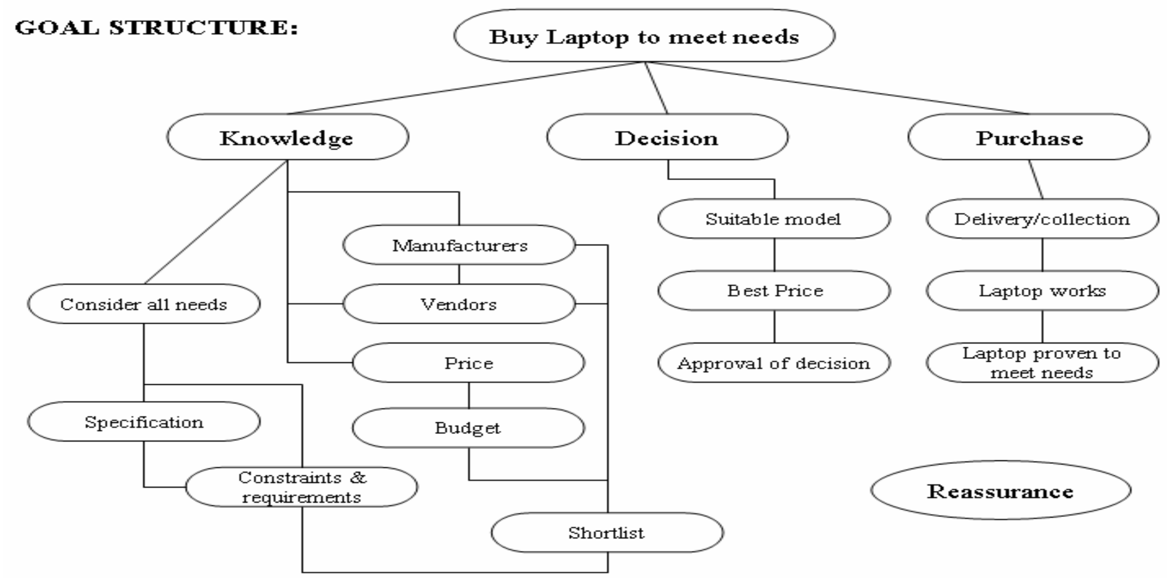


Fig. 1. TSK Model[1]

[1] TASK: Shopping for a Laptop; ROLE: Shopper (Researcher, Decision Maker, Funds provider).

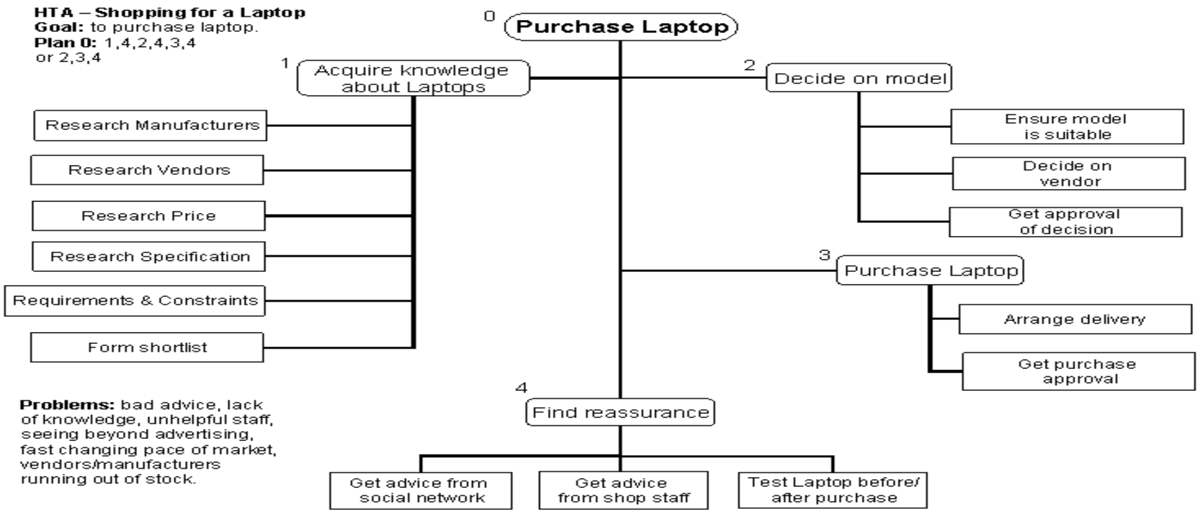

HTA – Shopping for a Laptop
Goal: to purchase laptop.
Plan 0: 1,4,2,4,3,4
or 2,3,4
0
Purchase Laptop
1
Acquire knowledge about Laptops
Research Manufacturers
Research Vendors
Research Price
Research Specification
Requirements & Constraints
Form shortlist
2
Decide on model
Ensure model is suitable
Decide on vendor
Get approval of decision
3
Purchase Laptop
Arrange delivery
Get purchase approval
4
Find reassurance
Get advice from social network
Get advice from shop staff
Test Laptop before/ after purchase
Problems: bad advice, lack of knowledge, unhelpful staff, seeing beyond advertising, fast changing pace of market, vendors/manufacturers running out of stock.

Fig.2. HTA Model 1[2]

TABLES

Table 1: Procedures for the consumer

Utilise internet	**Got to computer stores**	**Read Laptop magazines & books**	**Use social networks**
1. Search shopping sites. 2. Discuss in chat rooms. 3. Read review sites. 4. Use online buying guides. 5. Read manufacturer's sites. 6. Read vendor's sites 7. Purchase laptop.	1. Ask about manufacturers. 2. Ask about specification 3. Test display models. 4. Negotiate deal. 5. Seek advice. 6. Purchase laptop.	1. Look for manufacturer info. 2. Read laptop reviews. 3. Use laptop buying guides 4. Look at advertisements. 5. Read problem/letters pages. 6. Read articles on laptop development.	1. Ask advice. 2. Seek recommendations. 3. Purchase through network. 4. Find reassurance on decision. 5. Test laptops in network. 6. Get technical help

Table 2: Research Results Table[3]

STAGES	**PROCEDURES**	**OBJECTS**	**ACTIONS**
Decision to Shop			
-Is there a need? -Convince third party of need. -Budget for laptop. -Initial constraints/requirement.	- Assess need - Negotiate with third party - Use existing knowledge to form initial constrains/requirements	- Current situation - Third party -Existing knowledge.	- Negotiating - Assessing.
Research			
-Gain knowledge of laptop spec & components -Info on manufacturers -Info on vendors -Develop, constraints requirements & budget.	- Search WWW. - Use social networks. - Shopping, asking shop staff etc - Read magazines & literature. - Form shortlist.	- Chat room. - Review sites. - Online vendors. - Social networks. - Shop staff. - Laptops (display models). - Magazines. - Buying guides. - Knowledge.	- Reading. - Analyzing. - Comparing. - Discussing. - Searching. - Prodding.

[2] The HTA model 1 demonstrates the main tasks faced by anyone shopping for a laptop. Although not every task is necessarily undertaken, we have found that almost every shopper will invariably seek reassurance from a variety of sources at least once during their shopping task. This is represented by the reoccurrence of task 4 in the plan.

[3] Novice 1st time - Little or no previous experience of laptops or computers in general, has never bought (or been involved in buying) a computer of any type.

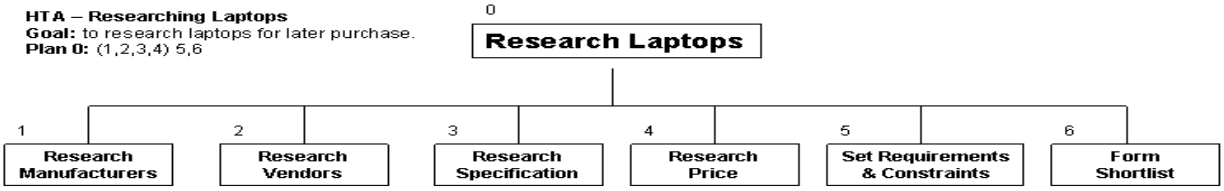
Experienced 1st time buyer - Good breadth of previous computer knowledge, but has never bought (or been involved in buying) a computer of any type.

Previous buyer - Good breadth of previous computer knowledge, has bought a computer of some sort previously.

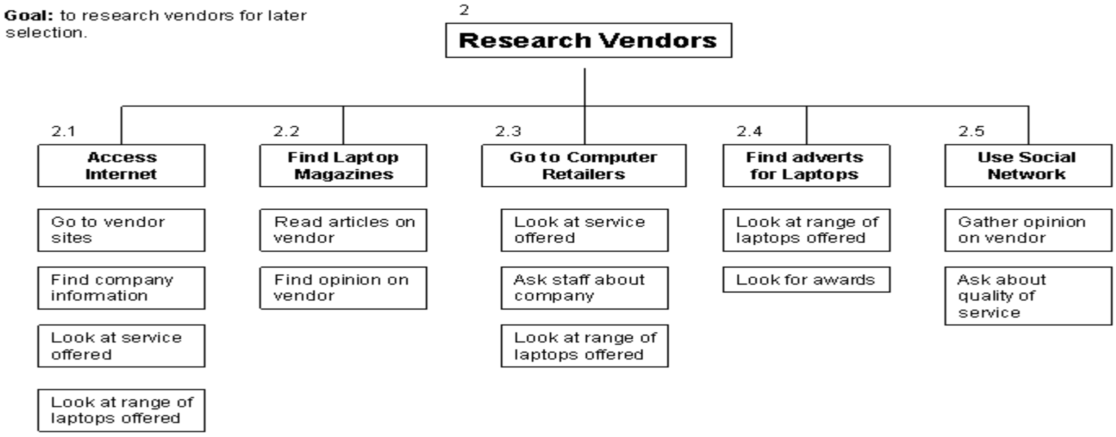
Expert buyers - Good breadth and depth of previous computer knowledge, has bought a number of computers before including laptops and/or has been heavily involved in helping others with the shopping process.

		- Shortlist. - Mail order brochures. - Vendor adverts. - Telephone.	
Decision to buy			
- Approval for purchase. - Decide on vendor. - Get best possible deal. - Try to ensure laptop will satisfy constrains / requirements.	- Online purchasing. - Purchasing in stores. -Purchasing through social network. - Negotiating with vendor. -Compare models on shortlist.	- Online vendors. - Computer stores. - Mail order vendors. - Social network. -Constraints/requirement. - Shortlist. - Telephone.	- Purchasing. - Negotiating. - Comparing.
Purchase			
- Ensure appropriate & quick delivery. - Purchase required laptop for best price / with most confidence.	- Purchasing laptop. - Arranging delivery.	- Purchased laptop - Vendor. - Delivery.	- Purchasing. - Arranging.
Post-purchase			
- Set up new laptop. - Check for defects. - Ensure laptop meets need.	-Installing software/hardware. - Testing laptop. - Returning laptop if defective / not appropriate.	- Laptop. - Laptop specification. - Software. -Constraints/ requirements. - Vendor.	- Installing. - Testing. - Checking. - Comparing.

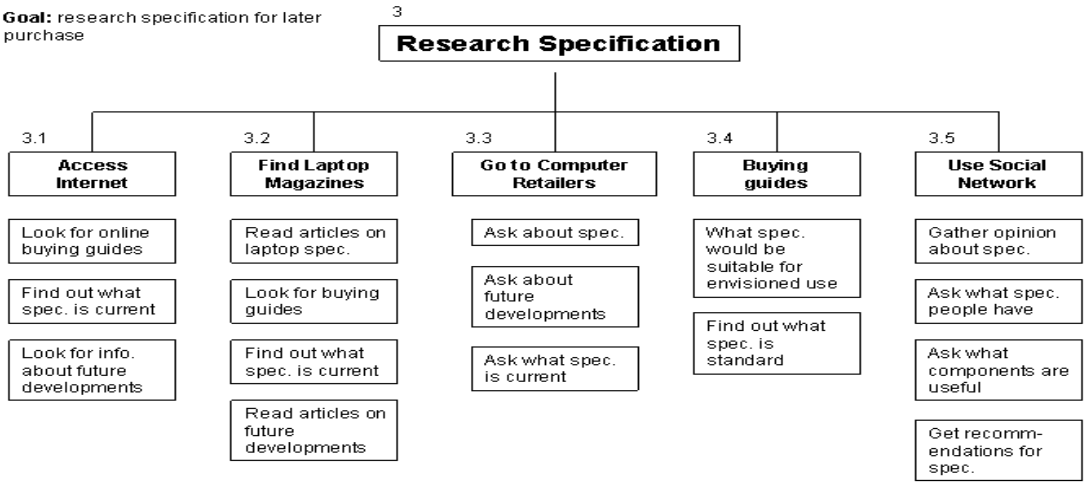
Appendix: HTA Model 2



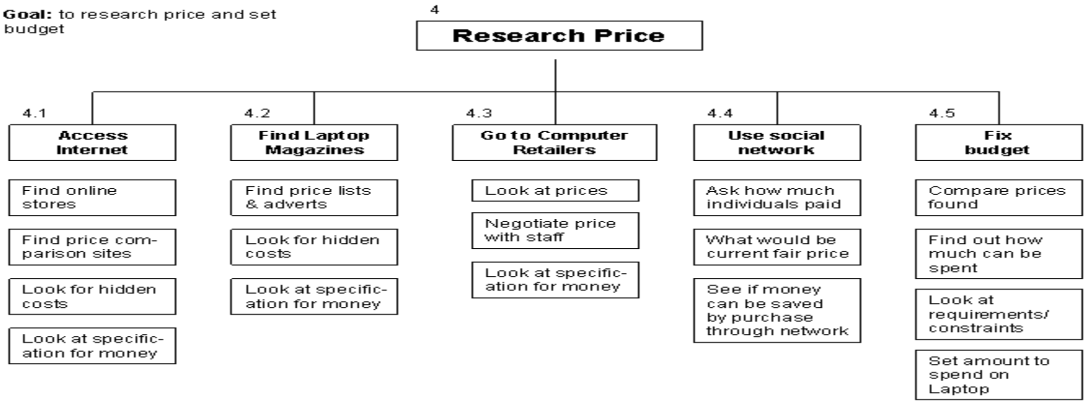
As the first four steps are normally carried out in conjunction they have been placed in brackets to indicate that they are undertaken together, although researching price may not occur until the first three tasks have been completed, or are underway. All of the tasks modelled above are iterative, for example when compiling a shortlist it may be necessary to further research manufacturers. It is worth noting that it is difficult to know when to stop researching due to the fast pace of the laptop sector and the complexity of the sector, it is for precisely this reason that reassurance is a necessary task when shopping for a laptop. Depending on the experience of the user all of these tasks may not be necessary, it is possible that when researching price a user could find a particularly attractive deal and move straight to the purchasing task.



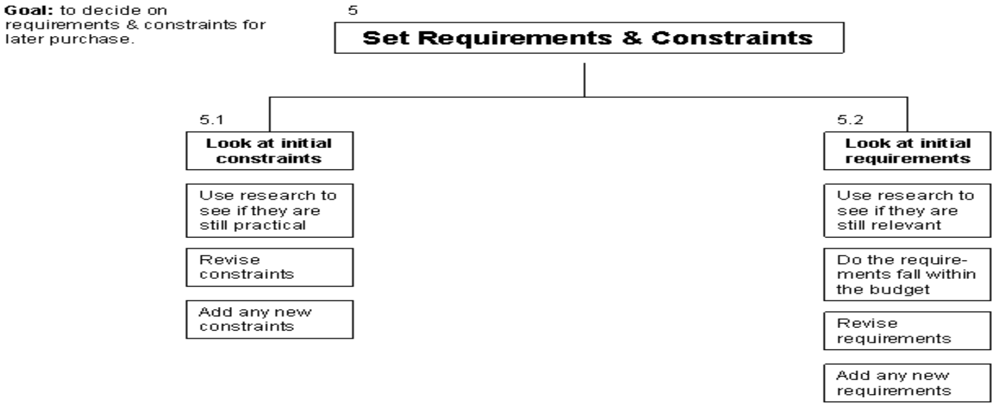
- There is no order to these sub-tasks.
- Sub-tasks are limited by shopping preferences; an avid online shopper is unlikely to consider retail outlets.
- It is not necessary for each of the sub-tasks to be completed.
- Most users will, however, complete at least one of the sub-tasks.


Goal: research specification for later purchase
3
Research Specification
3.1
Access Internet
Look for online buying guides
Find out what spec. is current
Look for info. about future developments
3.2
Find Laptop Magazines
Read articles on laptop spec.
Look for buying guides
Find out what spec. is current
Read articles on future developments
3.3
Go to Computer Retailers
Ask about spec.
Ask about future developments
Ask what spec. is current
3.4
Buying guides
What spec. would be suitable for envisioned use
Find out what spec. is standard
3.5
Use Social Network
Gather opinion about spec.
Ask what spec. people have
Ask what components are useful
Get recomm-endations for spec.

- There is no order to these sub-tasks.
- Not all sub-tasks are compulsory.
- Expert users are likely to skip this task completely as they may already have the required knowledge.

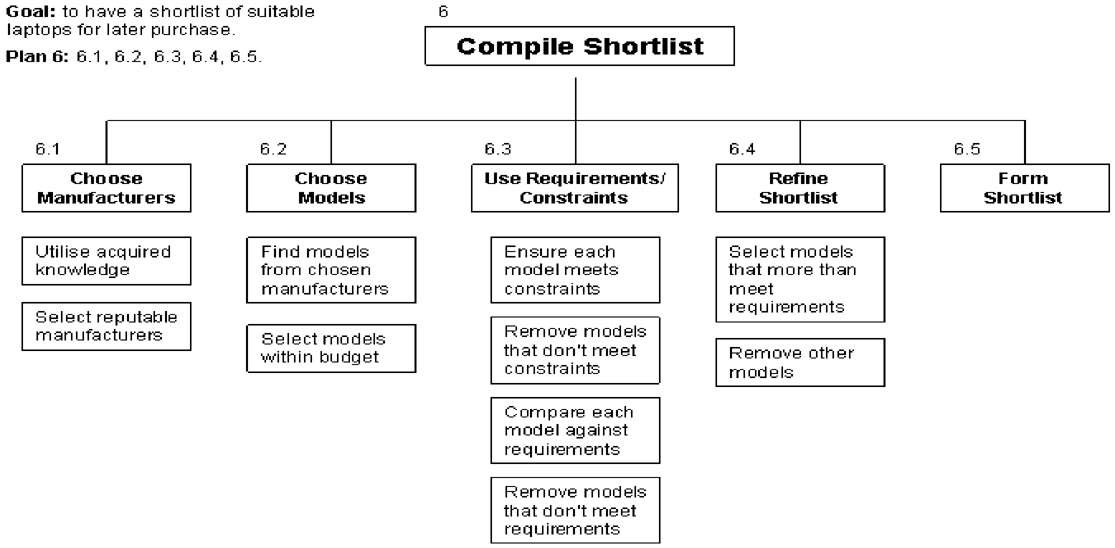

Goal: to research price and set budget
4
Research Price
4.1
Access Internet
Find online stores
Find price com-parison sites
Look for hidden costs
Look at specific-ation for money
4.2
Find Laptop Magazines
Find price lists & adverts
Look for hidden costs
Look at specific-ation for money
4.3
Go to Computer Retailers
Look at prices
Negotiate price with staff
Look at specific-ation for money
4.4
Use social network
Ask how much individuals paid
What would be current fair price
See if money can be saved by purchase through network
4.5
Fix budget
Compare prices found
Find out how much can be spent
Look at requirements/ constraints
Set amount to spend on Laptop

- The Fix Budget sub-task will invariably not occur until after the rest of the sub-tasks.
- The Fix Budget sub-task can also occur independently of any research tasks.
- This task will mainly occur concurrently with the rest of the research task, but not be completed until after the constraints and requirements have been set.


Goal: to decide on requirements & constraints for later purchase.
5
Set Requirements & Constraints
5.1
Look at initial constraints
Use research to see if they are still practical
Revise constraints
Add any new constraints
5.2
Look at initial requirements
Use research to see if they are still relevant
Do the require-ments fall within the budget
Revise requirements
Add any new requirements

- Both sub-tasks can occur concurrently.
- The majority of shoppers will have to carry out both tasks.
- A shopper may return to the research tasks in order to clarify a constraint or further detail a requirement.

- The research sub-tasks feed heavily into this sub-task and it therefore runs concurrently with them, but it will not be completed until after the research sub-tasks have been completed.


Goal: to have a shortlist of suitable laptops for later purchase.
Plan 6: 6.1, 6.2, 6.3, 6.4, 6.5.
6
Compile Shortlist
6.1
Choose Manufacturers
Utilise acquired knowledge
Select reputable manufacturers
6.2
Choose Models
Find models from chosen manufacturers
Select models within budget
6.3
Use Requirements/ Constraints
Ensure each model meets constraints
Remove models that don't meet constraints
Compare each model against requirements
Remove models that don't meet requirements
6.4
Refine Shortlist
Select models that more than meet requirements
Remove other models
6.5
Form Shortlist

This task’s sub-tasks have a more defined order. Task can be skipped if shopper has already decided on a particular model.

Count Internet Consumption and Selling Credit Factor for Personal Credit Scoring Model

Shujun Ye, Jing Liang
Economic and Business School
Beijing Jiaotong University, P.R.China, 100044
Email: shjye@bjtu.edu.cn

ABSTRACT

By applying Game Theory and analyzing the personal credit evaluation model of various banks within or outside China, this paper explains the reason and the importance of embedding Internet consumption and selling credit evaluation factor into current bank personal credit evaluation model. Moreover, it establishes a new integrated model to show the whole process of embedding Internet consumption and selling credit factor into existing model. The new integrated model not only results in a more accurate and objective evaluation but also considers the influence of technology development on personal credit evaluation. What’s more, it makes up the data deficiency for the potential customers of banks. On the basis of this new model, the author probe into the feasibility of importing Internet credit factor as well as bring forward several relevant suggestions.

Keywords: Game Theory, Internet Consumption and Selling Credit, Personal Credit Evaluation, the Integrated Model

1. INTRODUCTION

With the development of Chinese market economics, personal credit is becoming a premise of consumer market. It has close relationship with market justice and competition. The commercial banks in many countries have already established the personal credit scoring models .These models are aimed to help banks make a loan decision, and the accuracy of the model's result affects the bank's profit level directly. Therefore, it is critical to improve the evaluation model.

2. THE DEFICIENCY OF PERSONAL CREDIT SCORING MODEL OF CHINESE BANKS

At present, the personal credit scoring models of most banks are based on a hypothesis that a person's behavior is stable and foreseeable in a long period once it is formed. That means the future conduct can be judged by past behavior. So, it is obvious, the building of personal Credit scoring model is quite different between Chinese domestic banks and foreign banks.

Take the American national credit evaluation institutions for example, one of their credit scoring model is FICO, a econometrical model according to 'five C principle': Character, Capability, Capital, Collateral and Condition[1].In addition, American banks have their own independent credit databases with millions of loan records of North-America customers stored in them.

However, Chinese domestic banks do not have a uniform system for personal credit evaluation, and the credit factors in each bank differ. This inconsistency enables a bank to choose a particular credit scoring system, which may be suitable for it but may also lead liability and accuracy problems of credit evaluation in China. Some of the problems arise from modified statistic data, excess amount or obscure definition of credit score factors, which reduce the effectiveness of the scoring model. For example, some Chinese domestic banks divide the score factors, such as personal health, company economic development and career stability, into three simple levels, good, normal and bad, without clear definition. Another example is the ICBC（Industrial and commercial bank of China）personal credit scoring model. It takes 'personal impression' as a score factor, which mainly depends on subjective judgment.

What's more, Chinese domestic banks do not consider Internet credit and the influences of Internet technology progress as score factors for personal credit evaluation, and they only focus on quantity measurement of personal assets. However, the Internet credit factor can improve the liability and objectivity of the scoring model results and reduce the risks of personal loan, thereby enhance the bank's risk management level.

3. THE PERSONAL CREDIT REFLECTION ON INTERNET TRANSACTION

As on-line payment has become a emerging method for personal consumption, the Internet credit is playing an increasingly important role in consumer market. In some developed urban areas, more and more transactions happen on the Internet and Internet transactions begins to win the popularity among young generations who will be the future potential customers for Chinese domestic banks. According to China Internet development Statistic Report, in recent years, there is an evident rise not only in the total number but in the proportion of Internet consumers among on-line people. The Table 1. implies that the total number of Chinese Internet consumers from January to June in 2005 increased by 220.5% compared with that of the last six months.

Table 1. The change of the number of the Internet consumers in China [2] [3]

Investigation deadline	The number of Internet consumers（thousand）	Total percentage
2004-12-30	94000	6.7%
2005-6-30	103000	19.60%

In China, there are two common transaction models for personal Internet consumption. One is auction, the other is retail. The retail model is a simple pay-and-buy process, in which all the bargain information is available on the e-shop websites .In this paper, the Internet credit factor is discussed under the retail model, because it is more commonly used among Chinese Internet consumers.

The retail model has three different transaction methods: COD, advanced payment and delivery before payment. No matter which method is chosen, Internet consumption can not avoid the personal credit problems.

First, according to the Game Theory[4], when the buyer and the seller in Internet transaction are not able to make a COD, the Internet credit game will happen as is shown in

Table2. The first number written in each square of Table 2 presents the benefit of the seller, and the other number presents the benefit of the buyer (A>1,B<0,A+B<2).

The table shows that the buyer has two choices: pay or not to pay, and the seller can also make a delivery decision freely. What's more, it is supposed that both the buyer and seller are pursuing the maximum individual benefit and no one knows whether there will be further deals between them. In the advanced payment method, if the buyer is not willing to pay, then the game is over and no one receives benefit. If the buyer agrees to pay first, then the seller has the incentive to break the deal, because the individual benefit will be A, greater than 1, and the buyer will lost B because of the transaction cost, such as the length of Internet connection. So a sensible choice of the seller will be not delivering the goods without any personal credit; although the total benefits will be less than the maximum value which is 2.

Similarly, the under the delivery-before-payment method may still arose problems. The seller delivers the goods to the buyer first, the reasonable choice for buyer is not to pay, so that there will not be any personal credit for the buyer who wants to maximize his benefit. If he does not pay after the goods delivery, he will get A, greater than 1, and the seller will get the negative value B due to some transaction costs, such as transportation fees. If the seller predicts that the buyer will not pay, he will not deliver the goods, thus no transaction happens.

Table 2. The Internet credit game

Benefit		The buyer	
		Payment	No payment
The seller	Delivery	1, 1	B, A
	No delivery	A, B	0, 0

Even if both the seller and the buyer agree to adopt COD method, there are still many limitations. First, COD needs more transportation fees because a deliveryman needs to be employed to receive cash from the buyer. If this additional delivery cost is covered by consumers, the product price has to rise and the market sales of the seller will decrease. If the seller takes over the additional delivery cost, the sales profit will decline. Second, COD requires the buyer to pay cash face to face and it is not applicable to transactions of large amount of money. However, if the amount of sale is not big enough, the goods profit can not exceed the COD method cost, thus the transaction will not happen too.

In conclusion, an Internet transaction can not happen without personal credit, and this kind of credit reflection is defined as Internet selling and consumption credit factor, or Internet credit factor for short.

In addition, the empirical data also provide evidences for the above conclusion. The CNNIC statistics[2] indicates that nearly 50% Internet users choose 'lack of personal credit' as the biggest problem of Internet transaction, which includes information cheat(7.3%)and poor guarantee on product quality, customer service and seller's credit(42.4%). For the delivery method, only 24.7% Internet users would like to choose COD (shown in Table 3.), which is coincident with the analysis on COD's limitations and uncommonness.

Both the Game Theory and empirical data imply that the Internet transaction records are effective reflections on personal credit, and the Internet credit factor should be integrated into the existing credit scoring models to improve the quality of evaluation results.

4. NEW INTEGRATED SCORING MODEL WITH INTERNET CREDIT FACTOR

To embed the new factor -- Internet credit, the existing credit model has to be rebuilt and calculated in the process below, and the new established model is defined as the integrated model, which was introduced by Yajun Guo in his book concerning comprehensive evaluation [5].

First, the integrated model takes the Internet credit factor as an aggregation of limited variables whose data are available in the Internet transaction records, such as the number of successful deals, average amount of payment, or credit failure times. All these variables should be able to be recorded automatically from real Internet transactions by electronic systems, without any participation of either sellers or buyers.

Table 3. The choice of payment methods [3]

Method	Percentage
COD	24.7%
credit card	41.5%
mail	16.7%
remittance	16.7%
others	0.4%

Second, the integrated model embed a new factor by setting a sequence of importance. It supposes that personal credit is a system including limited evaluation factors{ $\mathrm{x}_{\mathrm{ij}}^{(\mathrm{k})}$ } $k = 1, 2, 3, \ldots, n$, and each factor is determined by a series of variables x_{ij} with the weight $W_j^{(k)}$, $j = 1, 2, \ldots, m\,; k = 1, 2, 3 \ldots, n$

Then the personal credit score can be calculated by equation (4-1):

$$Y = \sum_{i=1}^{m} W_j^{(k)} x_{ij}^{(k)}, i = 1, 2, \ldots, n_k; k = 1, 2, 3 \ldots, n \qquad (4\text{-}1)$$

The value of $W_j^{(k)}$ is calculated by the method 'G1'[5].This method do not have the coherence problem and it is more convenient to apply compared with A.H.P method (Analytic Hierarchy Process)[6].

Detailed steps are listed below: presume that Internet credit factors consist of m variables, which is $\left\{x_1, x_2, x_3, \ldots, x_i, \ldots, x_m\right\}$.

Step1: set the sequence according to variable importance. The expert selects the most important variable from $\left\{x_1, x_2, x_3, \ldots, x_i, \ldots, x_m\right\}$ and marks it by $x_i^{(1)}$ .Then he selects the second most important one from the rest m-1 variables marked $x_i^{(2)}$ …when there are *m-(k-1)* variables left，the expert selects $x_i^{(k)}$ and repeats this process until completes *m-1* selections. Therefore, a sequence of variable importance is set that is

$$x_i^{(1)} \succ x_i^{(2)} \succ x_i^{(3)} \ldots \succ x_i^{(k)} \ldots \succ x_i^{(m)} \qquad （4\text{-}2）$$

Here, $x_i^{(k)}$ refers to the number *i* variable of Internet credit factor in the " $\succ$ "sequence.

Step 2: give the value of variable importance. If the importance ratio of $x_i^{(k)}$ and $x_i^{(k-1)}$ is $W_i^{(k-1)} / W_i^{(k)}$, then:

$$w_i^{(k-1)} / w_i^{(k)} = r_{(k)}, k = m, m-1, m-2, ..., 3, 2 \qquad (4\text{-}3)$$

When the value of m is big, $r(m)$ in the equation（4-3）is equal to 1. In other cases, the value of $r(k)$ $r(k)$ would be able to be determined by Table 4. and $r(k)$ should meet the requirements in equation (4-4):

Table 4. The reference value of $r(k)$

$r(k)$	Explanation
1.0	$x_i^{(k)}$ has the same importance as $x_i^{(k-1)}$
1.2	$x_i^{(k)}$ is more important than $x_i^{(k-1)}$
1.4	$x_i^{(k)}$ is much more important than $x_i^{(k-1)}$
1.6	$x_i^{(k)}$ is far more important than $x_i^{(k-1)}$
1.8	$x_i^{(k)}$ is exceedingly important than $x_i^{(k-1)}$

$$r_{(k-1)} \succ 1 / r_{(k)}, k = m, m-1, m-2, ..., 3, 2 \qquad (4\text{-}4)$$

Step 3: count the weight $w_i^{(k)}$ of each variable. If the requirements in equation (4-4) are fulfilled, then $w_i^{(k)}$ can be calculated by equation (4-5) :

$$w_i^{(m)} = (1+\sum_{k=2}^{m}\prod_{i=k}^{m} r_i)^{-1} \qquad (4\text{-}5)$$

and the value of $W_i^{(k-1)}$ can be figured out by equation(4-6):

$$w_i^{(k-1)} = r_k w_i^{(k)}, k = m, m-1, m-2, ..., 3, 2 \qquad (4\text{-}6)$$

By building the integrated model, quantitative relationship has been successfully established between Internet credit factor and other factors in the existing credit scoring model, thereby forming a new model for personal credit evaluation.

5. THE FEASIBILITY OF IMPORTING INTERNET CREDIT FACTOR INTO EXISTING MODEL

It is commonly acknowledged that a new score factor should have the following characteristics: common in samples, comparable, measurable, independent and not capable of breaking the original systematic model. However, it seems that the Internet credit factor can not meet the operational requirements. The reason for it mainly lies in the first three required characteristics of a new factor. Firstly, in China, not all the bank personal customers are capable of making a deal on the Internet, which requires computer skills and infrastructural investments. Secondly, even if the bank customer is a Internet user, he or she may prefer other transaction methods instead of on-line consumption.

These difficulties will undermine the feasibility of new credit scoring model, but some measurements are capable of resolving the potential problem.

First, the universality problem of the Internet factor can be resolved by changing the score weight of the Internet credit factor. Detailed Steps are listed below:

Step 1: divide the bank customers into two different groups – Group A has already had personal credit records, Group B has no personal credit records;

Step 2: divide Group B into two groups named B1 and B2 – Group B1 has an average high rate of Internet transactions and group B2 doesn't have;

Step 3: different groups adopt different score weight of Internet credit factor. Group A should adopt smallest score weight while group B1 adopt the biggest one, and for group B2, the score weight value should between that of group A and group B2.

Moreover, this grouping pattern makes up the flaws of existing personal credit scoring model to a large extent, which attribute to two reasons: first, it provides the personal credit data for low-income people, especially for students who are the potential customers of banks. Second, the division criteria are available, because the Internet consumption and selling group has a clear characteristic in China. The CNNIC investigation shows that 68.6% Internet users' monthly income is below 1500 RMB, including half of the student users without monthly income or with monthly income below 500 RMB. What's more, among the entire student Internet users, up to 41% have lower monthly income which rang from 501RMB to 1000RMB, and only about 8% have monthly income above 1000RMB[3].

For the comparable characteristic of Internet credit factor, there are two methods to achieve it. One method is to calculate percentages and compare them by a score rank. For example, taking the number of successful Internet transactions as a variable of Internet credit factor, and the percentage of this variable equals to the number of successful personal Internet transactions dividing the total number of personal Internet transactions, and then multiplied by 100%. A bank can give a score rank as is shown in Table 5. according to its risk management ability.

Another method to make Internet credit factor more comparable is using the frame of reference, which means that a group of customers with similar characteristics will be selected as reference sample, and the rest of customers will be scored by being compared with the selected group.

Finally, the measurable character of Internet credit factor can be achieved by collecting data from large electronic transaction platforms. Some official regulations may be needed to guarantee the data authenticity, such as identity ID registration and random email investigation.

However, all the solutions introduced above need further exploration and can not promote the feasibility completely.

6. CONCLUSIONS

In summary, the Internet consumption and selling credit factor should be integrated into the existing personal credit scoring model, because it not only reflects the personal credit but considers the influence of technology development on transaction ways. What's more, the new integrated scoring model with Internet credit factor is able to make up the deficiency of some customers' credit data, especially for Internet student users. Thus, although the feasibility of this new model needs further discussion, the integrated model will result in a more accurate and objective personal credit evaluation.

Table 5. Percentage and score calculation

The percentage of successful transactions	Credit rank	Score
100%-90%	A	5
90%-80%	B	4
80%-70%	C	3
70%-60%	D	2
60%-50%	E	1
Under 50%	F	0

REFERENCES

[1] Tao Zheng, “Research on personal credit evaluation on consumption credit,” *postgraduate dissertation*, (in Chinese), pp.39-40, 2003.

[2] China Internet Network Information Center (CNNIC), *Statistical report on Chinese Internet development status*, (in Chinese), 2005.7.

[3] China Internet Network Information Center (CNNIC), *Statistical report on Chinese Internet development status* (in Chinese), 2005.1.

[4] John Von Neumann and Oskar Morgenstern, *Game theory and economic behaviour*(in Chinese),Beijing: life. reading. knowledge bookstore, pp.258-260,2004.

[5] Yajun Guo, *The theory and method of comprehensive evaluation* (in Chinese), Science Press, pp.18-19, 2002.

[6] Xin Zhao and Li Li, *Personal credit evaluation system built on AHP and GEM integrated arithmetic* (in Chinese), Rural Financial Research, vol. 4. Pp.21-24., 2002.

[7] Shuguang Li, *Personal credit evaluation* (in Chinese), P.H.D. dissertation, pp.83-85, 2003.

[8] Mingxia Wei, *A research on trust risk system of electronic commerce* (in Chinese), Forecast, vol.24, no.5, pp.49-52, 2005.

[9] Feng Li and Feng Yao, *Resolution for bottleneck of e-business* (in Chinese), Economic Tribune, no.12, pp.144-145, 2004.

[10] William F. Treacy, “Mark Carey, Credit risk rating systems at large US banks,” *Journal of Banking& Finance*, vol.24, pp.186, 2000.

[11] Lyn C. Thomas, “Survey of credit and behavioral scoring: forecasting financial risk of lending to consumers,” *International Journal of Forecasting*, vol.16, pp.152, 2000.

Research on Channel Coding Technology in RFID System*

Xiaohua Cao, Dexin Tao, Wenfeng Li
Logistics Department of Wuhan University of Technology
Wuhan, Hubei, 430063, China
Email: Tomm_cao@163.com

ABSTRACT

This work analyzes the communication process of RFID system, and establishes some simulation models of channel codes by Matlab/Simulink software tool. On basis of these models, some simulation experiments are done, and some useful results are obtained. The results show that Single Polarity Naught Code holds higher collision-detecting ratio, Manchester code holds lower BER in the same channel. In order to improve the identification credibility of RFID system, a suitable coding mode should be chosen, and the SNR should be bigger than 30 db in RFID system. The approaches and results in this work are helpful for quickening the development of RFID system and solving some practical application problems.

Keywords: Channel Code, RFID, Simulation

1. INTRODUCTION

1.1 Motivation

Recently, Radio Frequency Identification (RFID for short) has been one of the fastest developed technologies in the field of automatic identification and data collection. It carries out bidirectional and non-contact communication to achieve the aim of identification and data exchanging. There are two parts in a typical RFID system: tags and read-write devices. This system holds many obvious advantages, including non-contacting, non-optic, non-manpower, identifying more than one tag simultaneously. However, the inevitable channel noise during the RF communication is possible to depress the system's identification performance. So an appropriate channel code mode is needed to adapt the transmitted signal to the channel state in RFID system. The channel code is divided into two parts: Baseband Code and Error Control Code. The Baseband Code means how to express the binary numbers '1' and '0' with electric signal. The Error Control Code means how to convert the irregular or weakly regular digital signals into the regular ones.

The recent research on channel code technology in RFID system concentrates on spectrum characteristics, power supply of the Baseband Code, the correcting capability of Error Control Code. An analytical model of Baseband Codes was built to obtain theoretical power spectrum[1,2]. The RF input power of Passive RFID Transponder was maximized and optimized [3,4,5]. The correcting capability of Error Correcting Code was analyzed[6]. The channel codes were modulated into different signal in the wireless channels[7,8,9]. However, it is little to research the relation between the channel codes and the system's identification performance.

This work is focused on the analysis of the communication process of RFID system. A simulating model of RFID communication system is established by Matlab/Simulink software. The relation between the channel codes and the system's identification performance was researched.

1.2 Problem Statement

The usual communication process of RFID system is shown in Fig.1.

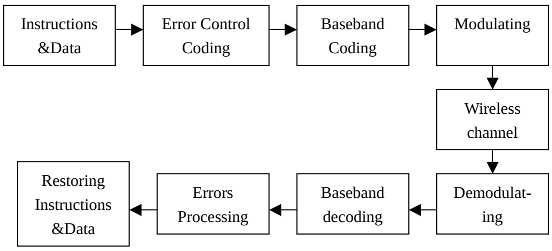


Fig.1. The communication process of RFID system

The instructions and data are encoded orderly into Error Control Code and Baseband Code, modulated and transmitted to the receiving unit through wireless channel. When the receiving unit detects the signals, it will demodulate these signals and decode orderly them into Baseband Code and Error Control Code according to the same principle as the sending unit's. The instructions and data will be restored in the receiving unit.

The channel code technology is one of the most important factors that influences the RFID systemic performance[1,6,7]. The general Baseband Codes contains NRZ Code, Manchester Code, Single Polarity Naught Code, Differential Bidirectional Code, Miller Code, Differential Code and so on. Their waves are shown in Fig.2.

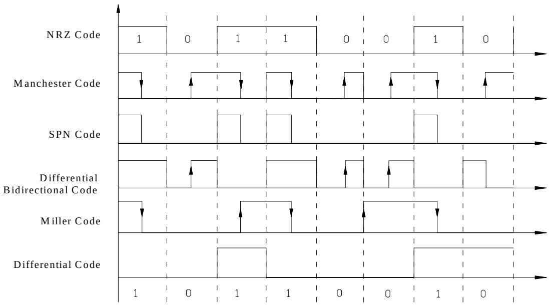


Fig.2. The waves of Baseband codes

The data transmission is non-contact, interfered easily during the wireless communication in RFID system. The interferences come from the inner thermal noise in the system and EMI outside the system, make the transmitted signals abnormal, as shown in Fig.3.

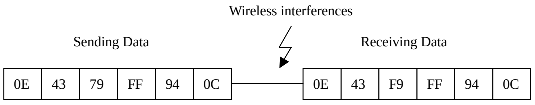


Fig.3. The errors coming from wireless interferences

The errors come from the distortion of transmitting signal, are tolerated impossibly in RFID system. There are two approaches

* This work was supported by Nature Science Foundation from Hubei province(2006ABA302)

to solve this problem. One is to increase output power of read-write device, but it will produce some electromagnetism pollution. The other is to use Error Control Code, add some checking bits into the original data. The Error Control Codes used often in RFID system contain Parity Check Coding, Hamming Code and CRC.

When the read-write device finds more than one tag simultaneously, a collision will be likely to happen. The read-write device has to judge if a collision has happened. Thus the collision-detecting ability of channel codes becomes very important in RFID system. When a collision happens, the signal wave will change, as shown in Fig.4. It's possible to detect the collisions according to the wave-changing characteristic.

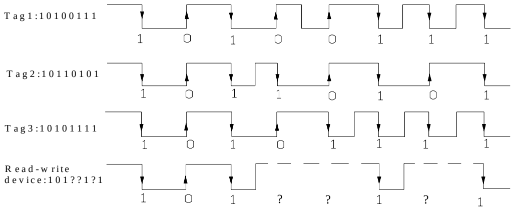


Fig.4. The collision wave of multi-tags

In order to improve the whole systemic performance, the collision-detecting and anti-jamming abilities must be strengthened. The different channel codes hold different collision-detecting and anti-jamming abilities. The key point of this work is how to design a simulating model for the RFID communication process, obtain the collision-detecting and anti-jamming abilities of different channel codes.

2. SIMULATION MODEL OF CHANNEL CODES

The communication process of RFID system is shown in Fig.1. It is feasible to design a simulation model to analyze the anti-jamming abilities of channel Codes. An integrated simulation model is established, as shown in Fig.5.

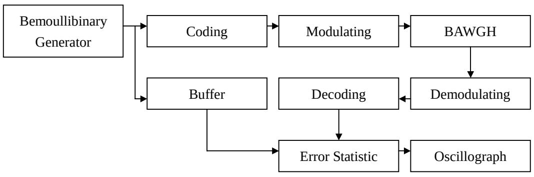


Fig.5. The integrated simulation model of RFID system

The Bernoulli-Binary Generator generates binary data. The data are encoded orderly into Error Control Code and Baseband Code, Modulated into high-frequency signals in ASK mode, transmitted in the channel with Gaussian White Noise. After these signals are received in the receiving unit, they will be demodulated, decoded and verified, then reverted into the original data. The Error Statistic module judges and calculates the whole systemic BER. The statistic result is displayed in Oscillograph module.

In addition, in order to gain Baseband Code's collision-detecting efficiency, another simulation model is established, as shown in Fig.6.

The error ratio of signal '1' is the same as that of signal '0' in the binary symmetrical channel. The sketch map of binary symmetrical channel is shown in Fig.7. The 1-P is the probability that '0' is changed into '1' or '1' into '0'.

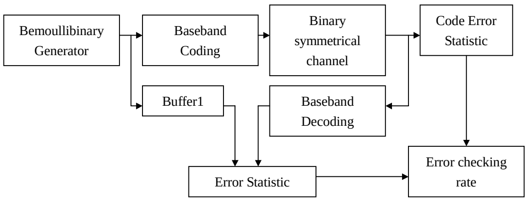


Fig.6. The simulation model of Baseband Code's collision-detecting efficiency

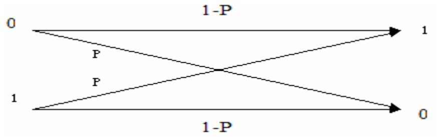


Fig.7. The binary symmetrical channel

The data are decoded into Baseband codes, transmitted through the binary symmetrical channel. By comparing the received data with the original data stored at the Buffer1, we can get the actual error number, symbolized by 'e1'. The Code Error Statistic module judges if the data collisions occur according to its demodulation wave, reckon up these errors as the detected error number, symbolized by 'e2'. The collision-detecting ratio can be obtained. The formula of collision-detecting ratio is:

$$\eta = \frac{e2}{e1}\% \qquad (1)$$

The Manchester code is taken for example to show the simulation process. Every inputted number '0' is coded into '01' and every inputted number '1' is coded into '10'. The Manchester coding simulation model is shown in Fig.8. A sequence of Manchester Codes can be produced through the simulation model.

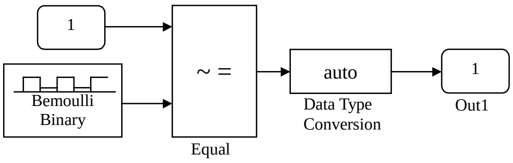


Fig.8. The Manchester coding module

Correspondingly, every pair of inputted '01' is decoded into '0' and every pair of input '10' is decoded into '1' in the Manchester decoding module. The simulation model is shown as Fig.9.

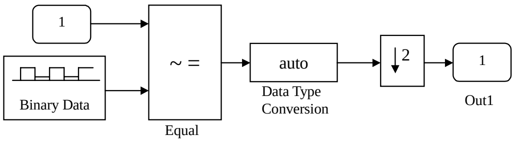


Fig.9. The Manchester decoding simulation module

In addition to Manchester Code, the simulation model of other coding modes, such as Single Polarity Naught Code, Differential Bidirectional Code and Miller Code, can be also established. The collision-detecting ratio can be also obtained.

3. ANALYSIS OF SIMULATION RESULT

3.1 The Simulation of Collision-Dear Sir or Madam:etecting Ability

In the practical RFID system, the length of every data frame is from several bits to dozens of bits, so the length of data frame is set as 50 in the simulation model shown in Fig.6. The Bit

Error Rate (BER for short) of binary symmetrical channel is changed from 1% to 50% in the simulation process. The simulation times is set as 2000. The data frames generated by signals generator and encoded into different channel codes, transmitted in binary symmetrical channel. The receiving unit compares the received data with the original data and adds up all errors. The result is shown in table 1.

Table.1. The Collision detecting ratio of different channel code

Coding mode	Factual error number(e1)	Detected error number (e2)	Collision detecting ratio(η)
Differential Bidirectional Code	1587	1143	72%
Single Polarity Naught Code	1602	1329	83%
Miller code	1568	0	0%
Manchester code	1602	1201	75%

The result shows that Manchester Code, Single Polarity Naught Code and Differential Bidirectional Code possess collision-detecting capability to some extent, but Miller code doesn’t. If the data are coded in Manchester Code or Single Polarity Naught Code or Differential Bidirectional Code, the wave characteristics can be grasped after a collision takes place. In opposition, if the data are coded in Miller Code, we can grasp little the wave characteristics after a collision. So there is no way to judge if there is a collision happens. Miller Code is not suitable for the work of collision-detecting in RFID system. The different coding modes hold different collision-detecting abilities. A better approach to increase the efficiency of anti-collision algorithm is choosing a suitable code mode.

3.2 The Simulation of Anti-jamming Ability

To simulate the practical wireless channel, the SNR (Signal Noise Ratio) of Gaussian White Noise channel is changed from 1db to 50db (the changing increment is set as 1db). The simulation times is set as 200. Some programs are designed to control the operation of simulation process shown in fig.5. The BER(s) of different channel codes are gained. The curve of simulation results are drawn out smoothly by linking every BER value, as shown in Fig.10~13.

The BER of different channel codes is different at the same channel. The BER of Manchester code is the lowest. The BER(s) of other three codes are approximate. When the SNR comes to about 33db, the BER of systemic data communication will come to zero.

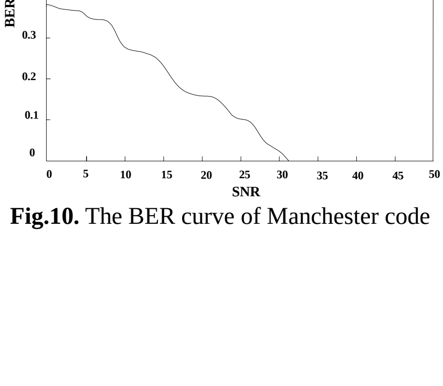


Fig.10. The BER curve of Manchester code

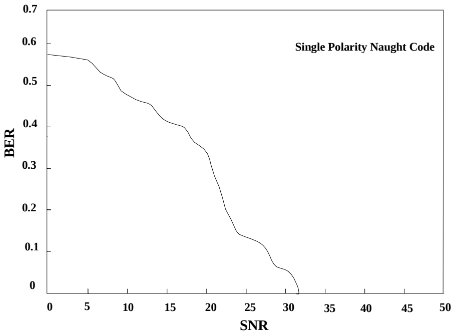


Fig.11. The BER curve of Single Polarity Naught Code

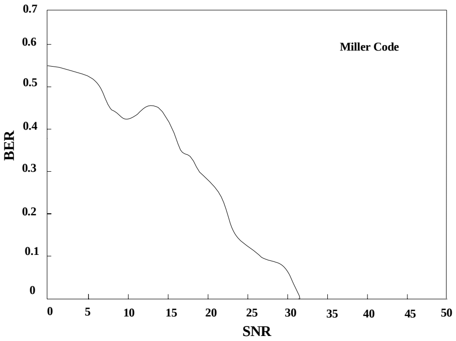


Fig.12. The BER curve of Miller code

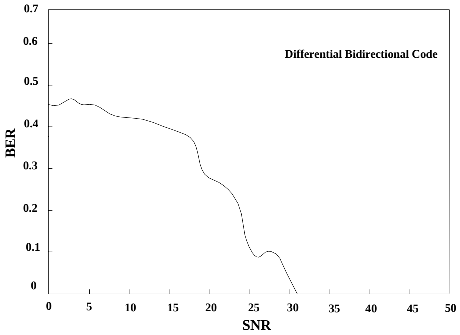


Fig.13. The BER curve of Differential Bidirectional Code

The reliability of data transmission will be up to the mustard in RFID system if the channel’s SNR is more than 30db. Thus the SNR should be kept bigger than 30db in RFID system. In order to weaken the influence coming from channel interference, a suitable channel code should be chosen.

4. CONCLUSIONS

In order to improve the identification credibility of RFID system, a suitable coding mode should be chosen and the transmitted signal should be adapted to the channel state in RFID system. After the analysis of communication process in RFID system, the simulation models are established through Matlab/simulink tool. The relation between the channel codes and the system’s identification performance was researched. The simulation results show that Single Polarity Naught Code holds higher collision-detecting ratio, Manchester code holds lower BER in the same channel and the SNR should be bigger than 30db in RFID system. The approaches and results in this work are helpful for quickening the development of RFID system and solving some practical application problems.

REFERENCES

[1] Simon.M.K, Million.S, “The Power spectrum of unbalanced NRZ and biphase signals in the presence of data asymmetry”, *TDA Progress Report* 42-126, JPI, NASA, August 15, 1996.

[2] Xu youyun,song wen tao,yuan min, “Wireless transmission technology to improve VHF band using rate”, *Radio Communication technology*, Vol.25, No.5, 1999, p51-53+64.

[3] Udo Karthaus, MartinFischer, “Fully Integrated Passive UHF RFID Transponder IC With 16.7-uW Minimum RF Input Power[J]”, *IEEE Journal of Solid-State Circuits*,2003, 38(10) p1602-1608.

[4] Bend Herbert sterassner II, “Microwave rectifying circuits and antennas for radio frequency identification and wireless power transmission applications”, *PH.D Dissertation*, Texas A&M University, August 2002.

[5] Inwhee Joe, “A Low-Power Hybrid ARQ Scheme for the RFID System”, *Lecture Notes in Computer Science*, v4208, 2006, p535 – 541.

[6] MacWilliams, Sloane, *The Theory of Error Correcting Codes[M]*,Amsterdam, North Holland Publishing Co., 1977, pp. 64～65 +482 +634～635.

[7] Ng, S.X., Guo, F. , “Jointly optimized iterative source-coding, channel-coding and modulation for transmission over wireless channels”,*VTC2004-Spring: Towards a Global Wireless World*, 2004, p 313-317.

[8] Hou, Jilei, Siegel, Paul H., *Capacity-Approaching Bandwidth-Efficient Coded Modulation Schemes Based on Low-Density Parity-Check Codes. IEEE Transactions on Information Theory*, v49, n9, September, 2003, p 2141-2155+2322.

[9] Yang, J., Sun, Y.; Senior, J.M. “Channel estimation for wireless communications using space-time block coding techniques. Proceedings-IEEE International Symposium on Circuits and Systems”, v2, 2003, p II220-II223.

Xiaohua Cao is a candidate for PhD in logistics department of Wuhan University of Technology. He graduated from Jiangsu University with the specialty of industry automation. His research interests are in RFID technology and port machinery remote monitoring technology.

Dexin Tao is a Full Professor and a head of Equipment Fault Diagnosis Lab, Vice-president of Wuhan University of Technology. He attended in a advanced studies in Hiroshima University of Japan (1983~1985). He has edited 4 books and published over 20 Journal papers. His research interests are in port machinery monitor state and fault diagnosis, port logistics technology and equipment. He is the principal of many tasks such as the failure analysis of wire ropes, performance tests of new-style pulleys.

A Parallel Mapping Algorithm for E-Commerce Web Pages to Semantic Concepts

Wenfang Yu, Yi Ouyang
College of Computer and Information Engineering, Zhejiang Gongshang University
Hangzhou, Zhejiang 310035, China
Email: jason_hi@163.com

ABSTRACT

How to find useful information is always a hot topic for researchers. Concepts recognition is a key problem in semantic information searching. A new semantic concepts framework EWO of web pages was built at first. Based on EWO, we proposed a parallel mapping algorithm for E-Commerce Web pages to semantic concept, which adopts two-stages concept mining method. The first stage is to implement local semantic schema; the second step is to implement global mining, and bridging relationship between local data semantic and global sharable ontology. Through combine ontology and web pages, the web pages and semantic concept can be mapping together for semantic searching. Experiments on several web pages sets show that it can outperform other methods in terms of precision and recall.

Keywords: Semantic Frame, Image Feature, Parallel Compute, Ecommerce

1. INTRODUCTION

The Semantic Web makes the computer to understand the Web information, and realizes intelligence interactive. The ontology is the semantic web foundation. As one kind of domain knowledge generalization and modeling method, it can be used to describe the computer processing data semantic information. At present, the ontology has become a standard for semantic web knowledge expression [1][2].

In order to achieve semantic information sharing, each domain has defined their ontology's standard respectively. For example, Cyc general knowledge Ontology database, enterprise information Ontology database, as well as biochemistry Ontology database and so on. However, the web distribution makes that each domain inevitably will define their ontology to describe their data. However, these different ontology structure have heterogeneous problems: two element of concepts have same meaning though their name are totally different, and at the same time, one concept object may be used as different names; Moreover the same domain concepts collection possibly defines the different ontology structure; In the different concepts structure, same example possibly uses the different expression method.

Semantic information mining over distributed ontology is one of the challenges about Web. This high level data representation reveals two possible methods for partitioning the data. The first method, document partitioning, slices the data matrix horizontally, dividing the documents among the subtasks. The N documents in the collection are distributed across the P processors in the system, creating P subclass of approximately N/P documents each. During query processing, each parallel process evaluates the query on the subclass of N/P documents assigned to it, and the results from each of the subclass are combined into a final result list. The second method, term partitioning, slices the data matrix vertically, dividing the indexing items among the P processors such that the evaluation procedure for each document is spread over multiple processors in the system.

In the semantic space, the concept object is represented in terms of the most frequent keywords. By mapping the concept object representations from the color feature space to their semantic representations, the system finds the associations between the low-level features and the semantic concept [3][4]. The associations are expressed as IF – THEN rules and could be used further to capture the semantic content and index new untagged images being added to the web pages database. The suggested system is also a powerful tool in reducing the semantic gap between the user's conceptualization of a query and the query that is actually specified to the system.

2. SEMANTIC CONCEPT FRAMEWORK OF WEB PAGES

Definition 1. Ecommerce Web Ontology (EWO) is a Five-tuple: EWO:=(W,C,M,Root(c)).

where W represents the keywords to describe the concept objects, C represents the objects of the real world, M is the mapping 2D matrix which reflect relation between the web page and concept; Root(C) is the root node of EWO structure.

Definition 2. Given two concept, C_i and C_j, the distance (=dissimilarity) of the two concepts is denoted by $D(C_i, C_j)$

The Web pages semantic information have some similarity, and these pages have some fixed format, such as, having title, introduction, hyperlink or advertisement figures. The arm of semantic searching is to find web information more efficient, and more reasonable. The hierarchical model of ECC have four layers, Keywords layer, Concept layer, Document layer and Fields layers, and the Concept layer is composed by a set of keywords which have same lexical link in WordNet[5].

A concept hierarchical may be defined based on an attribute field or on an attribute field set. Let a concept layer H defined on fields set Di,…, Dk, and different concept level has composed a hierarchical. Concept layer usually derived from concepts that arrange from general concepts to special concepts by partial order. The most general concept is the empty description, and the most special concept is the value corresponding in the database attribute. Its definition as follows:

The concept object is composed by several relative keywords, and the relationship between words can extracted from WordNet. Fig.1 shows an example of such a hierarchical structure. In this network, each node k_m represents the keywords in the collection, C_i represents the concept composed by a set relative keywords, each node d_j models a document, and the node f_n models the Ecommerce relative fields.

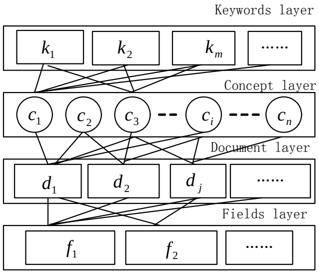


Fig.1. The structure of semantic concept layers.

Definition 2. Let N be the total number of WebPages in the system and n_j be the number of web pages in which the index term k_i appears. Let $freq_{i,j}$ be the raw frequency of term k_i in the webpage d_j. Then, the normalized frequency $f_{i,j}$ of term k_i in webpage d_j is given by $f_{i,j} = \frac{freq_{i,j}}{\max[\ freq_{t,j}]}$.

Definition 3. Let K ={$k_1, k_2,..., k_n$} be the set of nouns in a document, and R = {identity, synonym, hyponym, meronym} be the set of lexical relations. Let C = $\{c_1, c_2,..., c_n\}$ be the set of concept object in a document. Concept object c_j is composed of several keywords k_i, and each k_i and c_j have a weight that represents their respective degrees of semantic importance within a document.

Definition 4. Set H_n : $D_1 \times \ldots \times D_m$ => H_{n-1} => ..... => H_0, H_n represents the most primitive concept set; H_{n-1} represents those concepts that is higher than H_n in hierarchical structure. In the top layer, there are possibly including the most general concept " ANY ".

The structure of EWO contains five types of fundamental relations and three types of extended relations of ECommerce. The fundamental relations are defined as: is-a relation, part-of relation, associative relation, Synonymy relation and Meronym relation, whose definitions are the same as that of WordNet. Facing the special features of ECommerce, three extended relations are added into business concept entities: Geo-location relation, Apply-range relation and Apply-object relation, where each concept object is defined by a synonym set. To describe the relationship among the objects, we define two link Post-Link and Neg-Link. They are logical function. Given a data object sets S, if P,Q∈S and P,Q in the same class, then Post-Link(P,Q)=true;else they are exist in different class, e.g. Neg-Link(P,Q)=true. Formally:

(1)Post-Link and Neg-Link have symtem. For P,Q∈S
    Post-Link(P,Q)=true⇔ Post-Link(Q,P)=true
    Neg-Link(P,Q)=true⇔ Neg-Link(Q,P)=true
(2)Post-Link and Neg-Link have transitivity. For P,Q∈S
    IF Post-Link(P,Q)&&Post-Link(Q,R)=true THEN
        Post-Link(P,R)=true
    IF Neg-Link(P,Q)&&Neg-Link(Q,R)=true THEN
        Neg-Link(P,R)=true

3. Low-Level Feature Extraction

Methods that perform image searching based on features automatically extracted from the images themselves are referred as content-based image retrieval (CBIR) techniques[6][7][8][9]. The CBIR systems extract and compare primitive features (such as color, texture and shape) from stored and query images. The purpose of extracting image features lies in describing the images in the database with the features of low-level. These image features are useful for calculating similarity degree among different images, so it is obviously smaller than the primitive image on the size.

3.1 The color features

In the image searching fields, the color features is used the most extensive vision characteristic. The main reason lies in the color and object or scenes in the picture are closely related. In addition, compared with other vision characteristics, the dependence on size, direction, visual angle of the image itself of the color features is relatively small, thus the color features has better robust character. Researches on color features mainly adopt colored histogram, color moment, color collect, histogram refinement and color correlation diagram, etc.

The grayscale histogram extracted from an image is a vector that has 256 dimensions. All vectors are contained in histogram space S(H).

$$H = \{h_0, h_1,...h_m,...h_{255}\}^t, h_m \geq 0, 0 \leq m \leq 255$$

H represents the histogram, $h_m = \frac{B_i}{|B|}$ is the normalized value of the ith grayscale, B_i is the number of pixels corresponding to the ith grayscale. | B | represents the amount of pixels in an image. H can be regard as the feature vector of the color.

Where the feature vector of image is H', $d_k(H')$ is a measure of the distance between H' and the class centroid $H_K \in C_k$. The formula of the distance as follows:

$$d_i(H') = \sqrt{(H' - H_i)^t (H' - H_i)} \qquad (1)$$

3.2 Gabor texture feature

The Texture feature is a kind of characteristic not depending on the color or the luminance, and can reflect the visual characteristic of the homogeneity phenomenon in the image. Inherent characteristic that all object surface owns in common that it is, for example cloud, trees, brick, fabric, etc. All object have one's own texture characteristics. The texture includes the important information that the object's surface structure organization arranges, and these connections with surrounding environment of important information. Just because of this, the texture characteristic has extensive application in the content-based image retrieval, users can search other images with similar texture through submitting the image which include a certain texture[10]. Researches on the texture at present, commonly used methods have Tamura[11], wavelet transform[12][13], etc. Moreover, Gabor filters method can reduce the space and frequency to the greatest extent. Set the g(x, y) as a Gabor function and its Fourier transform as G(u,v),As follows:

$$g(x, y) = (\frac{1}{2\pi\sigma_x\sigma_y}) \exp(-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}) + 2\pi jFx) \qquad (2)$$

$$G(u, v) = \exp(-\frac{1}{2}(\frac{(u-F)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2})) \qquad (3)$$

Set $\sigma_u = \frac{1}{2}\pi\sigma_x$, $\sigma_v = \frac{1}{2}\pi\sigma_y$,and F is the complex modulation frequency of the Gaussian。g (x, y) is mother wavelet, pass g (x, y) correctly using g(x,y),we can get a series of wavelet filters by scale and rotate transform, and called it as the Gabor wavelet:

$$g_{m,n}(x, y) = a^{-m} g(x', y'), a > 1 \qquad (4)$$

where $x' = a^{-m}(x\cos\theta + y\sin\theta)$, $y' = a^{-m}(-x\sin\theta + y\cos\theta)$, and $\theta = k\frac{\pi}{n}$, a^{-m} as scale factor，n as direction number。

There is redundant information in the Gabor wavelet set image that processes by non-rothogonality method after filtering. In order to reduce redundant information, set U_h, U_l represent the centre high-frequency and centre low-frequency respectively, S represents the multi-resolution decomposition the number of times of the scale variety.

$$\sigma_u = \frac{(a-1)U_h}{(a+1)\sqrt{2\ln 2}}$$

$$\sigma_v = \tan(\frac{\pi}{2k})(U_h - 2\ln(\frac{\sigma_u^2}{U_h}))\sqrt{(2\ln 2 - \frac{(2\ln 2)^2\sigma_u^2}{U_h^2})},$$

where $a = (\frac{U_h}{U_l})^{-\frac{1}{(s-1)}}$,m=0,1,…S-1. An image can be represented by a convolution between the original image and the Gabor filter. Let I(x,y) is a picture, and $g_{m,n}(x,y)$ is Gabor wavelet, then the image is transform through the wavelet filtering result as follows:

$$M_{m,n}(x,y) = \int I(x,y)g'_{m,n}(x-x_1)(y-y_1)dx_1dy_1 \qquad (5)$$

After the Gabor filtering, the mean of the image that is $\mu_{m,n}$,and the square of the standard deviation is $\sigma_{m,n}$.

$$\mu_{m.n} = \iint |M_{m,n}(x,y)|dxdy$$

$$\sigma_{m,n} = \sqrt{\iint(|M_{m,n}(x,y)| - \bar{\mu}_{m,n})^2 dxdy}$$

Using the $\mu_{m.n}$ and $\sigma_{m,n}$ as components, these components constitute an eigenvector that can be using in the process of retrieval.

In the experiment, we let U =0.5,U =0.05, m =6, n =8. Based on computing the energy value of each filter and the image convolution, it can calculate the mean value of filtering wave energy value of every sub picture and variance ，which can be regarded as the texture feature of the sub image, such as:

$$X = [\mu_{00}, \sigma_{00}, \mu_{01}, \sigma_{01}, ... \mu_{57}, \sigma_{57}]$$

The distance $d(X_i, X_j)$ between two eigenvectors adopts the Euclidian distance to measure, where N is the number of eigenvectors dimension; the calculating formula as follows:

$$d(X_i, X_j) = \sum_{k=1}^{N}(X_{ik} - X_{jk})^2 \qquad (6)$$

4. Parallel Mapping Low-Level Image Features to Concept (PMLTC) Model

The system will reduce the semantic gap between the user's searching meaning of a query and user's keywords of the query that is actually specified. The system synthesizes the semantic features through a set of mapping low-level image features to semantic concept. First, an image analysis process is performed, consisting of two phases: low-level image features analysis phase, the basic objects composing a given image and their relative position are identified (such as computer, digital camera and mobile phone and so on). The high-level image analysis phase deals with image interpretation according to the [14][15]. At the end of the image analysis process, images are described in terms of the objects recognized, with associated belief and plausibility values, and the classes to which they belong. The mapping function is based on two things: domain semantics and statistical properties of low-level features. It then assigns them into mappings in the form of IF-THEN rules.

For each object a list is maintained. Each element of the lists of a pair(C,I),where C is the associated belief interval, representing the probability that the image considered really contains the object, and I is a pointer to the header of the image containing the object.

The development of the EWO query language has been driven by a number of requirements: first, it should be possible to easily navigate through the web page structure. Queries both on the content and on web page structure must be supported. In general, a semantic query has the form:

FIND PAGES TYPE *type-clause*
SCOPE *scope-clause*
COMPANY p*rovider-clause*
WHERE *condition-clause*
WITH *component-clause*

- The *type-claus*e allows the restriction of a query to web pages belonging to a specified set of types.
- The *scope-clause* restricts the query to a particular set of web pages. This set of web pages is either a user-defined web pages collection or a set of documents retrieved by a previous query.
- The p*rovider -clause* restricts the query to a particular company.
- The *condition-clause* is a Boolean combination of simple conditions on web page components. Predicates are expressed on conceptual components of web pages. Conceptual components are referenced by path-names. The general form of a predicate is:
  Component restriction
  where component is a path-name and restriction is an operator followed by an expression.
- The *component-clause* allows one to express structural predicates. Component is a path-name and the clause looks for all documents structurally containing such a component.

Different types of conditions can be specified in order to query different types of media. The EWO supports three main classes of predicates: predicates on data attributes, on which an exact match search is preformed; predicates on textual components, determining all objects containing some specific strings; and predicates on images, specify conditions on the image content. Image predicates allow one to specify condition on the class to which an image should belong or conditions on the existence of a specified object within an image and on the number of occurrences of an object within an image. The following example illustrates the basic features of the EWO query language.

Example *Consider the conceptual structure Generic-Web page, the following is an example of query:*

FIND PAGES TYPE IMAGE
SCOPE Page.Date < 01/01/2008)
COMPANY provider.name=" Lenovo"
WHERE *product_description contains "computer"
WITH *Company_logo

According to this query, the user looks for the all web pages, dated after January 2008, containing a company logo, having the word ‘Lenovo’ in the product provider, with the word “computer” in the product description section. Symbol ‘*’ indicates that the path-name is not complete, that is, it could identify more than one component. The conceptual structure of the type Ecommerce Web Pages is shown in Fig.2.

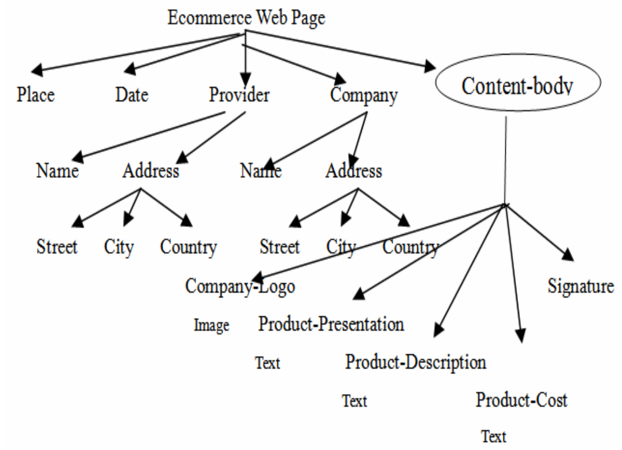


Fig.2. Complete conceptual structure of the type ECommerce Web page

So far, we have presented several ideas of how to create EWO structure and Semantic query language. The system searching process steps are as follow:

Step1: retrieve the query object Q from concept entities set EWO about the keyword K_i.
Step2: Mapping the query object Q into a point F(Q) in feature space.
Step3: using PCM() algorithm, compute Web pages and F(Q) similarity weight ,and retrieve all points within the desired tolerance e from Web pages.
Step4: Retrieve the corresponding objects, compute their semantic distance from Q.

Objects are mapping into 2D pointers(using the average and the standard deviation as features). Query becomes an f-D sphere in feature space, centered on the image F(S1) of S1. such query on multidimensional points is exactly what R-trees are designed to answer efficiently. In algorithm, the Cp is the concept collection of having Post-Link relation, and Cn as those concept collections of Neg-Link relation. More specifically, the parallel concept mapping (PCM) algorithm for a whole match query is as follow:

Algorithm PCM (E,C) //Parallel Concept Mapping
Input: E: concept hierarchical EWO ,C: the concept ready to query, rowcount is the web pages count.
Output: HIC

```
{
 Set K data objects as central point of K sets respectively,
 each subset number

 for (each processor Pi)
     for(i=1;i<N/P;i++)
          {
          Si←read ith subdomain;
          LoopCount=0;
            do{
               for(i=0;i<rowcount;i++) ADD(i);
                    PostProcess();
                    LoopCount++;
                    }while(LoopCount<Loops)
               }
}
function ADD(int recordNO)
{
          GetSubsets(recordNO);
          Switch(Dist())
          {
               Case 1: add the object data into subsets Cp;
               break
               Case 2: add the object data into subsets Cn;
               Break
               Case 3: add the new object data Cnew as the
               central concept object;
               Break
          }
}
function PostProcess
{     k=count(web pages);
      for(i=0;i<k;i++)
      {
          Computing the similarity between query object
          and each web page in subsets;
          feature1=F(Webpage(i), );
          If (feature1>e &&feature2>e)
          CreateMapping(webpage(i),

      }
}
```

The process of PCM perform can be separated into setting initial point and iteration two parts. The key step is the partitioning step, in which a set of concepts is based on EWO, the web page of the picture and the EWO structure node are created correlation.

In Algorithm PCM, the temporary results are stored in the variable Si, and each processor will be transmitted to others. After each processor receives the desired temporary result, these processors compute and update data. In other words, this algorithm divides m data of Cp, Cn into N/P sets and then computes them concurrently to obtain the value of Si within these sets. The execution time depends on the final finished processor.

5. EXPERIMENTAL RESULTS

This system has been implemented with a commercial web pages database including about 983,300 images.

The parallel PCM algorithm described in Section 3 was implemented and run into the five test images database. For each image, different sizes of the same image, ranging from 128×128px to 1024 ×1024px were used, and two measurements were taken. The total execution time,and computation time of the parallel algorithm. Based on these two measurements, the speedup of the algorithm, relative to a query image, was computed. The results of the empirical analysis are:

The dotted line denotes the real speedup obtained, while the solid line denotes the computational speedup obtained. $R_{speedup}$ denotes the real speedup, $C_{speedup}$ denotes the computational speedup, TE_i denotes the total execution time on i processes, and TC_i denotes the total time spent computing on i processes. We define real and computational speed as follows:

$$R_{speedup} = \frac{TE_1}{TE_p} \qquad (7)$$

$$C_{speedup} = \frac{TC_1}{TC_p} \qquad (8)$$

The results can be seen from above Fig.3.

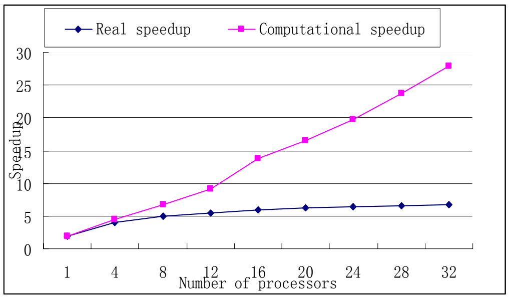


Fig.3. Speedup vs. Processors graph

Fig.3 shows that computational speedup running an advantage position with large parallel size than real speedup. Increasing the number of processors may improve the speedup in experiment, but the real mapping generated may not be efficient and needs optimization to get more performance.

The definitions of precision and recall are adapted by us to measure the experiment result. Recall describes the number of correct matching found in comparison to the total number of existing mappings. Precision we measure the number of correct mappings found versus the total number of retrieved mappings. The result of the experiment is as Table 1:

Table 1. Result of the experiment

Concept name	Recall	Precision
Clothing	0.688	0.695
Computer	0.712	0.701
Electronic	0.604	0.534
Flower	0.736	0.742
Jewelry	0.623	0.641
Shoe	0.727	0.743

6. CONCLUSIONS

This paper describes the definition of EWO and the designing of ontology. The key purpose is to improve the capabilities of semantic matching by measuring similarity. This paper proceeded to propose and implement a parallelized PCM, to parallel computing similarity of concept and web page for ontology mapping. The result of the test indicates that EWO can process practical Ecommerce concepts, but cannot totally express the semantic information of the commercial web pages. We will work from entity's structure of thinning EWO concept further, and the comprehensive characteristic of constructing the color, texture and shape will be set about.

REFERENCES

[1] Marc Ehrig, York Sure, *Ontology Mapping-An integrated Approach*, 6-7, 11-13.

[2] H.Takeda, K.Iwata,et al, “An Ontolgoy-based Cooperative Environment for Real World Agents, Proc”. *Int’l Conf. Multi agent Systems*, Dec.1996, 353-360.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q.Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D.Petkovic, D. Steele and P. Yanker, “Query by image and video content: The QBIC system,” *IEEE Computer*, no. 28, pp. 23–32, September 1995.

[4] Y. Rui, T.S. Huang, S. Methrotra and M. Ortega,“A relevance feedback architecture for content-based multimedia information retrieval system,” *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997: 82–89.

[5] Miller, G. “Nouns in WordNet: A Lexical Inheritance System”. *International Journal of Lexicography*, Vol.3, No.4, P245-264, 1990.

[6] A. Pentland, R.W. Picard, and S. Sclaroff, “Photobook: Tools for Content-Based Manipulation of Image Databases”, *International Journal of Computer Vision*,18(3),pp.233-254,1996

[7] A. Gupta and R. Jain, “Visual Information Retrieval,”*Comm. ACM*.vol. 40, no. 5, pp.70-79, May 1997.

[8] M. Flickner, H. Sawhney,W. Niblack, J. Ashley, Q. Huang, B. Domet al. “Query by Image and Video Content: The QBIC System” ,*IEEE Computer*, vol. 28, no. 9, 1995.

[9] J.R. Smith and S.F. Chang, “VisualSEEk: A Fully Automated Content-Based Image Query System,” *Proc. ACM Multimedia*,pp. 87-98, Nov. 1996.

[10] John R. Smith and Shih-Fu Chang. “Automated binary texture feature sets for image retrieval”. In Proc. *IEEE Int.Conf.Acoust.Speech,and Signal Proc*,May 1996.

[11] H.Tamura,N.Yokoya. “Image Database Systems: A Survey”,*Pattern Recognition*, 17(1),pp. 29-43,1984.

[12] A.Laine and J.Fan, “Texture classfiication by wavelet packet signatures,” *IEEE Trans Pattern Analysis and Machine Intelligence*, vol.15, no.11, pp.1186-1191, Nov.1993.

[13] T.Chang and C.C.Jay Kuo,”Texture analysis and classification with tree-structured wavelet transform,” *IEEE Trans. On Image Processing*, vol.2, no.4, pp.429-441, 1993.

[14] J.Barnett. “Computational methods for a mathematical theory of evidence”. In Proc. Of the 7th Int. *Joint Conference on Artificial Intelligence*, pages 868-875, Vancouver, Canada, 1981.

[15] J.Gordon and E. Shortliffe. *The Dempster-Shafer Theory of Evidence*. In Rule-Based Expert System, pages 113-138. Addison-Wesley Publishing Company, 1984.

Wenfang Yu is an associate professor in College of Computer and Information Engineering, Zhejiang Gongshang University. She is dean of computer application department. Her research interests are in computer application, database technique and ecommerce.

Research of the Object-Relational Mapping Based on NHibernate Framework＊

Ran Tan, Menghua Xiong
School of Computer Science and Technology, Wuhan University of Technology
Wuhan Hubei, 430063, China
Email: tr_99@163.com, menghua117@163.com

ABSTRACT

Following the computer technology rapid development，in particular network technology swift and violent innovation, now the society has been quietly sent in the information age. When extremely emphasizing the use of distributional enterprise calculation platform, we must protect the data that are object instances at any price and make them to exist forever even in face of the network trouble，memory leakiness，server collapse and other situations. So it has the extremely important practical significance to solve the storing conflict between the persistent objects, and to study the ORM and related mechanism.

Keywords: Object-Relational Mapping, Object Persistence, NHibernate, Relational Database.

1. INTRODUCTION

Most of the high-level programming language provides the abstract means, enables programmers freed from the model of memory allocation, we can fast and stochastically visit the data in RAM(Random Access Memory).The Senior programming language are mostly Object-oriented model, it has abstracted the real world territory well and has realized the nature description to the real world data which is easy to understand. But the RAM is a temporary memory. If need to cause the object lifetime to surmount the period of the running of the program, then must get support from exterior memory to cause the data to become permanent data.

Many enterprise-level applications request to make the service data be permanent storage and the state of object still be retained in order to be used at the next execution time or be shared between different application procedures, even if they are terminated. On the other side, the database development is especially important in the enterprise-level application and it generally contains two kinds of strategies which are the database based on object and the database based on relationship. Speaking of the present stage, the technology of database based on object is insufficiently mature at present and the relational database is still the first choice to the majority software system data storage. In summary, the development of application software will inevitably face the problem of the data persistence which means how to enable each kind of data to be preserved for a long time and be used continuously, so it is a very important question to design a reasonable persistent data layer.

The application system is always divided into three layer in the enterprise-level development, persistent layer, business layer and browser layer. The persistent layer wraps the behavior that is used to make object be persistent; The business layer carries the business on the entity which is from persistent layer according to the service logic and assembles the data to be transmission object for the representation layer; The browser layer will show the users the data that is from the business layer. The NHibernate framework studied in this paper is the tool of persistence for .NET development which can directly access the mapping object to realize data storage and needn’t to consider the database anymore.

2. NHIBERNATE

Working with object-oriented software and a relational database can be cumbersome and time consuming in today's enterprise environments. NHibernate is an object/relational mapping tool for .NET environments. The term object/relational mapping (ORM) refers to the technique of mapping a data representation from an object model to a relational data model with a SQL-based schema.

NHibernate not only takes care of the mapping from .NET classes to database tables (and from .NET data types to SQL data types), but also provides data query and retrieval facilities and can significantly reduce development time otherwise spent with manual data handling in SQL and ADO.NET.

NHibernate's goal is to relieve the developer from 95 percent of common data persistence related programming tasks. NHibernate may not be the best solution for data-centric applications that only use stored-procedures to implement the business logic in the database, it is most useful with object-oriented domain models and business logic in the .NET-based middle-tier. However, NHibernate can certainly help to remove or encapsulate vendor-specific SQL code and will help with the common task of result set translation from a tabular representation to a graph of objects.

3. FRAMEWORK ANALYSIS

NHibernate is a .NET-Oriented Object-Relational Mapping framework which provides a complete ORM framework that is more flexible than the generic ORM and the designer can completely implement the relational database just like an object one. Now a very high-level view of the NHibernate architecture is as follows:

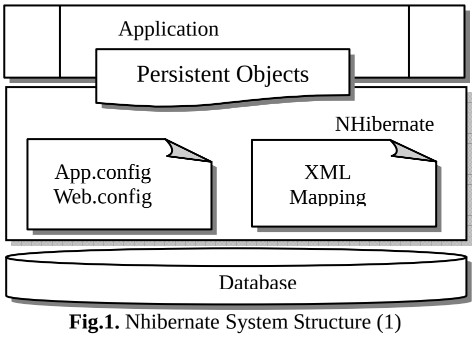


Fig.1. Nhibernate System Structure (1)

This diagram shows NHibernate using the database and configuration data to provide persistence services (and

＊ The work has been supported by the Natural Science Foundation of China Under Grant No.60572015

persistent objects) to the application. A more detailed view of the runtime architecture should be shown. Unfortunately, NHibernate is flexible and supports several approaches. Here two extremes will be shown . The "lite" architecture has the application provide its own ADO.NET connections and manage its own transactions. This approach uses a minimal subset of NHibernate's APIs:

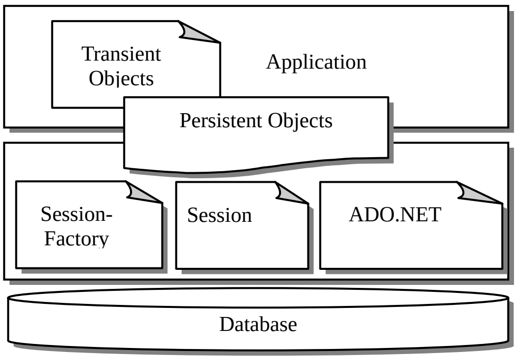


Fig.2. Nhibernate System Structure(2)

The "full cream" architecture abstracts the application away from the underlying ADO.NET APIs and lets NHibernate take care of the details.

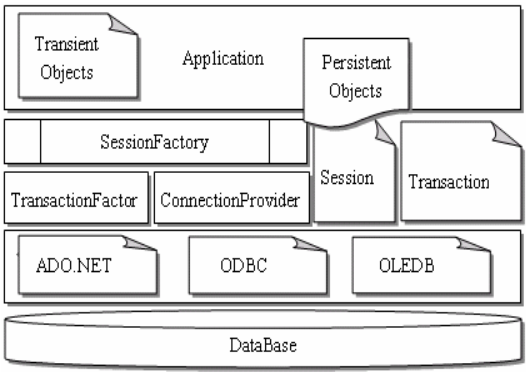


Fig.3. Nhibernate System Structure(3)

Here are some definitions of the objects in the diagrams:
ISessionFactory:
A threadsafe (immutable) cache of compiled mappings for a single database. A factory for ISession and a client of IConnectionProvider. Might hold an optional (second-level) cache of data that is reusable between transactions, at a process- or cluster-level.

ISession:
A single-threaded, short-lived object representing a conversation between the application and the persistent store. Wraps an ADO.NET connection. Factory for ITransaction. Holds a mandatory (first-level) cache of persistent objects, used when navigating the object graph or looking up objects by identifier.

Persistent Objects and Collections:
Short-lived, single threaded objects containing persistent state and business function. These might be ordinary POCOs, the only special thing about them is that they are currently associated with (exactly one) ISession. As soon as the Session is closed, they will be detached and free to use in any application layer (e.g. directly as data transfer objects to and from presentation).

Transient Objects and Collections:
Instances of persistent classes that is not currently associated with an ISession. They may have been instantiated by the application and not (yet) persisted or they may have been instantiated by a closed ISession.

ITransaction:
A single-threaded, short-lived object used by the application to specify atomic units of work. Abstracts application from underlying ADO.NET transaction. An ISession might span several ITransactions in some cases.

IConnectionProvider:
A factory for ADO.NET connections and commands. Abstracts application from the concrete vendor-specific implementations of IDbConnection and IDbCommand. Not exposed to application, but can be extended/implemented by the developer.

IDriver:
An interface encapsulating differences between ADO.NET providers, such as parameter naming conventions and supported ADO.NET features.

ITransactionFactory:
A factory for ITransaction instances. Not exposed to the application, but can be extended/implemented by the developer.

4. THE DESIGN AND REALIZATION OF NHIBERNATE

According to the analysis and discussion of the NHibernate architecture and it is easy to see the useful realistic value. Here take the customer (Customer) and the warehouse contract (Contract) as an example to introduce the concrete process of the persistent operation realized by NHibernate:

4.1 Programmatic Configuration of NHibernate

Add the quotation of the NHibernate in the respective project and set up a datastore from mappings defined in XML configuration files. The configuration file is named App.config in C/S(Client/Server) structure and Web.config in B/S(Browser/Server) structure and the details of the configuration file(B/S) is as follows:

```
<configSections>
  <section name="nhibernate"…/>
  </configSections><nhibernate>
  <add key="hibernate.connection.provider"
value="NHibernate.Connection.DriverConnectionProvider" />
     ……
  <add  key="connectionStringName"value="SQLServer2005"
/>
</nhibernate>
```

4.2 Establishment mapping files and persistent class

The mapping files are the files named by *.hbm.xml and the content of CustomerInfo.hbm.xml mapping file is as follows:

```
<?xml version="1.0" encoding="utf-8" ?>
   <hibernate-mapping xmlns="urn:nhibernate-mapping-2.0">
   <class    name="Yxck.Modal.CustomerInfo,    Yxck.Modal"
                       table="Customer">
      <id name="Id" column="Id" type="int">
      <generator class="native" /></id>
 <property name="Name" column="Name" type="String"/>
  <property name="Tel" column="Tel"    type="String"/>
```

```
      ……
  <property name="Cell" column="Cell" type="String" />
  <bag name="Contracts" inverse="true"
          order-by="Modif-yTime Desc"
          cascade="all-delete-orphan">
      <key column="CustomerId" />
  <one-to-many
        class="Yxck.Modal.ContractInfo,Yxck.Modal"/>
      </bag></class></hibernate-mapping>
```

The content of Contract.hbm.xml mapping file is as follows:

```
      ……
<class name="Yxck.Modal.ContractInfo,
                  Yxck.Modal" table="Contract">
   <id name="Id" column="Id" type="int">
      <generator class="native" />
   </id>
<many-to-one
      name="Customer" column="CustomerId" not-null="true"
      class="Yxck.Modal.CustomerInfo,Yxck.Modal"/>
<property name="CargoCategory" column="CargoCategory"
       type="string"/>
<property name="ContractNo" column="ContractNo"
       type="string"/>
      ……
```

The content of CustomerInfo is as follows:

```
      public   class CustomerInfo
      {public CustomerInfo(){}
                private Int32 id;
                private String name;
                …
      public Int32 Id
        { get { return this.id; }
           set { this.id = value; }}
      public String Name
        { get { return this.name; }
           set { this.name = value; }}
      ……
      public ContractInfo CurrentContract
              { get{if (contracts == null)
                          { return null; }
      IEnumerator ie = contracts.GetEnumerator();
      while (ie.MoveNext())
      {if(((ContractInfo)ie.Current).IsCurrent)
        { return (ContractInfo)ie.Current; }}
         return null; }}
```

The content of ContractInfo.cs is as follows:

```
      public class ContractInfo
        { public ContractInfo()
                { IsCurrent = true;}
                private string cargoCategory;
      ……
      public string CargoCategory{
      get { return this.cargoCategory;   }
      set { this.cargoCategory = value; }}
      public Yxck.Modal.CustomerInfo Customer
      {get { return this.customer; }
        set { this.customer = value; }}
```

4.3 Realization of persistent class

The persistent operation can be realized after the establishment of mapping files and persistent class:

Define the Customer object:

```
      CustomerInfo c=new CustomerInfo();
```

Evaluate for Customer object:

```
      ……
```

Define the Contract object:

```
      ContractInfo ct=new ContractInfo();
```

Evaluate for Contractobject:

```
        ……
```

Establish the relation between the object:

```
      ct.CustomerId＝cst;   c.ie.Add(ct);
```

Save the object:

```
//Define the transaction
Using(NhbTransactionUtils trans=new NhbTransactionUtils)
{
      Try{
      Trans.SaveOrUpdate(cst);
      //Save or Update the object
      Trans.Commint();}
      Catch(Exception ex){
      Trans.Rollback();
      ……
      Trans.delete(cst);
```

5. CONCLUSIONS

NHibernate is an object-relational mapping tool for .NET environments, it can help the developers to save the persistent data easily and needn’t spend too much time in choosing how to store or how to configurate, Thus it can greatly enhanced the development efficiency. But O/R Mapping framework also has Limitation, comparing with the relational database that is complete mature in all theory and practice at present, it only has made the partial beneficial supplements and improvement and it does not need to user O/R mapping technology in the aspect of on-line analysis processing and so on. However, its simple usability has brought the enormous convenience to handle the storage of business object model.

REFERENCES

[1] Yang Fuqing; Shao Weizhong; Liu Junfei. “A Study on the Store Technique of Persistent Object [J],” ACTA Electronica SINICA, 1994,(08) .

[2] Zhang Jiaming Zhou Boxin Song Wenzhong. “Translation of Relation Database Schema into OODB Schema [J],” *Journal of southeast university (Natural Science Edition),* 1998,(02) .

[3] Huang Chenliang Gao Jianhua, “Computer Applications and Software [J],” I*mplement ORM Using The Method Of Embedded Polymorphic SQL.*2002, (03).

[4] Scott W.Ambler, Senior Object-Oriented Consultant AmbySoft. Inc, http://www.Ampysoft.com/persistenceLayer.pdf, 1998.4

[5] NHibernate for .NET [EB/OL]. http://www.Hibernate.-org/343.html.

[6] Tan Ran，Lei Jingming and Xue Yanxin．”Infinite Tree Structure Realization in Programming．”*Proceedings of the International Conference on Computer Science & Educati-on*. Xiamen University Press. Jul, 2006, Xiamen, China. P43-46.

Ran Tan （1961- ）is a Associate Professor and a head of CSCW Lab of Wuhan University of Technology, she was born in Wuhan, and got her master degree in Wuhan University of Technology. Her major is in Computer Network Technology, Long Range Data Processing and Software system structure.

Menghua Xiong（1983- ）, is a master graduate student of Wuhan University of Technology. He was born in Nanchang, and his major is in Computer Network Technology.

Research on E-business System with Dynamic Service Composition

Yang Xia[1], Qiaofen Gao[1], Zhao Xu[2]
[1]School of Computer Science and Technology, China University of Mining and Technology
Xuzhou, Jiangsu 221116, China
[2]School of Information and Electrical Engineering, China University of Mining and Technology
Xuzhou, Jiangsu 221116, China
Email: xia-y@163.com

ABSTRACT

In order to improve the web service resource sharing and the collaboration in e-business system, the concept of Dynamic Service Composition (DSC) based framework is put forward. By integrating many simple services into one complicated service, the distributed autonomous resources could be regarded as one logical unit so that more services could be provided and Quality of Service (QoS) could also be improved.

Keywords: Dynamic service composition, E-business, Web service, Daml_s

1. INTRODUCTION

With the development of Internet technology, e-business [1] will mainly process back-end business transactions. These mutual operations mostly intervene among the computer system, business application procedure and the software module. This is called the dynamic electronic business. In the framework of e-business system, the problem of how to get various services from the network needs to be solved by making use of web service.

Web service technology is an application integration approach based on standard. It can encapsulate the information, behavior, the data performance and the business flow without considering application environment. But web services mainly concentrate on static standard of data exchange and service publication, and intend many people to participate in the executable process. So, it brings new challenges to the automatic administration of supply chain: firstly, how to find out correlative web service required by service requester. Corresponding to the e-business, it means how to look for potential provider and consumer; secondly, how can single atomic service be composed into more complex service in order to integrate all shared service resources to provide the customer with better service.

The problem of web service composition is that how can the registered atomic service be composed into complex service by according to the request of service requester. Semantic web technology is needed. The application of semantic web can make users comprehend web service easily, and then e-business services can be more extensively applied.

2. THE DEFINITION OF SERVICE COMPOSITION AND DYNAMIC E-BUSINESS

Service composition [2] uses basic service offered by the system resources, and makes these basic services compose in certain sequence. In addition, it changes the sequence to construct a more advanced new service to satisfy the request of the customer. Service composition is divided into static service composition and dynamic service composition. Static service composition implies that participated resource services, their mutual operation and the mutual rules among them are confirmed in advance. However, dynamic service composition chooses participated resource service on the basis of current state of resource. Because of the change of resource state and service performance, even the same request of the user can result in different basic services that are participated in service composition.

The definition of dynamic e-business puts great emphasis on the integration of B2B and basic establishment. It creates perfect benefit for the inner and the external enterprises by regulating Internet standard and general basic establishment. Actually, it is e-business flow and related system that can flexibly adapt constant variation of business strategies. It reflects the concept of applying dynamic integration under certain condition, and exhibits the real value of e-business.

In the e-business system, enterprises can reconstruct application and business flow with web service. As such, enterprises make full use of the dynamic characteristic of web service to integrate person, flow and information more easily, so the business flow period can be shortened and the reaction speed can be increased; at the same time, it can improve the quality of customer service by using better extensibility to provide key data for more applications and users synchronously. Enterprises will be endowed with better celerity and agility. Therefore, dynamic e-business has the following basic principles [3]:

- The integration of software resources must be loose-coupling form;
- The service interface of software resources must be released publicly and be accessible;
- Transferring messages among procedures should observe open Internet standard;
- It constructs applied procedure by integrating the core business progress, outside software modules and resources;
- Improving usability of grain software resources would make business progress more flexible and individual;
- Reusing outside software resources will decline cost and improve the productive efficiency for service consumer;
- The software can be sold as service.

3. E-BUSINESS SYSTEM BASED ON DYNAMIC SERVICE COMPOSITION

The technology of web service applied in e-business system has solved the contradiction between increasing business activity and current technology. All business activities are encapsulated to independent service, which can be operated on a platform irrespective of operating system and development languages. Thus, on one hand we can encapsulate enterprise application by uniform format to avoid incompatibilities of bottom technology; on the other hand it can provide convenience for developers.

In the paper, a dynamic e-business system structure is introduced which is on the basis of the technology of web service and service composition. It can construct a new web service based on existing services. It implies that it does not reconstruct web service. It creates web service which satisfies the requirement of the customer [4] to adapt the extensive application request of web service.

3.1 Dynamic E-business System Structure

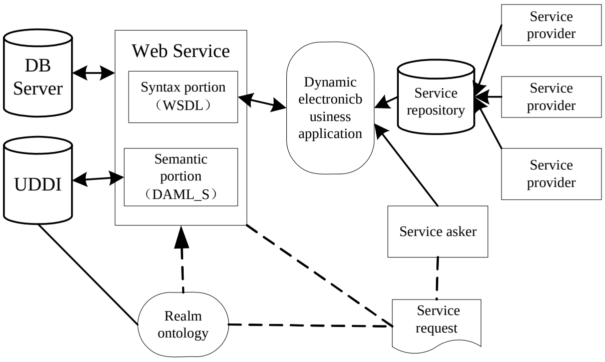


Fig.1. Dynamic e-business system structure

Existing service providers construct service description information with WSDL [5] .These description information supplements ontology portions, and makes them point to special realm ontology in order to complete the semantic description. Web services with semantic description are registered in the UDDI registration center. Service requester puts forward the description of requested service. The requested description equally quotes the same realm ontology. It makes nonstandard request be converted into standard request by appropriate semantic conversion, and applies for the registration center. UDDI uses dual standard of phrasing and semantics to discover needed service. Then it returns search result to service requester, and fulfills binding the special web service [6] according to concerned information with service.

Thus, this system structure can automatically decompose some complex operations (composition service) to logical simple operation. Each simple operation is completed by specific Web service. Through a chain of decomposition-detection-implementation process, we can achieve web integration automatically [7].

3.2 Web Service Discovery

Service requester sends out the request, in other words, utilizing function requirement to describe the work that service needs to complete. If there is service needed by service requester in existing registration service, it is called that existing service matches to the requirement of service [8].

Matching method of existing service mainly has two kinds. One is a method like UDDI using key words to compare; the other is service description with semantics based on OWL_S. It uses knowledge of realm ontology to realize semantic match between Input and Output of descriptions. This is called web service discovery with semantic.

UDDI uses the method of comparing its key word to match service according to service description with WSDL. But this kind of matching method has many shortages. So, many research organizations hope that they can make use of semantic description information and reasoning ability of semantic web to match service.

3.3 Semantic Extension of Web Service

In order to realize semantic web service discovery, we need a service description language and matching arithmetic between existing service and service requirement. Thus, when using service description with semantic information such as OWL_S, we can utilize this semantic information [9] to improve the matching ability through comparing two services accurately.

DAML_S defines three basic ontologies for web service exclusively [10]. Then, it increases semantic description for web service, and solves the shortage of description semantic and performances information with WSDL. We combine WSDL and three basic ontologies to extend current web service description model. In the extense model, it utilizes profile to relate to special realm ontology. Therefore, UDDI can accurately find out the needed web service, and then make it present to service requester. The description model of Web service is shown as the following diagram:

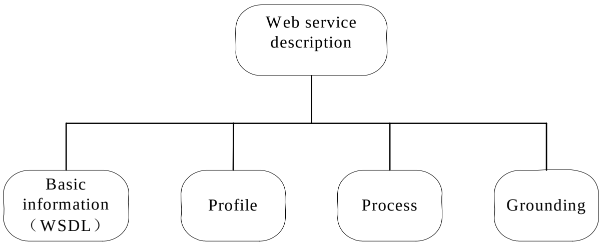


Fig.2. Web service description model

The description of web service with WSDL has great weakness: It only provides syntax layer description, but not semantic layer; It only provides operation layer information, but not concerned semantic information. Owing to the lack of semantic information about service, it brings many difficulties to web service application such as service discovery、service transfer and service composition etc.

3.3.1 Adding Semantic Information to WSDL Description File

The main semantic information that are demanded to join in the WSDL description file includes:

1) The concept of reflecting operation to ontologyEach WSDL file includes a group of operations, and each operation completes a special function. Service discovery needs to confirm whether the searched WSDL file has the requirement operation, and reflects operation to related concept in ontology. The concept is used to describe function of the operation.
2) The concept of reflecting information to ontologyThe information is divided into two species: Input information and Output information. In the WSDL, information is defined by XML schema structure. The information consists of several members, and each member should be reflected to the corresponding concept in ontology. As a result, the semantic of information is not only comprehended by service provider but also all the users who want to use web service.
3) Increasing semantic information method for expressing premise and conclusion.

Each operation should have a set of premise and conclusion.

The premise defines the condition that executing operation must satisfy. But the conclusion defines changed state after executing this operation successfully.

4) Other semantic information

We can also add other semantic information, such as semantic information about service provider and service category semantic information etc.

3.3.2 Semantic Description of Service with DAML_S

The DAML_S is ontology of web service, and it uses DAML and OIL language to realize semantic description of web service. It defines a service class to describe web service from three aspects, including service- profile, service- model and service- grounding. Its emergence makes it possible for automatic discovery, invocation, composition, interoperation and the execution monitoring of web service.

Service- profile is high level description of service. It is used to describe the function and characteristics of service. Users can use it to choose service and to orient service. The most important information among them includes input service, output service, preconditions and post-conditions. Service-model describes how service works. Service-grounding describes how service accesses to, and provides the information on the binding layer. In addition, it can use the mechanisms such as SOAP or the Java RMI to make users access to services.

3.4 Web Service Dynamic Composition Module Structure

In dynamic service composition system, service provider affords services, and releases services to the UDDI registration center. However, service requester directly uses services after searching it in UDDI. The most important information is that we can make some services compose a new service and then use it.

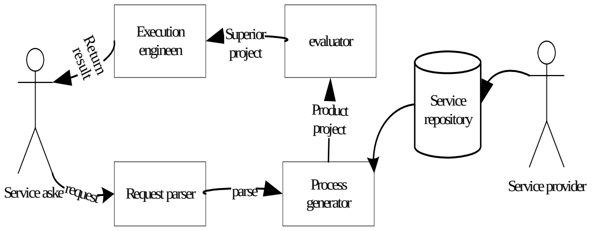


Fig.3. Web Service Dynamic Composition Module Structure

Mainly including five modules [11]:

1) Request Parser: When service requester demands a composition service, it sends a composition request which includes an external specification of composition service (ES means external specification in fig.3).The external specification is the description of composition service, including several aspect specifications such as function, stipulation condition and performance request of composition service.

The mission of Request Parser is to parse this request. It makes request parse into internal specification (IS means internal specification in fig.3) which is comprehended by Process Generator (Process Generator).Finally, It submits internal specification to Process Generator.

2) Process Generator: It produces all feasible composition projects.

3) Evaluator: When Process Generator produces some feasible composition project; it evaluates all composition projects according to certain rule, and selects the superior project.

4) Execution Engine: Evaluator selects the best composition project, and gives it to Execution Engine. Execution Engine executes this composition project and returns the result to service requester.

5) Service Repository: It stores all feasible services in the process of service composition. In fig.3, SS means Service specification.

4. CONCLUSIONS

As above-mentioned dynamic e-business system structure considers semantic restriction, it gains web service in middle course more accurately. The match relationship between front and back Web service will be more clearly so as to perfectly meet the requirements of service requester. Meanwhile, it solves the problem of service discovery and service matching as well. In addition, it can construct a new web service on the basis of existing web service to dynamically construct web services which satisfy the request of the user. It makes e-business service be applied widely, and has super feasibility and research value. With the continuous improvement of web service and dynamic service composition, it will get the extensive application on the market.

REFERENCE

[1] QIN Zheng, YAN Yan, WANG Li. “Dynamic E-commerce based on Web Service” [J]. *Application Research of Computers*. 2003(9):155-157.

[2] NI Wan-cheng, LIU Lian-chen, WU Cheng, LIU Wei. “Practical service grid framework based on dynamic service composition,” [J].*Computer Integrated Manufacturing Systems*.2006, 12(8):1327-1333.

[3] LUO Hong. “The Study and Application of Web Services-based E-commerce” [D].Graduate, Sichuan University 2005.

[4] Sun Kai, Chen De-ren. “Dynamic E-business Model Based on Web Services,” [J].*Computer Engineering and Application*.2003 (26):172-174

[5] Zhu Xiang-liang. “Study and Practice of the Security Mechanism Based on Web Service Electronic Commerce” [D]. Graduate, Chongqing University, 2004.

[6] JIANG Hong, YU Qing-song, GU Jun-zhong. Study on Dynamic E-business System Based on Web Service [J].Computer Engineering.2003, 29(2):195-197.

[7] Qin Xue-jie. “Research and Application of Business Process-Integration Based on Semantic Web Services” [D].Graduate, HeHai University, 2006.

[8] JING Li-qiang. “Service Information Discovery and Integration for E-Commerce” [D].Graduate, Tongji University, 2006.

[9] LIN Qing-ying. “UDDI based semantic web service discovery research” [J].*Computer Engineering and Design*.2006, 27(12):2215-2218.

[10] Wang Feng. “Semantic-based web services integration” [D].Graduate, Southeast University, 2004.

[11] He Zhi-hua. “Research on Technology of Automatic Semantic Web Service Composition” [D].Graduate, Tongji University, 2006.

Yang Xia: Male, 1962.12, Xuzhou Jiangsu China, Ph.D. Candidate, Associate Professor, Post Graduate Student

Supervisor, Study: E-Commerce and E-government, Network and Database Technology, Middleware Technology

Qiaofen Gao: Female, 1981.7, Shijiazhuang Hebei China, graduate student, Study: Web Services Technology, E-Commerce

Zhao Xu: Male, 1955.1, Shuyang, Jiangsu, PRC, Professor, Study: Communication and Information System, Broad Band Access Network Technology, Fiber Optical Communication Technology

VAR Framework for Financial Development and Income Inequality

Renxiang Wang, Jie Gao, Ping Yu
Department of Economics, Wuhan University of Technology
Wuhan, Hubei 430070, China
Email: gaojie129@hotmail.com

ABSTRACT

This paper investigates the relationship between financial development and income inequality for the case of Wuhan city over the period 1993–2005. After considering the time series characteristics of the dataset, a multivariate vector autoregressive (VAR) framework is used as an appropriate specification and the long-run relationship among financial development, income inequality and other key factors is analyzed in a theoretically based high dimensional system by identification of co-integrating vectors through tests of over-identifying restrictions. Our results strongly support the view that financial development and income inequality are mutually causal, that is, causality is bi-directional. These findings suggest the need to accelerate the financial reforms to improve the efficiency of the financial system to stimulate saving/investment and, consequently, narrows the gap between the rich and poor within Wuhan.

Keywords: Financial Development, Income Inequality, VAR, Granger Causality, Economic

1. INTRODUCTION

Vector autoregressive (VAR) models are widely used for multivariate time series analysis, especially in econometrics; see Sims (1980), Lutkepohl (1993, 2001), Reinsel (1993), Hamilton (1994), Hendry (1995), Gourie′ roux and Monfort (1997), Dhrymes (1998) and Clements and Hendry (2002)[1].One reason for this popularity is that VAR models are easy to estimate and can account for relatively complex dynamic phenomena. Important features and applications based on such models include forecasting, causality analysis (in the sense of Wiener, 1956, and Granger, 1969), impulse responses, cointegration, etc[2].

This paper aims to examine the relationship between financial development and income inequality for the case of Wuhan over the period 1993–2005. Specifically, what kind of role has financial sector played to narrow the gap between the rich and poor? What is the nature and direction of the relationship between financial development and income inequality? What kind of effect, positive or negative, has financial development exerted on income inequality? We attempt to answer these questions empirically and try to shed some light on the roles of financial development as well as other conditional variables in determination of income inequality. A theoretically based multivariate VAR framework is used as an appropriate specification and whether proxy measurement of financial development is associated with income inequality is identified in a co-integrating framework through tests of over-identifying restrictions. We find that there exists in Wuhan over the period 1993–2005 a unidirectional causality from income inequality to financial development, results that depart distinctively from those in the existing literatures.

2. ECONOMIC DEVELOPMENT AND EQUALITY

Income inequality is different than poverty. The former represents the distribution of income among the different classes of people (here, within the same country), while the latter describes the percent of the poorest category of the population. Poverty rates are measured by the percent of the population living below the national poverty line (the international poverty line is below $2 a day, and below $1 a day for extreme poverty). Income inequality can be measured in different ways (see Dollar and Kraay, 2000; Ravallion, 2001; Deaton, 2001)[3].The most popular one is the Gini coefficient calculated from the Lorenz curve, other measures (and perhaps more reliable) are quintile shares, median as percent of poorest decile, and richest decile as percent of median.

Some studies argued that more economic development narrows the gap between the rich and poor within a country (Easterly, 2001; Deininger and Squire, 1997; Mo, 2000)[4]. Other work (Amos, 1988; Ram, 1997) found a U-shaped relationship between inequality and growth; thus supporting the aforementioned Kuznets hypothesis. Alternatively, Ravallion (2001) argued that inequality measures (mainly, the Gini coefficient) are not reliable, especially across countries, and therefore would disable any significant finding. He called for more microempirical work on growth and distributional change[5].

Conversely, Bhatta (2001) found that an increase in initial inequality increases subsequent growth in the metropolitan areas of the US. Similarly, Fields (1979) predicted a positive relationship between growth and income distribution.

3. ECONOMETRIC METHODOLOGY-VAR FRAMEWORK

Consider the following VAR model of order P

$$y_t = A_1 y_{t-1} + ... + A_p y_{t-p} + Bx_t + \varepsilon_t \qquad t = 1,2\cdots,T \qquad (1)$$

Where y_t is a k×1 vector of the first-order integrated variables; x_t is a d×1 vector of the first-order integrated variables; ε_t is a vector of normally and independently distributed error terms.

$$\begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{kt} \end{pmatrix} = A_1 \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \\ \vdots \\ y_{kt-1} \end{pmatrix} + \cdots + A_p \begin{pmatrix} y_{1t-p} \\ y_{2t-p} \\ \vdots \\ y_{kt-p} \end{pmatrix} + B \begin{pmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{dt} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{kt} \end{pmatrix} t = 1,2\cdots,T \qquad (2)$$

This can be reparameterised as:

$$A(L)\tilde{y}_t = \tilde{\varepsilon}_t \qquad (3)$$

where

$$A(L) = I_k - A_1 L - A_2 L^2 - ... - A_p L^p \qquad (4)$$

Take income gap in the urban and rural areas (URIG) as an example, it set up VAR model with financial development scale (FDS), financial development efficiency (FDE), urbanization level (CI) and economic degree of opening (XM),etc. other VAR model of income gap index can be constructed in the

* Chenguang Programe of Wuhan City, 20065004116-18.

same way. Among them lag behind in the steps and count p chosen by SC criterion, It is estimated that the result is as follows:

Table 1. The finance development and the urban and rural income gap with VAR model.

Vector Autoregression Estimates

Date: 12/20/06 Time: 20:58

Sample (adjusted): 1995 2005

Included observations: 11 after adjustments

Standard errors in ( ) & t-statistics in [ ]

	URIG	FDS	FDE	CI
URIG(-1)	-0.355095	0.690147	0.075613	-0.028066
	(0.21383)	(1.70165)	(0.18266)	(0.00322)
	[-1.66060]	[0.40558]	[0.41397]	[-8.72503]
URIG(-2)	0.345605	-0.101654	-0.139779	0.020202
	(0.07402)	(0.58901)	(0.06322)	(0.00111)
	[4.66930]	[-0.17259]	[-2.21085]	[18.1442]
FDS(-1)	0.447241	-0.530454	-0.815105	-0.012890
	(0.13915)	(1.10729)	(0.11886)	(0.00209)
	[3.21419]	[-0.47906]	[-6.85786]	[-6.15833]
FDS(-2)	0.337692	-1.021161	-0.414794	0.004380
	(0.18485)	(1.47098)	(0.15790)	(0.00278)
	[1.82686]	[-0.69420]	[-2.62702]	[1.57512]
FDE(-1)	-0.661753	-0.272976	0.363974	-0.009957
	(0.15540)	(1.23662)	(0.13274)	(0.00234)
	[-4.25842]	[-0.22074]	[2.74201]	[-4.25950]
FDE(-2)	-0.423005	-0.937577	-0.682934	-0.050761
	(0.29344)	(2.33516)	(0.25066)	(0.00441)
	[-1.44152]	[-0.40150]	[-2.72457]	[-11.4994]
CI(-1)	-18.26772	28.23892	20.18163	1.140580
	(3.40111)	(27.0652)	(2.90519)	(0.05116)
	[-5.37111]	[1.04337]	[6.94675]	[22.2934]
CI(-2)	11.86965	-15.42831	-18.53864	-0.261916
	(3.42935)	(27.2899)	(2.92931)	(0.05159)
	[3.46120]	[-0.56535]	[-6.32866]	[-5.07716]
C	6.144091	-5.398015	1.191714	0.143480
	(2.18315)	(17.3730)	(1.86483)	(0.03284)
	[2.81432]	[-0.31071]	[0.63905]	[4.36896]
XM	10.33040	6.233894	15.47163	0.719972
	(3.98746)	(31.7312)	(3.40605)	(0.05998)
	[2.59072]	[0.19646]	[4.54240]	[12.0030]
R-squared	0.999100	0.972287	0.999345	0.999990
Adj. R-squared	0.991000	0.722875	0.993448	0.999899
Sum sq. resids	0.000123	0.007781	8.97E-05	2.78E-08
S.E. equation	0.011085	0.088211	0.009469	0.000167
F-statistic	123.3433	3.898309	169.4825	11040.27
Log likelihood	47.10395	24.28834	48.83758	93.26935
Akaike AIC	-6.746173	-2.597880	-7.061378	-15.13988
Schwarz SC	-6.384451	-2.236157	-6.699655	-14.77816
Mean dependent	2.324084	1.051408	0.879685	0.594617
S.D. dependent	0.116844	0.167566	0.116980	0.016622

Determinant resid covariance (dof adj.)	0.000000
Determinant resid covariance	0.000000

This is the estimation result of the finance development and the urban and rural income gap with VAR model.

4. Employ VAR Model to Carry On Granger Cause and Effect Analysis

A very important application of VAR model is to analyze the causality between the array variable of economic time, Granger solves the problem that whether x causes y, mainly sees in what great intensity present y can will be explained whether the lagging value which joined x enabled and explained the intensity higher by the past x. If x is helpful in the prediction of y, or the coefficient correlation of x and y is prominent in counting, can say " y is caused by x Granger ".
In one p steps VAR model :

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} \end{bmatrix} \begin{bmatrix} y_{t-2} \\ x_{t-2} \end{bmatrix} + \ldots + \begin{bmatrix} a_{11}^{(p)} & a_{12}^{(p)} \\ a_{21}^{(p)} & a_{22}^{(p)} \end{bmatrix} \begin{bmatrix} y_{t-p} \\ x_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{5}$$

As and only when all of the are a_{12}^{q} 0, variable x can not Granger cause y, Judge at this moment Granger reason's direct method utilizes F test to test following jointly supposed:

$$\begin{cases} H_0 : a_{12}^{(q)} = 0, q = 1, 2, \ldots\ldots, p \\ H_1 : at \quad least \quad exist \quad a \quad Q \quad make \quad a_{12}^{(q)} \neq 0 \end{cases} \tag{6}$$

Its statistic is:

$$S_1 = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(T-2p-1)} \sim F(p, T-2p-1) \tag{7}$$

Incomplete difference square sum of y equation

$$RSS_1 = \sum_{t=1}^{T} \hat{\varepsilon}_{1t}^2 \tag{8}$$

$$y_t = a_{10} + a_{11}^{(1)} y_{t-1} + \ldots + a_{11}^{(p)} y_{t-p} + \hat{\varepsilon}_{1t} \tag{9}$$

Apply to VAR model, an asymptotic equivalence is examined and can be provided by the following formula:

$$S_2 = \frac{T(RSS_0 - RSS_1)}{RSS_1} \sim \chi^2(p) \tag{10}$$

If S_2 greater than $x^2(p)$, refuses originalsuppose; Otherwise accept original suppose: X can not Granger cause y. Now we employ VAR model to carry on Granger cause and effect analysis according to the above-mentioned step.

4.1 The Income Gap in Urban and Rural Areas and Financial development

According to the actual conditions, among income gap in urban and rural (URIG), finance develop scale (FDS), the finance develops efficiency (FDE) and the urbanization (CI), is there any prominent Granger causality? Its result is as shown in Table 2.

Table 2. The granger causality test of the income gap in urban and rural and financial development

	Original suppose	Statistic χ^2	DF	P
URIG equation	FDS can't Granger cause URIG	12.13904	2	0.0023
	FDE can't Granger cause URIG	29.78945	2	0.0000
	CI can't Granger cause URIG	28.97022	2	0.0000
	FDS,FDE,CI,XM can't Granger cause URIG	339.6851	6	0.0000
FDS equation	URIG can't Granger cause FDS	0.164535	2	0.9210
	FDE can't Granger cause FDS	0.332868	2	0.8467
	CI can't Granger cause FDS	1.092852	2	0.5790
	URIG,FDE,CI can't Granger cause FDS	4.597144	6	0.5964
FDE equation	URIG can't Granger cause FDE	5.181950	2	0.0749
	FDS can't Granger cause FDE	49.53477	2	0.0000
	CI can't Granger cause FDE	55.81440	2	0.0000
	URIG,FDS,CI can't Granger cause FDE	125.3710	6	0.0000
CI equation	URIG can't Granger cause CI	331.1517	2	0.0000
	FDS can't Granger cause CI	44.48497	2	0.0000
	FDE can't Granger cause CI	224.7575	2	0.0000
	URIG,FDS,FDE can't Granger cause CI	1361.091	6	0.0000

The result from Table 4 be able to see the financial development scale (FDS), financial development efficiency (FDE), urbanization (CI) Granger cause the income gap (URIG) in urban and rural areas; In urban and rural areas income gap can Granger cause finance develop scale; Under 8% significance level, in urban and rural areas income gap can Granger cause finance develop efficiency; Income gap and urbanization in urban and rural areas have two-way Granger causality.

4.2 The Income Gaps of Urban Residents And The Financial Development

Set up VAR model with the income gap of urban residents (UIG), the finance develop efficiency (FDE) and the financial development scale (FDS). The urbanization lever(CI) and economic degree of opening (XM) are regarded as the external variable. Its Granger cause and effect analysis result is as shown in Table 3.

Table 3. The granger causality test of the income gap of urban residents and financial development

	Original suppose	Statistic x^2	DF	P
UIG equation	FDS can't Granger cause UIG	0.507170	2	0.7760
	FDE can't Granger cause UIG	0.461997	2	0.7937
	FDS,FDE can't Granger cause UIG	1.985664	4	0.7384
FDS equation	UIG can't Granger cause FDS	4.768960	2	0.0921
	FDE can't Granger cause FDS	1.654488	2	0.4373
	UIG,FDE can't Granger cause FDS	17.85079	4	0.0013
FDE equation	UIG can't Granger cause FDE	12.34876	2	0.0021
	FDS can't Granger cause FDE	4.641420	2	0.0982
	UIG,FDS can't Granger cause FDE	23.15191	4	0.0001

We can see financial development scale,financial development efficiency can not Granger cause the income gaps of urban residents (UIG) from Table 5, its P value is up to 0.78 and 0.79 separately; Under 10% significance level, income gap of urban residents can Granger cause finance develop scale (FDS); and the income gaps of urban residents (UIG) are more prominent Granger reason of the finance develop efficiency (FDE).

5. CONCLUSIONS

Our results strongly support the view that financial development and income inequality are mutually causal, that is, causality is bi-directional. [6]These findings suggest the need to accelerate the financial reforms to improve the efficiency of the financial system to stimulate saving/investment and enlarge the financial development scale, consequently, narrows the gap between the rich and poor within Wuhan.

REFERENCES

[1] Jean-Marie Dufour, Tarek Jouini, "Finite-sample simulation-based inference in VAR models with application to Granger causality testing",in *Journal of Econometrics*,135,2006,pp.229-254.

[2] Qi Liang, Jiang-Zhou Teng,"Financial development and economic growth: evidence from china",in *China Economic Review*,17,2006,pp.395-411.

[3] Ravallion,M,"Growth, inequality, and poverty: looking beyond averages",in *World Development*,29,2001.

[4] Easterly,W, "The middle class consensus and economic development,"in *Joumal of Economic Growth*, 6, 2001, pp.317-335.

[5] Rock-Antoine Mehanna, "Poverty and economic development: not as direct as it may seem",in J*ournal of Socio-Economics*,33,2004,pp.217-228.

[6] Suleiman Abu-Bader, Aamer S. Abu-Qarn, "Financial development and economic growth: The Egyptian experience," in *Journal of Policy Modeling*, 2007.

Jie Gao is a graduate student of department of economics in Wuhan University of Technology. Her major is Financial Engineering and undergraduate course graduated from Wuhan University of Technology in 2005 with Economics. Her research interests are financial engineering and bank supervising.

The Challenges and Key Points on the Successful Application of E-commerce in Cotton Textile Enterprises

Jie Wan, Xin Chen
Wuhan Yinpeng Company, Wuhan, Hubei, China
Email: sincerelyajie@yahoo.com.cn

ABSTRACT

As one of the traditional economic modes, merchandise trade has been developing for many years. Trade companies, to some extent, bear the traditional styles and features. It is rather prevailing for textile enterprises which are restructured from planned economy. The modern information network is developing so dramatically that it is crucial for enterprises to make full use of network resources and choose the appropriate e-trade pattern. For many cotton textile enterprises, their first choice is e-commerce. However, how can they improve their work efficiency and benefit from it in the long run? What are the challenges and focus? This paper elucidates the application and development of e-commerce in Wuhan Yinpeng Cotton Textile Co., Ltd (Yinpeng Company).

Keywords: Cotton Textile, Trade Companies, Traditional, Application, E-Commerce, Tools, Platform, Agent, Competitive Purchase, Competitive Selling, Transaction

1. INTRODUCTION OF MAJOR MODES AND THE APPLICATION OF E-COMMERCE IN COTTON TEXTILE ENTERPRISES

In general, there are three major modes of e-commerce application: B TO B, B TO C and C TO C, among which B TO B is the most commonly utilised mode for online order and procurement conducted between enterprises. The trade volume and value amount for the highest. B TO C means that based on the web platform and Internet technologies, enterprises provide their clients with commodities and services through the website. C TO C is different from the former two modes. The enterprise itself sells no products or goods, instead, it just offers a virtual online platform for the clients to conduct their transactions.

Cotton is often traded in large quantity with large sum of money. Cotton textile industry is rather traditional. At present, the sole successful e-commerce platform for all the cotton textile exchanges is China National Cotton Exchange (CNCE) in the mode of B TO B. CNCE is employed for e-matching, competitive purchase and selling. B TO B and B TO C will be incorporated into the future development and research as main modes. Since 2001, Yinpeng Co., Ltd has been engaged in online exchange in accordance with severe laws and regulations both on the quality of commodities and management, which fully guarantees the credit standing and security of exchange. There are mainly two ways to form prices: auction (a low price offered, and then purchasers/buyers offer their prices, transaction will be down with the highest price) and intense price-competing (In a certain period, all the potential buyers and sellers offer their expected prices which will be recorded by computer in a special way and revealed on the exchange terminal systems. The automatic online service will match the transactions on the basis of the policy “price and time first”). However, the ups and downs of price will put the traders at risk. Therefore, severe rules and regulations on price fluctuation and deposit are made to ensure that traders will live up to their commitment.

2. BRIEF INTRODUCTION OF YINPENG COMPANY AND THE DEVELOPMENT OF ITS E-COMMERCE

Wuhan Yinpeng Cotton Textile Co., Ltd was developed from the former Wuhan Cotton Company which was restructured and re-organised on Dec, 6th in 2001. It is a comprehensive cotton textile corporation whose services consist of cotton purchase, processing, e-commerce, information service, online agent, actuals delivery, etc. So far, Yinpeng Company has got through a long and awkward journey to develop from a traditional company to a modernised company. It is the same case with the development of its e-commerce.

During the planned economy era, the cotton circulation system was not reformed. The cotton purchase, consolidated procurement and price-forming were under the country’s control. Companies had little right to manage their business. They were passive and shoulder little pressure. However, in 1997, with the reform and transition from planned economy to market economy, all the companies were mobilised to reform and take more responsibilities so as not to be washed out by their rivals. At the same time, the government advocated building up a standardised, modernised, well-informed and well-organised exchange market. In 1998, the government declared that a computerised and Internet-connected cotton exchange web, which takes China Cotton Exchange Market as centre and main cotton production areas as foundation, should be formed for market price-forming and information exchange. By the end of 1998, it began to be built. It was first put into use in October 1999, which marks the beginning of e-commerce based on B TO B.

Yinpeng Company has been adapting itself with the reform and development in the area of cotton circulation and CNCE. In 2000, CNCE started to serve for e-commerce. Beijing was set up as the exchange centre, and 22 standing service stations were established in those main cotton producing and selling areas. What’s more, the centre and the branches can serve for the clients at the same time. Under the principles “open, fair, honest and faithful”, the functions of CNCE are as following: to organise exchange, to find price, to avoid risks and convey information. It offers the traders with trade balance, commodities exchange, quality check-out, stocking and transportation, information, consultation, training and some other services. All of the branch stations were at provincial level. In 2001, Yinpeng Company tried every effort to become the first city-level working station. And then, it began to develop e-commerce for cotton exchange based on the mode of B TO B.

Then the working stations were not only committed to do cotton purchase and selling but to provide online exchange agent service for their clients. C-S system was employed; therefore, the only approach for clients to enjoy online service is to log on the website by dialling long-distance call. It was so complex that clients should come to Yinpeng Company when online cotton purchase or selling started. Then Yinpeng Company would deal with the oral declaration forms. Considering that some clients were from afar, the company launched the B-S online agent system, which is the current Online Transaction Matching Agent System, to make it more convenient for clients. Although the B-S system made it possible for clients to do their business by logging on the website, Internet was not available for some clients because it was a fledgling. Yinpeng Company established a mini inter-company network equipped with 8 computers. With the help of the 8 exchange seats, transactions can be carried out in Yinpeng Company's exchange hall with clients' seat numbers. However, it did not make it possible for clients to do business at home. At that time, the decision-makers of Yinpeng Company determined to invest more to build its own wide web, and offer exchange terminals, computers, to clients. All together, more than 30 computers reached the clients and Yinpeng Company adjusted the network for them. Up to then, clients could carry out online exchanges at home in-net dialing. From then on, e-exchange of cotton developed so dramatically in Yinpeng Company that its service spreads all over the country. Yinpeng Company transplanted the system to Wuhan Cotton Network, which fulfilled the ideal that exchanges can be carried out at home.

In June 2004, cotton futures agreement was listed in Zhengzhou Exchange. Yinpeng Company renewed its system on Wuhan Cotton Net. Futures information and agent system of warrants trade which improved the exchange ability of deferred actuals on Wuhan Cotton Network.

On the whole, Yinpeng Company has overcome a lot of difficulties and challenges in the development and application of e-commerce. Although there was no advanced technology to depend on, during the first few stages, the wise decision-makers and the young staff's enthusiasm and innovation conquered numerous difficulties.

3. THE CHARACTERISTICS OF THE DEVELOPMENT OF E-COMMERCE IN YINPENG COMPANY

It's really not easy for Yinpeng Company to develop quickly in a few years from an enterprise that has run traditional business for decades into a modern one that applies the most modern means of information, e-commerce to operate very traditional agricultural products. In fact, the uneasiness of e-commerce in Yinpeng Company does not mean the company has realized an amazing technology. It refers to the innovation of this operating thought and the uneasiness for Yinpeng Company to develop clients and make achievements. Yinpeng Company has overcome the obstacles in clients' operating thought, and fostered clients very patiently. Education was also supplied to clients step by step, transferring them from computer illiterates to frequent users of e-commerce. Gradually a new operating thought and mode has nurtured in the clients' mind. This process is extremely difficult. In the course of this development, the members of Yinpeng Company experienced two leaps. One leap is that the members of Yinpeng Company themselves have jumped out of the traditional operation thought to truly feel and understand this emerging business mode, e-commerce, and to make full use of their innovation continuously that only after their own thorough understanding can others be influenced; the other leap is the development from scratch of a batch of clients through arduous efforts. Their traditional business thinking has been converted. E-commerce mode has been acknowledged, felt and accepted by the clients. Through years of efforts, Yinpeng Company has developed more than 100 clients, providing them with B2B e-commerce services. The company has been taking e-commerce and online trading seriously ever since, although it still runs actuals. The annual trade volume is quite large. The trade volume of 2004 constituted 1/10 of the whole nationwide market. And the client growth rate is also extremely fast.

In 2005, the online transaction volume reached 55,500 tons, the amount of money exchange was 780 million yuan.

In 2006, the online transaction amounted to 62,000 tons, the amount of money exchange was 832 million yuan.

The main revenues of e-commerce in Yinpeng Company are profits of matching from self-operation and agent, the website membership income, and text messaging services.

4. DIFFICULTIES AND PROBLEMS YINPENG COMPANY ENCOUNTERED IN E-COMMERCE DEVELOPMENT

With the establishment of the national network infrastructure, the improvement of network technology, and e-business application software being updated, Yinpeng Company has gradually broken through the technology bottleneck of e-commerce development in its entire process of developing e-commerce.

The unfavourable factors in development at present mainly come from the industry itself, i.e. the development of clients' awareness of e-commerce and the criteria of quality in online transaction(actually it's an issue of integrity). The quality criteria of cotton is rather complex, and the value per transaction is very high. Usually it's purchased by ton. The transaction price per ton is more than 10,000 yuan, and the price differential between different levels of cotton amounts to over 100 yuan per ton. Thus if quality problems arise, the risk is relatively large. That's why in traditional cotton marketing the both sides always take a close look at the cotton, and identify the quality level, and then sign the contract. However, in online transaction, the buyer and the seller cannot personally look at the quality of goods. They can only judge the quality indicators given by network. This requires the membership clients to give firm trust to e-commerce service providers. Meanwhile, cotton business experienced a long era of planned economy. So far the reforms in many cotton ventures have not been completed yet. Thus there is still deeply rooted traditional business thought. It's hard for those ventures to accept the application of e-commerce to cotton industry, a high-value, complex-quality, staple, agricultural product. Besides, the limitation of information, network, and application level, caused by the industry itself, makes the application of e-commerce more difficult. Apart from quality problem as unfavourable, in reality, the traditional way of remittance remains in use due to the big sum of payment. Thus the efficiency is not high. Moreover, the current focus of logistics and distribution is warehousing. In terms of transportation, warehousing and transportation are not efficiently integrated owing to fund and resource problems. Therefore, individual transportation retains (i.e. clients find their own transportation vehicles). In this way the efficiency/cost and security of delivery cannot be protected. In transactions, the contracts used are the traditional fax contract and face-to-face handwritten one. Electronic contracts are not in use. And efficiency is yet to be enhanced. In sum, the difficulties and problems are as follows:

1) Practitioner's understanding of information and application level is a little low;
2) The enterprises' traditional business thought is still deeply rooted. It's very hard to change;
3) There is no scientific, standard quality indicator and identifying method for the online commodities, making it difficult to convince everyone of the quality;
4) The way of cash flow in online transaction is traditional and low-efficiency;
5) Logistics needs to be further integrated; the transportation section needs improvement; and information knowledge needs to be enhanced in logistics, because the over 40 nationwide delivery warehouses cannot communicate effectively, which adds difficulties for the headquarter to supervise each warehouse;
6) The application of electronic contract needs enhancing.

5. THE FOCUS OF E-COMMERCE'S FUTURE DEVELOPMENT IN YINPENG COMPANY

It's been five years since Yinpeng Company developed e-commerce, relying on the national cotton market. In the past five years, Yinpeng Company has not only accumulated rich experience in e-commerce (i.e. cultivating a number of e-commerce operating personnel) but also converted the traditional concept of cotton business, and adapted to the development of the modern information society. Meanwhile, a good information platform has been built— Wuhan Cotton Network (among the first four places inner-industry). Raw cotton purchase and processing plant has been established, which ensures the supply of resources; a logistics centre of Yinpeng Company has been completed as well. At the same time, storage transportation is in application on an information-based management. During the past five years, the development of e-commerce in Yinpeng Company has never stopped. However, as the market develops, competition becomes more and more fierce, the market changes faster and more complex. Consequently Yinpeng Company needs to encounter the challenges continuously in its development.

In Yinpeng Company's further development, the three transaction modes of e-commerce, B TO B, B TO C, and C TO C will be allocated and integrated to the full extent. The transaction system is mainly based on competitive purchase (the buyers set a highest purchase price, then the sellers cut it down, finally the lowest contracts) and competitive selling (the sellers set a lowest price, then the buyers increase it, finally the highest contracts). The transaction commodities include all the raw textile materials like cotton and chemical fibre. The system timely launches irregular trading market according to changes in the market; also an online cotton ordering system is launched, it announces the raw textile materials as commodities on the Internet, then selected independently by the buyers. Yinpeng Company, as the provider and organizer of resources, in addition to its own resources, organizes various resources (resource agents) to enter the cotton ordering system, and provides the most comprehensive variety of resources to buyers, who order directly through the network. Yinpeng Company then organizes the quality identifying and the contract signing. The execution, clearing, and delivery to final clients are all included in the company's one-stop service. Orders can be accepted around the clock. Clients can view, search and order the raw textile materials provided at any time by logging onto the company's website.

6. CONCLUSIONS

After the research and exploration of e-commerce development in Yinpeng Company, the cotton textile enterprises' problems lie in the following focuses and difficulties, according to the market environment of e-commerce development and technology conditions of hardware and software, and in terms of the characteristics of cotton textile industry:

1) The industry in which the enterprise belongs to and its own characteristics need to be clarified, as well as the market and the macro environment it stands in. A careful analysis of features of properties the venture operates needs to be made, as well as the efficient way to find the attributes of digitalized commodities. Analysis of the features of venture and market is also to be done so that the best e-commerce mode can be made finally.
2) Considering the venture's characteristics, fully analysing and using the existing industry tools and e-commerce platform is to be selected. Meanwhile, new e-commerce tool or platform is to be developed to enhance the competitiveness of enterprises. This is a difficulty in development.
3) In view of the low information level of practitioners in cotton textile industry and the status quo that their information sense is not sensitive, it's important and difficult in the development of e-commerce in cotton textile industry to guide and develop clients to trade electronically using e-commerce platform and tools, to convert their transaction habits, to develop online marketing in a large scale, and to enlarge the scale of clients and transactions in order to cut down operating cost. It also determines its application success and whether there is obvious effect or not.
4) A key factor that determines the application success and development of e-commerce tools and platform is the insurance of the security and stability of a rich and complete variety of transaction resources.
5) Another difficulty affects the development of the venture is the safety of electronic payment and its punctual arrival, as a result of the relatively big sum of money exchange in transaction of cotton textile industry.
6) Another key factor that influences the successful application of e-commerce tools and platform lies in the macro environment of logistics, logistics infrastructure, and the management level of the development of logistics information.
7) The successful application of e-commerce tools and

platform in cotton textile industry also depends on the development and cultivation of e-commerce personnel.

REFERENCES

[1] Fang Meiqi, *An Introduction to E-commerce*, Tsinghua University Press, 1999

[2] http://www.whcotton.com/

[3] Cao Shuilian, Liu Jiagang. “An Analysis of Main Application Modes of E-commerce in China”, *Market Modernization*.

A Trust-Oriented Security Model for Workflow in Business Process*

Shihui Wang [1], Wei Liu[2], Wei Du[2]
[1]Faculty of Mathematics and Computer Science
Hubei University,Wuhan China 430062
[2]College of Computer Science and Technology
Wuhan University of Technology, Wuhan China 430063
Email: wliu@whut.edu.cn

ABSTRACT

Workflow in business process is becoming very popular and is being used to support many of the day to day workflows in large organizations. One of the major problems with workflow management systems is that they often use heterogeneous and distributed hardware and software systems to execute a given workflow. This gives rise to decentralized security policies and mechanisms that need to be managed. Since security is an essential and integral part of workflows, the workflow management system has to manage and execute the workflows in a secure way. At the same time, the stability and security of the virtual enterprise depends on the right balance of trust and distrust. In this paper we provide and discuss a trust-oriented security model that is grounded in business process trust characteristics. Our proposed model allows it to decide which other partners' opinions they trust more and allows partners to progressively tune their understanding of another partner's subjective recommendations.

Keywords: Trust-Oriented Security Model, Business Process, Trust Management, Workflow

1. INTRODUCTION

The business processes of different organizations need to be integrated seamlessly to adapt the continuously changing business conditions and to stay competitive in the global market [1]. To enable such business collaboration, research efforts have been put on improving current workflow technologies for supporting collaborative business processes [2-6]. Web service technology has also emerged partly for this purpose and has been deployed for implementing inter-organizational workflows [7,8]. However, both of them deal with the highest values by only considering certain circumstances, rather than considering users' expectations thus unable to provide potential choices by their bidding evaluation criteria for dynamic refinement later. We consider flexibility is also a very important issue.

The challenge is to be able to make trust assessments in the workflow of business process environment, and this boils down to finding methods for receiving reliable evidence about systems and remote transaction partners in computer networks[4]. A transaction partner can be someone you already know but it can also be someone who is totally unknown and with whom you have never interacted before and never will again.

The remainder of the paper is organized as follows. In section 2 some related works are mentioned and remarks on them are proposed. Section 3 is the introduction of trust-oriented security model. The definitions for formalization of trust-oriented security model and the correlative cases are put out in section 4. And the detailed evaluation and discussions are depicted. Finally the work is summed up and future work is discussed in section 5.

2. RELATED WORK

Public-key infrastructures (PKI) simplify key management and distribution but create trust management problems[9,11]. A PKI refers to an infrastructure for distributing public keys where the authenticity of public keys is certified by Certification Authorities (CA). A certificate basically consists of the CA's digital signature on the public key together with the owner identity, thereby linking the two together in an unambiguous way. The structure of digital certificates is standardized by the ITU X.509 standard [1]. In order to verify a certificate the CA's public key is needed, thereby creating an identical authentication problem. The CA's public key can be certified by another CA etc., but in the end you need to receive the public key of some CA, usually called the root CA, out-of-band in a secure way, an various solutions can be imagined for that purpose.Trust management includes methods for assessing policies regarding issuance and handling of public-key certificates and for determining whether these policies are adhered to by CA and users[12].

There are many ways of describing trust. In the context of e-commerce and IT security we will define trust in principals as the expectation or belief that they will behave according to a given policy and without malicious intent, and trust in systems as the expectation or belief that they are secure and will resist malicious attack. Trust is thus a belief and we assume trust to be based on evidence, experience and perception. In the physical world trust in things and in other people is based on our experience with them, information we have received about them and how they appear to us. All this makes trust a very subjective phenomenon, meaning that I don't necessarily trust the same things or the same people as you and vice versa. The number of people we can potentially relate to within a physical world is also limited by distance and physical constraints. Trust management can be defined as the activity of collecting, codifying, analysing and presenting security relevant evidence with the purpose of making assessments and decisions regarding business environments.

3. THE TRUST-ORIENTED SECURITY MODEL

We use a trust model based on sociological characteristics of trust, as described in previous sections. In particular, our model supports the following properties of social trust, as outlined as follows:

a) Trust is context-dependent.

b) Supports negative and positive degrees of belief of an agent's trustworthiness, although on a short range of values.

c) Trust is based on prior experiences. Agents are able to identify repeated experiences with similar contexts and with the same agents.

* Supported by Wuhan University of Technology Fund (Xjj2005078).

3.1 Direct Trust

We represent an agent's belief in another agent's (a) trustworthiness within a certain context (c) to a certain degree (td) by the following:

t (a, c, td), where td {vt, t, u, vu}.

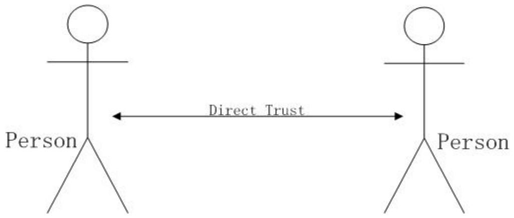


Fig.1. Direct Trust

The semantics for direct trust is given in Fig.1 above. Additionally, we leave the context variable c open so that agents are able to define their own contexts when using this trust model.

3.2 Recommender Trust

The agent may also believe that another agent (b) is trustworthy to a certain degree (rtd) for giving recommendations about other agents with respect to a context (c), represented as:

rt (b, c, rtd).

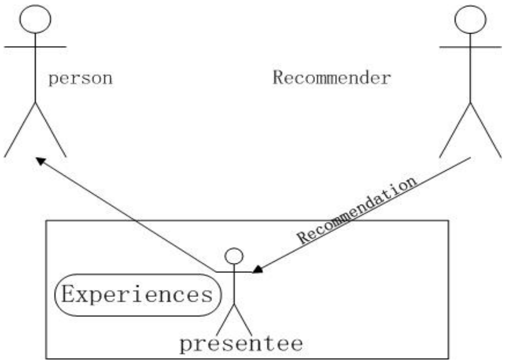


Fig.2. Recommendation Trust

A recommendation need not necessarily represent the belief of the recommending agent. Therefore we assume that recommenders may lie or give out contradictory recommendations to different agents. The value of rtd indicates the 'semantic distance' between the recommendation and rt's own perception of the recommended agent's trustworthiness. In other words, it is a value that is applied to 'what the recommender said' to obtain 'what I think she really means'. For example, the recommender's perception of 'very trustworthy' may only equate to what rt perceives to be 'trustworthy', thus when a recommendation of 'very trustworthy' is made, we can apply the rtd value to obtain 'trustworthy'. As with direct trust, we leave the context variable c open. The semantics for recommendation trust is given in Fig.2 above.

3.3 Data Structure

An agent's opinion about another is based on their previous interactions. An agent x maintains this in two separate sets: the set Q for direct trust experiences and the set R for recommender trust experiences. Assume that C= {c1, …, cn} is the set of contexts known to an agent x and A = {a1, …, an} is the set of agents that x has interacted with (either directly or as a recommender).

Further, assume that the 'grade of outcome' of an experience, e, is a member of the ordered set E = {vg, g,b, vb}, representing 'very good', 'good', 'bad' and 'very bad' respectively.

This is the set for trusted recommender agents. For experiences with recommender agents, the result is different. The goal is to obtain a similarity measure, referred to as the semantic distance, of an agent's recommendation and x's perception of the outcome. As a simple example, if a recommends to x that agent b is 'very trustworthy' with respect to context c, and x's evaluation of its experience with b is merely 'trustworthy' (a grade lower than 'very trustworthy'), then future recommendations from a can be adjusted accordingly. In this example, we say that x's experience with bdowngrades a's recommendation by one (or that the difference is –1). The domain of possible adjustment values is given by the set G = {-3, -2, -1, 0, 1, 2, 3}.

4. THE DEFINITIONS FOR THE TRUST-ORIENTED SECURITY MODEL

We identify a set of relations which represent agents possessing some properties for further inferencing. We do not distinguish role from service for the reason that every role's function is to provide service(s).

Therefore, we may describe the geometric object within a certain range to simplify the relations of these three dimensions which stand for an agent's capability, resource expectation and role to .ll. Here x, y, z are normalised to interval [0, 1]. The endpoint is the extreme case which seldom happens. Only extreme case < 1, 1, 1 > is reasonal be here.

Consequently, an agent, say i is depicted by 3-tuple< Σi, Ψi,i >. We define Γ as the universe of discourse of an agent, Γ = < Σ, Ψ, >. Agents consult each other using performatives of Speech Act such as assertives (informing), commisives (promising), directives (requesting or querying) and declaratives to query the meaning of the concept if it is a part of a shared ontology. Primitives are as follows: intent = {term | term ∈ {performatives, assertives, commisives, directives }}

In terms of TBox (for terminological knowledge of DL) and ABox (for assertional knowledge of DL) of DLs, we have :

Intent (i, j,C1 _ C2) = {< intent(i, j,CI1, CI2 > | i, j ∈ Γ and CI CI2)

intent(i, j, α : C) = {< intent(i, j, αI, CI > | i, j ∈ Γ and αI ∈ CI)

where α is an individual name; C1, C2 denote any complex concepts. I denotes an interpretation which is a pair of a domain I and an interpretation function • I , namely I = (I , • I). The semantics of above formulas can be explained as concept C1 is more speci.c than C2 for the .rst formula, and the second one is about an instance α of concept C [2][8].

4.1 Capability, Resource and Role

The difference of b1 and a1 depicts the range of the capability dimension which. is denoted as the a principal agent's criterion for potential partner(s). They are defined by human users/agents at the present stage. Therefore, some partners will be excluded for their low credits to be able to full the requirements.

The structure of the capability dimension is as follows. It is a k-tuple schema.

AgID (IndusCode, GDes, att1 ,att2 ,. . . ,attk2), (k ≥ 2) where AgID: Agent ID;

IndusCode1: Industry it belongs to. We assume the codes match exactly (assigning 1 for the similarity between them) for two or more organisations to alliance in the wofkflow.

GDes: goal/subgoal description.

att1 ,att2 ,. . . ,attk2, ($k \geqslant 2$): list of its capabilities, $k \in N$, N denotes natural number.

Both similarity assessment and similarity measurement definitions for the other two, namely resource and role, are similar. We omit them in this paper to avoid redundancy.

4.2 Evaluation

Based on the calculation, we assume that all the interested potential partners will be ranked by their numeric values, which are between 0 and 1 by ascending order. By doing so, every interested partner denoted by AgID (agent ID) has a numeric number related to every dimension, namely capability dimension SCP, resource dimension SRS and role dimension SRL respectively, where SCP, SRS, SRL are sets including a group of AgIDs. Partners within the range of values are derived from the intersection of the three sets: = {SCP ∩ SRS ∩ SRL} The following is a brief description of the process for evaluation (see Fig.3 about its state transition):

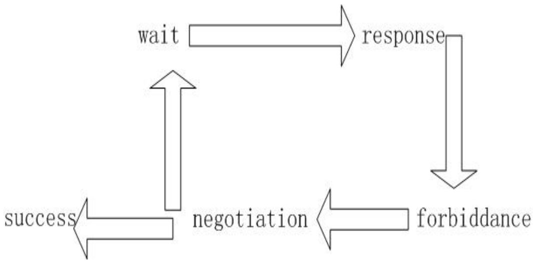


Fig.3. State Transition

(1) derive new ψ from the waiting state every time when it starts a new bidding announcement

(2) if ψ _= Φ then continue bidding, bidding terminates according to principal agent's criteria and followed by negotiation which will lead to a success or waiting state

(3) else (if $\psi = \Phi$) relax constraints, i. e., to let more candidates enter, go to (1)

(4) if still unable to match (such as running out of time limit, no agreement reached so far), just waiting for new entrants, go to (1)

Certainly, "malicious" is the keyword and that definition is reasonable for the domain of security. However, considering the complimentary aspect of competence or correctness is useful.

A person with perfectly good intention or a software component without malicious behaviors can still break things horribly. This is because the person might not be competent to do the job or the software is not "fit for purpose", i.e. it does not behave maliciously but still not correct.

To provide better security for e-commerce, it is desirable that there are mechanisms to support trusting people's competence and trusting software's correct behaviors. These types of trust do exist in normal commerce. High standards of requirements for professional bodies memberships provide the beliefs and expectations that the members are competent to do their job.

4.3 Trust Negotiation Protocol for Workflow in Business Process

The coordination system could be based–at the higher levels–on clauses modeled from the business contract. For instance, if the contract states that partner A shall send to partner B, every Friday evening, a report on the production status regarding some distributed business process, two coordination plans can be defined, one for each partner, in order to automate the fullment of this clause. The coordination plan in partner A would include the activities necessary for gathering the production status data in the right time, formatting these data and sending the report to partner B. On the other hand, a coordination plan running in partner B would include the activities to supervise the arrival, in time, of the expected report and, in case it does not arrive, start some contingency procedure. (supposing K^+ A and $K.^-$A respectively express using public key and private key of n to encrypted and decrypt).

1. B→S :{ Key ID_ B ,{[Reg_B]} } K^+ B

The first step: the virtual enterprise node B sends challenge request. The request information includes Key ID_ B and Reg_B(register sequence number) assigned by autonomy area server S.

2. B→S :{ Key ID_ B ,{[N_s]} } K^+ B

When the server S receives the challenge request from the sender node B. the server S parses the Key ID_ B and Reg_B from the request information. If the Key ID_ B and Reg_B of the request information is the same with them which is saved in server S, the server S will send response of challenge request to client node B. The response information is composed of Key ID_ B and N_s (the N_s is random unsigned integer and unused before).

3. B→S :{ N_s, Key ID_ B ,{[$B_{message}$]} } K^+ s

When the virtual enterprise node A receives challenge response from the server S, the node B parses response information and get Key ID_ B and N_s . If the Key ID_ B in response information is equal to it saved in the node A, the node A saves s N and send information of algorithms and keys negotiation with the node B to server S. the negotiation information includes Key ID_ B , N_s , the node B information, and algorithms and keys of request. If it is different, the node A will drop it.

4. S→B

When server S receives negotiation request from the node B, server S compares the Ns in negotiation request with N s that saved in sever S. if the comparing result is equal, the server S queries information of the node B in its database, else the sever S doesn't process it. If the node B has registered into server S, the server S will send fixed large number to the node B, else the server returns error information to the node B.

4.4 Case Study

Let us take an e-shop, an online book sale, as the example. The precondition of our discussion is that all the agents have the same IndusCode, namely we omit the similarity measurement of the goal/subgoal description by assuming that they are looking for each other in that case (under same business pattern).

E-shop is a good example to study what are involved in the workflow. Suppose it includes several agents such as customer Agents, supplier Agents, e-shop manager A gent, and bank Agents. The supplier (α , where α is an individual name) has a feature C which is a complex concept, denoted by (α :C).

Ask (i,j, Computer Science.trusted Computing Best-Seller.Book); asking if there exists a concept Trusted Computing which belongs to the Computer Science that is implied by an existing book object which is a best-seller one.

5. CONCLUSIONS AND FUTURE WORKS

Trust forms the basis of interaction in any society, including virtual ones. In this paper we looked at the issues of trust in society and outlined a model for supporting trust in virtual communities, which is based on experience and reputation. An example application was then given for illustration. We acknowledge the adhoc nature of certain aspects of the model, namely the trust degrees and the weightings. Future research will attempt to identify a more concrete representation for these metrics.

The most problem of workflow in business process is the security. The challenge is to be able to make trust assessments in the business environment, and this boils down to finding methods for receiving reliable evidence about systems and remote transaction partners in computer networks.

Our proposed model allows it to decide which other partners' opinions they trust more and allows partners to progressively tune their understanding of another partner's subjective recommendations.

REFERENCES

[1] ITU. Recommendation X.509, "The Directory: Authentication Framework. International Telecommunications Union," in T*elecommunication Standardization Sector (ITU-T)*, 1996.

[2] ABA, "Digital Signature Guidelines: Legal Infrastructure for Certification Authorities,"in *American Bar Association*, 1995.

[3] A.Jøsang, "Trust-based decision making for electronic transactions,"in *the Proceedings of the Fourth Nordic Workshop on Secure Computer Systems (NORDSEC'99)*, Stockholm University, Sweden, 1999.

[4] M.Jakobsson and M.Yung,"On Assurance Structures for WWW commerce," in *Proceedings of Financial Cryptography 98*, Springer,1998.

[5] Byrne, J. A., Brandt R. And Bort, O. (1993),"The Virtual Corporation," in *Business Week*,.8,pp36-40,1993.

[6] Y.-H. Tan and W. Thoen,"Towards a Generic Model of Trust for Electronic Commerce,"in *Proceedings, 12 International Bled Electronic Commerce Conference*, Bled Slovenia,1999.

[7] L. Rasmusson and S. Jansson,"Simulated Social control for Secure Internet Commerce (position paper),"in *Proceedings, New Security Paradigms Workshop*, Lake Arrowhead, 1996.

[8] Horrocks, I., Sattler,"U. A Description Logic with Transitive and Inverse Roles and Role Hierarchies," in *Journal of Logic and Computation*,9(3),pp385-410, 1999.

[9] R. Yahalom, B. Klein, T. Beth,"Trust Relationships in Secure Systems - A Distributed Authentication Perspective," in *Proceedings, IEEE Symposium on Research in Security and Privacy,*Oakland,1993.

[10] Oliveira, E. and Rocha, A. P,"Agents' Advanced Features for Negotiation in Electronic Commerce and Virtual Organisation Formation Process,"in *European Perspectives on Agent Mediated Electronic Commerce*, Springer Verlag, 2000.

[11] D. H. McKnight, L. L. Cummings and N. L. Chervany," Trust Formation in New Organisational Relationships," in *Proceedings, Information and Decision Sciences Workshop*, University of Minnesota,1995.

[12] D. H. McKnight, N. L. Chervany. *The Meanings of Trust. Technical Report 94-04*, Carlson School of Management, University of Minnesota, 1996.

[13] U. Maurer, "Modelling a Public-Key Infrastructur," in *Proceedings of European Symposium on Research in Computer Security*, 1996.

Design and Implementation of Database Generation System From UML Class Diagram to Relational Databases in 3NF

Dawei Du, Minghe Huang, Bin Guo, Gaocai Jiang
Software College, JiangXi normal university,
Jiang Xi, Nan Chang, 330022, China
Email:jxnuit@163.com

ABSTRACT

Transforming the object-oriented model of UML into relational database has the wide and important significance when database is in the core position of the software. This paper introduces a system based on the theory of relational database. It has successfully solved three issues: What is the mapping regulation in complicated condition from UML class Diagram to high quality database; how to build the middle model and how to enhance the flexibility of generating codes. And then it also has created the high quality database in 3NF, and through the middle specification in CSQL improved the flexibility and expansibility of the system greatly.

Keywords: theory of relational database, automatic generate, mapping regulation, class diagram

1. INTRODUCTION

With the popularity of UML and the establishment of core position of database in the software architecture, there is an important meaning of transforming object-orientated class diagram model in UML into corresponding high-quality relational database model in 3NF. Reference [1] claimed a simple mapping regulation, and advanced three problems waiting for solving: what is the mapping regulation in complicated condition; how to build the middle model and how to enhance the flexibility of generating codes. By defining the mapping regulation from UML class diagram to the middle specification in CSQL and building the relational database model translator matching 3NF and the automatic code generator, we have successfully designed a system to solve these three issues. This thesis will introduce the design and implementation ideas, software structure, mapping regulation and algorithms of the system. Two methods solving this issue in [2] [3].

2. RELATIVE WORKS

There are two methods solving this issue introduced in [2] [3]. In [2], a simple mapping regulation is mentioned, but it can't build any tools to implement it, and it only maps the objects to the entity patterns that are matched 1NF. In [3], they also only map the objects to the 1NF entity patterns and the middle model they created is not so flexible that transforming it to the high quality database is very difficult.

3. SYSTEM STRUCTURE

In order to generate codes matching multiple database products, the system is divided into three layers as Fig.1 shows.

The first layer: we define the mapping regulation to transfer the UML class diagram to CSQL, a specification describing relation of entities. The second layer: we use a translator to transform CSQL specification to relational model in 3NF, based on the axiom of Amstrong and the algorithm of pattern decomposition. The third layer: we use code generator to transform the relational pattern generated by second layer to corresponding codes according to different DBMSs.

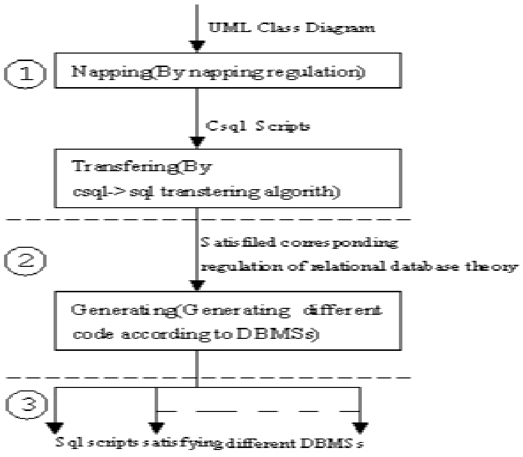

UML Class Diagram
1
Napping(By napping regulation)
Csql Scripts
Transfering(By csql-> sql transtering algorith)
2
Satisfiled corresponding regulation of relational database theory
Generating(Generating different code according to DBMSs)
3
Sql scripts satisfying different DBMSs

Fig.1. System structure

4. DESIGN AND IMPLEMENTATION

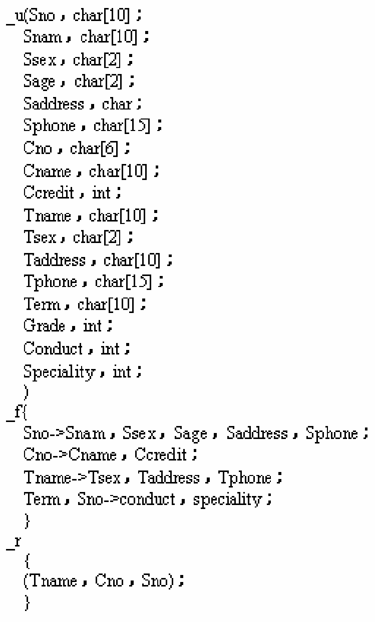
In the whole system, the middle specification is the most important. The mapping regulation from class diagram to it and the transformation from it to relational model are also two key factors.

4.1 The Format and Functions of CSQL Specification

Now we define the syntax of CSQL specification as follows:

1. <S> —> _u （<u_state>） _f{ <u_state>} <s_rear>
2. <S_rear> —> _r {< r_state> } | E
3. <u_state> —> <Id>, <type> ; <u_state> | E
4. <f_state> —> <Ids> —> <Ids> ; <f_state> | E
5. <r_state> —>(<Ids>) ; <r_state> | E
6. <Ids> —> <Id> <Idment >| E
7. <Idment> —> , <Id> <Idment> | E
8. <Id> —> [a-z A-Z]+([<NUM>] | E)?
9. <NUM> —>[1- 9]+[0-9]+
10. <type> —> int <identities> | float | char | datetime|double|datetime|money
11. <identities>—>Identities(<NUM>,<NUM>) | E

The specification has three attributes sets: u, f, and r. Set u records all of the variables and their types of the classes; Set f describes the dependence relationship of variable under normal circumstance; Set r is a special dependence relationship collection, only generated by multiple relationship among classes. The code in Fig.2 is a simple sample of CSQL specification script:



```
_u(Sno , char[10] ;
   Snam , char[10] ;
   Ssex , char[2] ;
   Sage , char[2] ;
   Saddress , char ;
   Sphone , char[15] ;
   Cno , char[6] ;
   Cname , char[10] ;
   Ccredit , int ;
   Tname , char[10] ;
   Tsex , char[2] ;
   Taddress , char[10] ;
   Tphone , char[15] ;
   Term , char[10] ;
   Grade , int ;
   Conduct , int ;
   Speciality , int ;
   )
_f{
   Sno->Snam , Ssex , Sage , Saddress , Sphone ;
   Cno->Cname , Ccredit ;
   Tname->Tsex , Taddress , Tphone ;
   Term , Sno->conduct , speciality ;
   }
_r
   {
   (Tname , Cno , Sno) ;
   }
```

Fig.2. A sample of CSQL

The specification has three fields, _u, _f, and _r, according to three sets. For example, in the _u field, “sno char[10];” means that the set u owns the “sno” attribute, and its type is char[10]. In the _f field, “Cno-> Cname , Ccredit ;” means attribute “Cno” will determine attributes “Cname” and “Ccredit”. In the _r field, “(Tname ,Cno ,Sno)” denotes that the object represented by the attributes of “Tname” , “Cno” and Sno” owns multiple to multiple relationships

According to three sets, the specification has three fields, _u, _f, and _r. For example, in the _u field, “sno char[10];” means that the set u owns the “sno” attribute, and its type is char[10]. In the _f field, “Cno-> Cname , Ccredit ;” means attribute “Cno” will determine attributes “Cname” and “Ccredit”. In the _r field, “(Tname ,Cno ,Sno)” denotes that the object represented by the attributes of “Tname” , “Cno” and Sno” owns multiple to multiple relationships

4.2 Mapping Regulation

The UML specification defines multiple diagrams [4] [5] [6] [7], and we can construct system models in different aspects. The class diagram is one of the diagrams which are used to construct models according to static system designing view. Fig.3 is a simple class diagram:

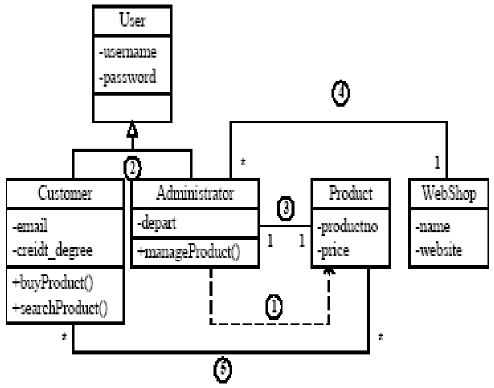


Fig.3. Class diagram

The diagram includes five classes, User, Customer, Administrator, Product, and WebShop. The relationships among classes can be divided into “dependency”, “inheritance”, and “association”. In Fig.3, ①indicates dependency ,② indicates inheritance, ③、④、⑤ indicate associations: ③ is a one-to-one association, ④ is a one-to-multiple association, ⑤ is a multiple-to-multiple association. Our mapping regulation is based on these three relationships.

We define the mapping regulation as follows: let denote class diagram, C denote Class, “attr” denote attribute, and “type” denote the corresponding type of “attr”.

[Regulation 1]

$$\frac{\forall C_i \in \xi}{[\_u]attr_j, type_j; [\_u]} \bigg| \forall attr_j \in C_i$$

In the above diagram the character means the CSQL specification code mapping from $"\forall C_i \in \xi"$ to $"attr_j, type_j"$;[_u] denotes that the mapping results are generated in _u field. $"|\forall attr_j \in C_i"$ means that it is possible to trigger mapping only under condition of $"\forall attr_j \in C_i"$.

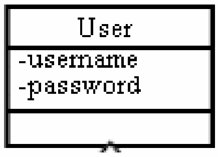


Fig.4. A solitude class

Under Fig.4’s condition, we will add u attributes “username, char[20]; password char[20] ;”in set u .

[Regulation 2]

$$\frac{\forall C_i \in \xi}{[\_u]C_{i\_id}, \text{int} \quad identity(1,1); [\_u]} \bigg| \forall attr_j \in C_i \ \wedge (1)$$

$$[\_f]C_{i\_id} \to attr_j; [\_f]$$

$"\forall attr_j \in C_i \ \wedge (1)"$ means that mapping can only be triggered by the matching conditions of $"\forall attr_i \in C_i"$ and Regulation 1. For example, in Fig.4 we need to add u attributes “User_id, int indentities(1,1) ;” and f attributes “User_id-> username ; User_id -> password ;” based on the regulation 1 and $"\forall attr_i \in C_i"$.

[Regulation 3]

$$\frac{\forall C', C \in \xi \ \wedge \ C' \xrightarrow{*} C}{[\_f]C_{\_id} \to C'_{\_id}; [\_f]} \ | (1) \wedge (2)$$

“ $C' \xrightarrow{*} C$ “ means that C’ inherits from C.

According to Regulation 3, two f attributes are generated: “User_id->Customer_id;” and “User_id->Administrator;”.

[Regulation 4]

$$\frac{\forall C', C \in \xi \ \wedge \ C \ |\!\leftrightarrow\ |C'}{[\_f]C'_{\_id} \to C_{\_id}; [\_f]} \ | (1) \wedge (2) \wedge (3)$$

$$\frac{\forall C, C' \in \xi \ \wedge \ C \ |\!\leftrightarrow * C'}{[\_f]C'_{\_id} \to C_{\_id}; [\_f]} \ | (1) \wedge (2) \wedge (3)$$

$$\frac{\forall C', C \in \xi \ \wedge C * \leftrightarrow *C'}{[\_r](C_{\_id}, C'_{\_id}); [\_r]} \ | (1) \wedge (2) \wedge (3)$$

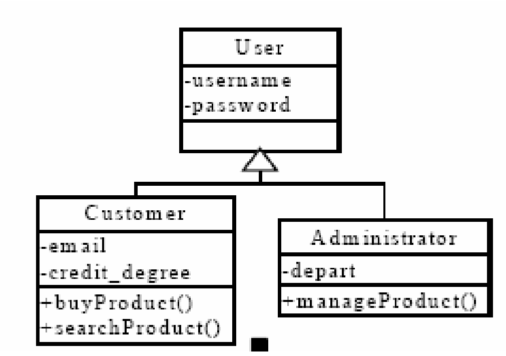


Fig.5. Three classes in inheritance relationship

In this regulation , $C\ |\leftrightarrow|\ C'$ means a one-to-one association relationship between C' and C. $C\ |\leftrightarrow\ *\ C'$ means a one-to-multiple association relationship between C and C'. $C\ *\ \leftrightarrow\ *\ C'$ means a multiple-to-multiple association relationship between C and C'.

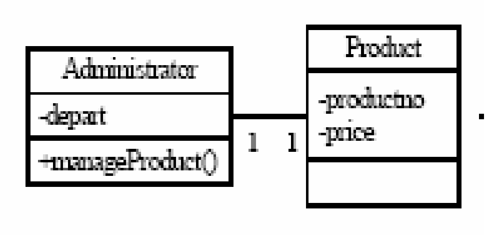


Fig.6. Two classes in one-to-one association relationship

According to 1st rule of Regulation 4, the one-to-one association is transformed to f attributes "Administrator_id -> Product_id;" or "Product_id ->Administrator_id" in Fig.6 .

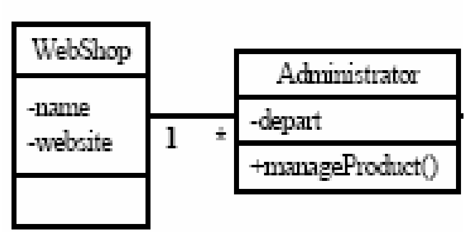


Fig.7. Two classes in one-to-multiple association relationship

According to 2nd rule of regulation 4, the one-to-multiple association is transformed to f attribute "Administrator_id -> WebShop _id;" in Fig.7.

According to 3rd rule of regulation 4, the multiple-to-multiple association is transformed to r attribute "(Customer_id , Product_id)" in Fig.8.

We could combine the description of set f as follow: $\forall a,b,c \in u\ \wedge\ a \to b \in f\ \wedge\ a \to c \in f$ into $a \to b,c$ (u represents the set u ， f represents the set f), making the csql script more concise and easy to understand.

4.3 Transforming Algorithm

After generating the CSQL scripts, we will use CSQL-SQL|3NF translator to build relation pattern matching 3NF.

Steps:

Step 1, By scanning and parsing, put all the attributes and types of set u into a hash table. (Here we call this table as basic information table.)

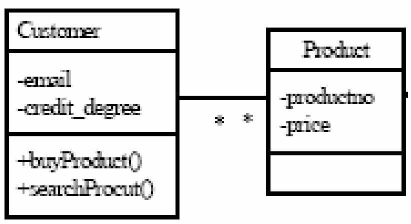


Fig.8. Two classes in multiple-to-multiple association relationship

Step 2, Push the dependence relationships in set f into an adjacent table (we call it as dependence relationships graph), at the same time we separate the left attributes (1st Armstrong axiom [8]).

Step 3, Get every node's closure by using BFS, and then delete redundant attributes and redundant function dependence (2nd and 3rd Armstrong axiom [8]).

Step 4, Divide nodes into 4 categories, N, L, R, LR, and get the prime key by using state space searching to LR node.

Step 5, Take apart every dependence relationship graph into several sub-dependence relationship graph recursively until each one matches regulation 3NF by applying pattern decomposing algorithm [9][10][11].

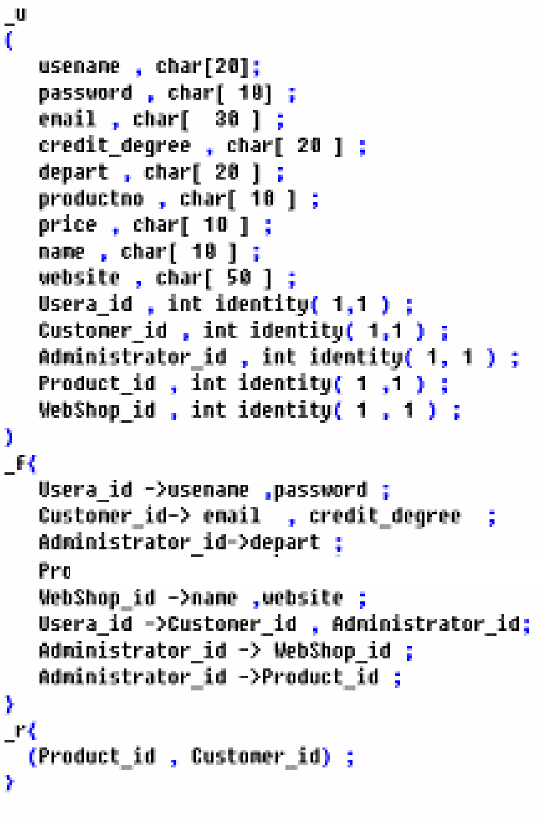
4.4 Generating Codes

Firstly, read the attributes from set r singly, and generate code for building a single table for each r attribute. The prime keys of the tables are all attributes of themselves. Secondly, generate code for sub-dependence relationship graph. Lastly, generate code for building foreign keys automatically. If an attribute has an identity setting in its type, the generator could generate this setting when it is the primary key in the current table.

5. DEMONSTRATION

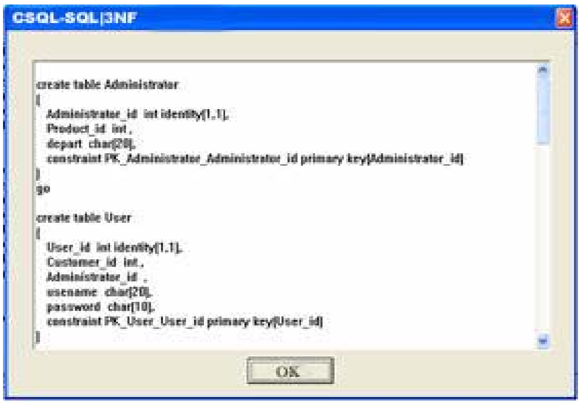
We demonstrate the sample in Fig.3. According to the mapping regulation defined in 3.2 we can get CSQL scripts of that class diagram in Fig.9 easily.

Firstly, load it into our CSQL-SQL|3NF translator, and translate the CSQL script into the relational pattern matching 3NF. Then exchange the name of entities in pattern which have been transformed according to their attributes' names, and generate the SQL code in Fig.10.



```
_u
{
   usename , char[20];
   password , char[ 10] ;
   email , char[  30 ] ;
   credit_degree , char[ 20 ] ;
   depart , char[ 20 ] ;
   productno , char[ 10 ] ;
   price , char[ 10 ] ;
   name , char[ 10 ] ;
   website , char[ 50 ] ;
   Usera_id , int identity( 1,1 ) ;
   Customer_id , int identity( 1,1 ) ;
   Administrator_id , int identity( 1, 1 ) ;
   Product_id , int identity( 1 ,1 ) ;
   WebShop_id , int identity( 1 , 1 ) ;
}
_f{
   Usera_id ->usename ,password ;
   Customer_id-> email  , credit_degree  ;
   Administrator_id->depart ;
   Pro
   WebShop_id ->name ,website ;
   Usera_id ->Customer_id , Administrator_id;
   Administrator_id -> WebShop_id ;
   Administrator_id ->Product_id ;
}
_r{
  (Product_id , Customer_id) ;
}
```

Fig.9. The CSQL code of Fig. 3



```
create table Administrator
(
  Administrator_id  int identity(1,1),
  Product_id  int ,
  depart  char(20),
  constraint PK_Administrator_Administrator_id primary key(Administrator_id)
)
go

create table User
(
  User_id  int identity(1,1),
  Customer_id  int ,
  Administrator_id  ,
  usename  char(20),
  password  char(10),
  constraint PK_User_User_id primary key(User_id)
)
```

Fig.10. Code generating

Set up the database according to the sql code generated above by using the tool attached to our translator automatically.

Finally, we could see our database's relationship graph in SQL Server Enterprise Manager in Fig.11.

The relationship of database that our system has generated not only matches the relationship of 3NF regulation but also satisfies the description of UML class diagram's relation obviously. The object-oriented model of UML is transformed into high quality database truly.

The sample demonstrated above is very simple. Our system could calculate any complicated models accurately and rapidly.

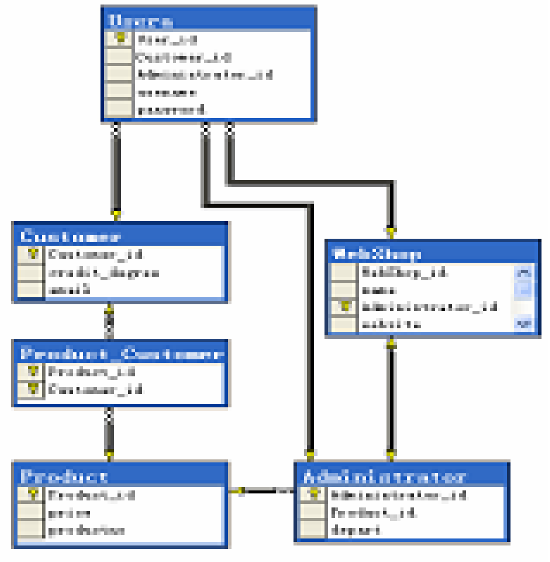


Fig.11. Relationship graphs in SQL Server Enterprise Manager

6. CONCLUSIONS

This system can greatly improve the quality and efficiency when it is used to develop database products. The definition of the middle specification CSQL can improve the flexibility and expansibility of the system significantly. What we will do next is to build a unified model criterion on the basis of CSQL specification, making sure that every model tool like UML can be transformed into this specification, and then the CSQL scripts can be transformed to high quality database subsequently. This can make the CSQL specification independent of any DBMSs, and transparent to any relational database theories, such as Armstrong Axiom, pattern decomposition. As a high-level unified database designing specification, users only need to pay their attention to the relationships between external objects when applying CSQL to design their databases.

REFERENCES

[1] GU Ying ying, GAO. "Jianhua Method of Code Generaion from UML Class Diagram to Tables of Related Database,"Computer Engineering,May 2005.

[2] Wolfgang Keller. "Mapping Objects to Tables – A Pattern Specification,"EA Generali, Vienna, Austria, 1997.

[3] Stefan Mitterdorfer, Egon Teiniker, Christian Kreiner, Zsolt Kov′acs, and Reinhold Weiss." A UML Model to Relational Database Mapping for Dynamic CORBA Component Model Persistency," in Rex Gantenbein and Sung Y. Shin, editors, 17th Intl. *Conference on Computers and Their Applications*, pages 43–48, 2002.

[4] Blaha M, Premerlani W. *Object-oriented Design of Database Application*. *Rose Architect* 1, 2, 1999.

[5] Booch G, Rumbaugh J, Jacobson I. *The Unified Modeling Specification User Guide* SECOND EDITION, Addison Wesley Professional,May 19, 2005.

[6] Stephens M. Automated Code Generation.http://www.softwarereality.com, May 2002.

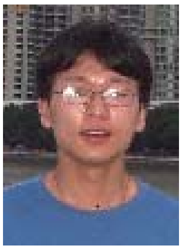
[7] Miliev D. “Automatic Model Transformations Using Extended UML Object Diagrams in Modeling Environment.” *IEEE Trans*. Software Eng, 2002, 28(4),pp.413-431.

[8] Armstrong W W,*Dependency Structures of Data Base RelationShips*,in *Proceeding of IFIP Congress*,1974.

[9] Bernstein P,“Synthesizing Third Normal Form Relations From Functional Depenencies,” *ACM TODS*,1:4,Dec 1976.

[10] Delobel C. “An Overview of the Relation Data Theory,” IFIP, 1980.

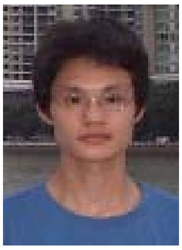
[11] Maier D. “The Theory of Relational Databases.” Rockville Md.: Computer Science Press, 1983.



Dawei Du is a undergraduate student in Software college of JiangXi normal university of China. His current interests include database system, AI, operating system and compiler system.

Minghe Huang is a full professor and a dean of Software College in Jiangxi Normal University, a member of the CPPCC Jiangxi Province Committee, a young and middle-aged academic director in Jiangxi Province. He graduated from Jiangxi Normal University in January of 1982, and then in 1986 he attended an advanced study in the department of computer science of Fudan University. His research interests are in distributed parallel processing, algorithm design and analysis, object-oriented modeling technology, etc. He has published several books and 50 more papers on magazines such as "computer research and development", "computer science", "computer engineering and science" and etc.

Bin Guo is a Lecturer of Software College in Jiangxi Normal University, He graduated from Jiangxi Normal University in July of 1994. He received his Master of Sfotwar Engineering degree from Tongji University in 2005. His research interests are in computer networks and database system.



Gaocai Jiang is a undergraduate student in Software college of Jiang Xi normal university of China. His current interests include database system and network system.

Using Agent and Ontology into automated negotiation system in e-commerce

Qiumei Pu, Qianxing Xiong, Luo Fang, Min Xiao
School of Computer Science and Technology,WuHan University of Technology
Hubei Province, 430063, CHINA
Email: puqm@ whut.edu.cn

ABSTRACT

Negotiation plays a fundamental role in electronic commerce activities, allowing participants and take decisions for mutual benefit. Recently, software agents are believed to be playing an increasing variety of roles in the research of e-commerce. Meanwhile, semantic web and ontology can provide interoperability from syntactic level to semantic one not only for human users but also for software agents. This paper proposed an agent-based system that integrates ontology. The aim of it is to provide a solution to automating negotiation process in the electronic marketplace. In this approach, agents need very little prior knowledge of the protocol, and acquire this knowledge directly from the marketplace.

Keywords: Electronic Commerce, Ontology, Automated Negotiation, Intelligent Agent

1. INTRODUCTION

Recently, software agents are believed to be playing an increasing variety of roles in the research of e-commerce. Informally speaking, software agents help to automate a variety of tasks including those involved in buying and selling commodities over the Internet, offering great flexibility and improved performance. Meanwhile, semantic web and ontology can provide interoperability from syntactic level to semantic one not only for human users but also for software agents. This brings about an historic and revolutionary opportunity to the development of e-commerce.

Negotiation plays a fundamental role in electronic commerce activities, allowing participants to interact and take decisions for mutual benefit. Recently there has been a growing interest in conducting negotiations over Internet, and constructing large scale agent communities based on emergent Web service architectures. The challenge of integrating and deploying negotiation agents in open and dynamic environments is to achieve effective communications.

Domain ontology and knowledge based systems have become very important in the agent and semantic web communities. As their use has increased, providing means of resolving semantic differences among ontology has also become very important. When agents interact, for instance, to cooperate, negotiate or even to compete, they should be able to communicate. In multi agent systems (MAS), languages such as agent communication language (ACL) (FIPA 1997) [1] and knowledge query and manipulation language (KQML) [2] provide the standard for agent communication. The ontology used in the communication are however not standard. Thus ontology negotiation to enable cooperation among agents that are based on different ontology is essential. Ontological interoperability research is crucial to the success of the semantic web effort and multi agent development.

The remainder of this paper is organized as followers. Section 2 presents related work with the subject of this research. Section 3 introduces the architecture of proposed e-commerce negotiation system. Finally, conclusions are drawn and future work is addressed in Section 4.

2. RELATED WORKS

In order to get better effect in the electronic commercial activity and obtain more interests, the buyer (or supplier) hopes to transact with multiple suppliers (or buyers) at the same time and select the best bargaining from each other. The key to realize the multi-agent system cooperate and collaborate is negotiation, and it is a mutual mechanism to build agent communication language.

It can reach a common understanding through negotiation. In addition, negotiation is one of the important stage in the mode of e-commerce field in consumers buying mode of behavior (CBB) [3], it assign limited resources fairly through adopting the market mechanism. It is difficult to realize automated negotiation in e-commerce. At present there are some negotiation system adopts simple protocols. For instance, online auction system named AuctionBot [4] offers a public application program interface in negotiation mechanism. AuctionBot system let users set up the negotiation tactics by oneself and establish its bid way according to one's own partiality, agent will accord with these tactics to negotiation. Some related papers listed as followed:

Tamma et al. (2005) [5]: In this paper the author present the negotiation protocols which are expressed in terms of a shared ontology in contrast to being coded within agents participating in negotiation, thus making this approach particularly suitable for applications such as electronic commerce. In order to permit interoperability, the protocol is defined in terms of a shared ontology of negotiation which provides the basic vocabulary that agents must share in order to discuss the terms of the participation in the negotiation session. In this approach, agents can negotiate in any type of marketplace regardless of the negotiation mechanism in use. Meanwhile, the special-purpose terminology that the knowledge ontology offers can infer negotiation protocol, component and limitation of rule. The ontology is also used to tune agents’ strategies to the specific protocol used. The paper presents this novel approach and describes the experience gained in implementing the ontology and the learning mechanism to tune the strategy.

Grosof and Poon (2004) [6]: This research focus on how to deal with the exception in the automated negotiation protocols which based on the knowledge. The paper brings forward three contributions: (1) It builds upon the situated courteous logic programs knowledge representation in RuleML, the emerging standard for Semantic Web XML rules. Here, it extend the SweetDeal approach by also incorporating process knowledge descriptions whose ontologies are represented in DAML+OIL (the close predecessor of W3C's OWL, the emerging standard for Semantic Web ontologies), thereby enabling more complex contracts with behavioral provisions, especially for handling exception conditions (e.g., late delivery or non-payment) that might arise during the execution of the contract. (2) It integrates knowledge ontology to rule languages and utilize the classification and attribute of knowledge as the description in rulebase. This provides a foundation for representing and automating deals about services － in particular, about Web Services, so as to help search, select, and compose them. (3) According to MIT Process Handbook, the system is the first to combine emerging Semantic Web standards for knowledge representation of rules (RuleML) with ontologies (DAML+OIL/OWL) with each other, and moreover for a practical e-business application domain, and further to do so with process knowledge.

Pokraev et al. (2004) [7]: This research put forward a software framework for the dynamic aggregation of buyers and/or sellers in an electronic market. The framework provides an architecture for automated negotiation between alliances. The author aims at semantic perceive bringing two research questions: (1) Most alliances are formed as a result of a negotiation process between the companies that form an alliance. The format of negotiation messages and semantic description is lack of common understanding. (2)The author believes that negotiation messages are the most important evidence in companies. But most alliances are lack of common understanding on how to automate negotiation. So it allows for the semantic description of negotiation objects and their attributes, and provides a mean for the exchange of negotiation messages unambiguously interpretable by all parties involved. The proposed framework supports ad-hoc alliances by allowing parties with a common interest to negotiate on the proposal they want to make to other market participants first.

In all words, ontology negotiation is becoming increasingly recognized as a crucial element of scalable agent technology [8]. This is because agents, by their very nature, are supposed to operate with a fair amount of autonomy and independence from their end-users. Part of this independence is the ability to enlist other agents for help in performing a task. Negotiation software is better suited to e-commerce transactions than other basic online platforms. Most Internet commerce transactions involve an array of variables, such as market fluctuations, distance from supplier to buyer, availability of materials, and applicable government regulations. These variables can make basic online exchange systems useless in Internet transactions.

3. ARCHITECTURE OF THE AUTOMATED NEGOTIATION SYSTEM

In this paper, we develop a software framework for the dynamic aggregation of buyers and/or sellers in an electronic market. The framework provides an architecture for automated negotiation between alliances. The framework relies entirely on the use of ontologies as a mean of formally representating the knowledge on a particular domain of interest. This provides clear semantics and represents shared understanding of the issues being negotiated on. Our aim is to define a common vocabulary for the agents to communicate with each other. So the communication is facilitated through shared ontology, which defines the concepts and the relations between the concepts of a particular domain. In our work, a shared ontology is used to design the negotiation protocols. The shared ontology provides the basic vocabulary that agents in negotiation should share in order to discuss the issues that can arise in the negotiation. In the open environments like Internet, agents should not be forced to commit to a specific negotiation protocol, but their interaction can be regulated by the shared ontology of protocols.
The system architecture is shown in Fig.1:

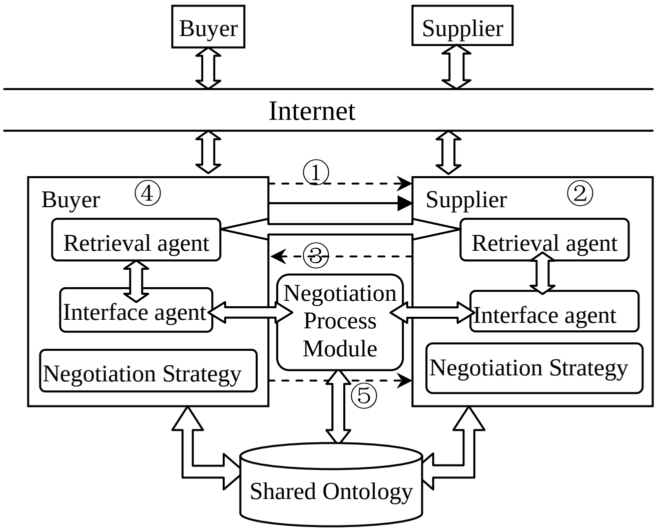


Fig.1. Structure of the e-commerce system

The system work flow is depicted as followed:
①Request: Represent the request and offers; the request of the consumer and the counter offer of the provider are represented as vectors.
②Evaluate: Learn preferences over interactions, requires incremental learning algorithms, evaluate Request and learning; Learn about consumer's preferences based on requests and counter offers; Learn preferences as concept.
③Provide Service or Offer alternative: Possible combination of values associated to the negotiation attributes which represent an expression of will.
④Evaluate the offer: Estimate similarity between the request and available services;
⑤Accept or Re-request: Revise requests or offers based on incoming information

The architecture of the proposed agent-based system is designed to help users buy products from distributed resources based on their interests and preferences. Each agent is autonomous, cooperative, coordinated, intelligent, rational and able to communicate with other agents to fulfill the users' needs.

Fig.1 shows the three types of agent in the proposed three-tier architecture from the sellers' and buyers' perspectives. Interface agents are computer programs that provide personalized assistance to users with their computer-based tasks. Most interface agents achieve personalization by learning a user's preferences in a given application domain and assisting him according to them. In this work the interface agent is a stationary agent that resides on every host machine. It keeps track of a user's profile, interacts with the user and other agents, creates retrieval agents and provides them with parameters, handles incoming retrieval agents and interacts with the user using a graphical user interface. The retrieval agent is a mobile agent that can travel to remote hosts and it is instantiated by the interface agent. When mobile agents are created, they will be dispatched to search for the commodity information from the respective seller (or buyer). It communicates with the interface agent at the remote host where it may engage in negotiation with other agents. The agents communicate using a Knowledge Query Manipulation Language (KQML) like format. The user interacts with the system through a GUI to submit queries and specify requests.

3.1 Interface Agent

The interface agent provides the web interface and implements the bidirectional communication between the buyer and customers. Interface agents are viewed as a facility that allows the user to interact with the online trading environment. Their main responsibility is to fulfill the users' requests by receiving queries and delivering results. The interface agent accepts a query in a form described by a set of words that includes the product of interest and a set of constraints, such as the desired response time. The interface agent obtains the goals from the submitted query and arrives at a solution that best fits the user's needs. The agent’s goals might be locally achievable (i.e. the product is available from the local database) or require interaction with other agents through the generation of retrieval agents. The interface agent acquires and builds user profiles, which are represented as a set of all parameters entered in a transaction.

3.2 Retrieval Agent

The retrieval agent is a buyer agent. It can travel to remote hosts and negotiate on behalf of the interface agent. It should be noted that the negotiation at the remote host destination is constrained by a time limit as specified by the interface agent. When the retrieval agent travels it carries all the required components, including a negotiation model, parameter tables and a sales transaction log. The parameter tables contain all the necessary parameters to perform negotiations. Examples of such parameters include: time constraint, user preferences and preference evaluations. The sales transaction log contains all the transactions made during the negotiation session. It also contains the final transaction if an agreement is reached.

In order to maintain a conversation, agents must have a common language. In our work, a shared ontology is used to design the negotiation protocols (the protocol defines the allowed interactions between the involved parties during a negotiation process over a single or a bundle of knowledge services). The ontology provides the basic vocabulary that either buyer agent or seller agent should share in order to discuss the multi-issues in the negotiation. Based on the ontology, agents need not be forced to commit to a single negotiation protocol, but can choose the negotiation protocol that is most suitable to the type of interaction they participate in. The ontology acts as a general framework that permits agents to interact with each other and reach agreement. This framework defines the rules describing the conditions under which an agreement can be reached and what the agreement offer is. The ontology-based protocol can be defined in terms of some concepts such as Object (the object and/or service the terms of the use of which are under negotiation), which will be specified according to the specific rules of the protocol adopted, and Property, which is associated with Object and filled with the values describing the protocol.

The negotiation ontology provides a general framework that permits negotiation participants to reach agreement by establishing a shared understanding of the negotiation protocols. The ontology is as generic as required for the inclusion of all possible options that may be used in a knowledge service transaction, without excluding further extensions. The idea is based on where the goal is to permit agents to negotiate with most of the negotiation mechanisms, posing as fewer constraints as possible on the agent implementation as possible, in order to ensure flexibility. We have adopted the negotiation ontology in order to decrease its complexity ensuring the appropriate level of expressiveness that a knowledge service transaction requires.

The negotiation ontology specifies the interactions that may occur among participants in a knowledge service transaction and it virtually implements part of the business model that a provider follows. As an example, a “take it or leave it” strategy, where everything is fixed (implemented by a one-step negotiation process) and a multi-attribute negotiation over price, time constraints and user group definition, unveil different business models that can be supported by the negotiation ontology.

4. DISCUSSION AND FUTRUE WORK

The proposed system in this paper provides a new solution to the development of intelligent e-commerce applications. In our work, various information technologies are integrated to automate the Business trading process. It shows the potential toward the future electronic market.

The negotiation ontology specifies the interactions that may occur among participants in a knowledge service transaction and it virtually implements part of the business model that a provider follows. Our future work will be focused on developing some security mechanisms to provide security services for the system [9]. There are many further issues that need to be addressed in the proposed system to be able to introduce a reasonable assistant for its user such as more powerful strategy which has been explored is to introduce the agent in a society of agents.

REFERRENCES

[1] “Foundation for Intelligent Physical Agents”, http://www.fipa.org.

[2] Finin, T., Labrou, Y. and Mayfield, J., “KQML as an Agent Communication Language”, In *Bradshaw J.M. (Ed.) Software Agents*, Cambridge, MA: AAA/MIT Press, pp. 291-316, 1997.

[3] R. Guttman, A. Moukas and P. Maes, “Agent-mediated electronic commerce: A survey,” Knowledge Engineering Review, 13(2) (June 1998) ,pp.147–159.

[4] AuctionBot: <http://auction.eecs.umich.edu>.

[5] V. Tamma, S. Phelps, I. Dickinson and M. Wooldridge, “Ontologies for Supporting Negotiation in E-Commerce,” *Engineering Applications of Artificial Intelligence*, Vol. 18, No. 2, 2005, pp. 223-236.

[6] B. N. Grosof and T. C. Poon, “SweetDeal: Representing Agent Contracts with Exceptions Using Semantic Web Rules, Ontologies, and Process Descriptions,” *International Journal of Electronic*

Commerce, Vol. 8, No. 4, 2004, pp. 61-97.

[7] S. Pokraev, Z. Zlatev, R. Brussee and P. van Eck, “Semantic Support for Automated Negotiation with Alliances,” *Proceedings of the ICEIS 2004: 6th International conference on enterprise information systems*, 2004, pp. 244-249.

[8] D. Fensel, “Ontologies and Electronic Commerce”, *IEEE Intelligent Systems*, January /February, 2001, pp 8.

[9] D. M. Chess. “Security issues in mobile code”, G.Vigna(Ed), *Mobile agent and security, Lecture notes in Computer Science 1419*, springer, Berlin, 1998.

A Reliable Economic Framework in P2P File-sharing Systems*

Qiubo Huang[1], Guangwei Xu[1], Qiying Cao[1]
[1]School of Computer Science and Engineering, Donghua University, Shanghai, China, 201600
Email:huangturbo@dhu.edu.cn

ABSTRACT

In P2P networks, an economic measure is efficient to incentivize nodes to provide services to others. As the payment scheme is essential in the economic measure, an efficient and reliable payment scheme is our goal. In this paper, we propose a mechanism that the peers transfer a file in pieces, and the length of the piece may be 1 MBytes. After getting one piece, the downloader pays an electronic check to the source node. The amount of tokens in the electronic check is for one piece of the file that the source node has just sent to the downloader. The electronic check is signed by the payer's private key and can be redeemed into tokens by the payee. As the transferred unit is one file piece, even if the source node fails during transaction, the downloader can resume downloading from other nodes. As a result the payment scheme can counterwork the accidents. Furthermore, our economic framework can well handle the inflation, the deflation, the sybil attack and the cheating, and is well adaptive to P2P dynamic environment. Therefore, the economic framework is reliable in P2P networks.

Keywords: P2P, File Sharing, Payment Scheme, Electronic Check, Economic Framework.

1. INTRODUCTION

A fundamental difference of P2P applications compared to traditional distributed systems is the fact that decisions of the individual peers are based on their own self-interest and this in principle leads to inefficient system operation. In particular, a rational peer would wish to participate in the system without sharing any resources, and following the so-called 'free riding' strategy [4],[5]. Adar et al. [4] report that nearly 70% of P2P users do not share any file in a P2P community. Instead, these users simply “free ride” on other users who do share. Since the users who are willing to share or provide services to others are few, nearly 50% of all file searching responses come from the top 1% of information sharing nodes. In order to incentivize participating nodes to contribute resources to a global pool, such as to be willing to upload files in file-sharing system, lots of works have been done.

Shneidman et al. in [9] advocate the use of mechanism to make users behave in a globally beneficiary manner. The price-based scheme [1], [2], [3], by tracking each user's resource contribution, aims to do the same.

Golle et al. [3] revealed that, this kind of scheme will motivate peers to share their resources to their maximum extent. As the payment scheme is essential in the price-based scheme, many efforts have been taken, including [1], [2], [11], [12], [13], [14]. In [1] and [2], a distributed accounting system is introduced. [1] and [2] solve the problem to keep account information in decentralized form and transfer tokens(karmas) in transaction. But they have common drawbacks: (1) The transfer is based on whole file transfer and can not adapt to accident, such as computer or network failure; (2) It is hard to realize to avoid fraud such as the source node rejects to provide files after getting the tokens.

Yang et al. [11] and Zou et al. [12] use transferable coins in P2P systems. However in Yang's scheme [11], the performance of the system will be deteriorated when lots of coin owners are off-line. Also, the scheme can not avoid collusions. And in Zou's scheme [12], each coin should be signed reliably, resulting in much cost to generate and validate the signature, so it's time-consuming to realize.

To overcome these drawbacks, we introduce a protocol to transfer tokens (virtual currency) based on transferring files in pieces. In our system, one file is separated into pieces (say 1 MBytes), and the file is transferred piece by piece. After receiving one piece, the downloader pays an electronic check to the source node. The protocol can resume under conditions that the computer and network may fail, and can counterwork attacks. Furthermore, we propose an accusation center to avoid false file or file pollution.

The remainder of this paper is organized as follows: Section 2 introduces the parties in the system, and describes how the nodes can become a member of the system; Section 3 is the most important part in our paper, and describes the payment scheme during the transaction. Section 4 describes how to deal with the inflation and deflation. Section 5 introduces a mechanism to avoid the false file, and section 6 shows how to avoid a sybil attack. Section 7 are our conclusions.

2. OVERVIEW

2.1 The Parties in the System

The system has one centralized registry server, which is responsible for admitting nodes to enter the system. Each node should register first, then get a signed account and the system's public key. The account contains an initial amount of tokens (The tokens are virtual currency, and the amount may be different for nodes at different period according to inflation or deflation of the system.) and is signed with threshold signature scheme [10] by a quorum of Trusted Nodes (TN), who possess one part of the system's private key). Everybody can check the signature but can't counterfeit or tamper the account information, so the account information can be kept by the node himself.

Each node should also have public-private key pair. The private key is used to sign the electronic checks and the public one to check the signature. For convenience, the public key is encapsulated in a certificate and can be distributed to everyone. So the system should have a CA (Certificate Authority) to generate the certificate for nodes. In practice, the CA and the registry server can be the same node.

For each consumption, the node should pay tokens. When the tokens are used up, the node should earn some, as a result the nodes are incentivized to provide services to others. During each transaction, TN (Trusted Nodes) will update the payer and payee's account information. To prevent the payer to use a former account during transaction,

* Supported by “Research of location sense in pervasive computing”, the Ministry of Education Technology Research Key Foundation of China (No. 104086).

we need Account Holders (AH) to keep the updated account information. Compared to the node himself AH need not keep the detailed information. Instead they need only keep the ID, timestamp and token quantity of the corresponding account, and these information will also be updated during the node's transaction.

From above, we can know that, the participating parties in our system are:

1) Registry server. It is only responsible for admitting nodes to enter the system, and doesn't participate in the transaction.
2) CA (Certificate Authority). It generates the public key certificate for nodes. In the following description, the CA and the registry server resides in the same node.
3) Downloader (i.e. Payer). He consumes services and should pay tokens to the provider.
4) Provider (i.e. Payee). He provides services to the downloaders and gets the payment.
5) Trusted nodes (TN). They are a quorum of nodes who are trusted to do something. TN possess one part of the system's private key and are eligible to sign the account information for payers and payees. During the transaction, one of the trusted nodes will be requested by a payer/payee to sign the account information, and he will select several other trusted nodes to sign the account information together. After that, he sends each selected trusted nodes the account information and get the returned partially signed result, and then he combines all the results to acquire a complete signed account information, which will be sent back to the payer/payee.
6) Account Holders (AH). To suit the dynamic characteristics of peer-to-peer networks, AH are a quorum of nodes. They keep the updated account information for a node. AH may be selected according to a fixed algorithm, such as the method, which chooses one's closest nodes as his AH in the Pastry ring [16]. AH can check the validity of an account by checking the ID and timestamp of the account. If the account is invalid, the transaction will not go on.

2.2 Electronic Check

The above parties are essential for the downloader to initiate a transaction. During the transaction, the downloader will not pay tokens in each consumption, instead, he issues the electronic checks to the provider (the electronic check is signed by the downloader, and can pay one piece of file or some pieces). After getting the electronic checks, the provider can redeem them into tokens by sending his request to TN. Therefore the electronic check should not be able to be counterfeited, tampered, or double redeemed. These should be all guaranteed by the public key signature.

The electronic check may contain the following fields (Fig.1):

1) ID is the unique identifier of the check, and the TN can use ID to check the double redemption. ID is a sequence number, and the sequence number in a new check is always one greater than that in the last one when the payer issues the electronic checks.
2) Value shows the value of the check.
3) Payer identifies who issues the check.
4) Payee identifies who is qualified to redeem the check. This can help to prevent others from redeeming the check.
5) Expire is the expiration time for the check, that is to say, the check should be redeemed before the expiration time.
6) Signature is a digital signature generated by the payer's private key and is used for avoiding forgery.

ID	Value	Payer	Payee	Expire	Signature

Fig.1. Electronic check structure

2.3 Initialization

A new node should first become a member of the system, then can start the transactions with others.

When a node N wants to enter the system, it has to generate public-private key pair first, and register in the centralized registry server (which is also the CA). The registration procedure is illustrated in Fig.2, which is interpreted as the following 6 steps:

① Node N generates public-private key pair, and sends the server the registration information which includes the public key.

② CA verifies the validity of the registration information, if node N is permitted to register in, CA will generate an account which includes one number indicates how many tokens are initially issued for node N.

③ CA sends node N's account information to a quorum of trusted nodes.

④ Trusted nodes signed account information with threshold cryptosystem[8] and return the result to CA.

⑤ Trusted nodes inform account holder(AH) of node N the updated account information. In future, AH can check validity of node N's account in transaction.

⑥ CA returns node N's public key certificate and account information which is signed by trusted nodes to node N, and node N stores them for future use.

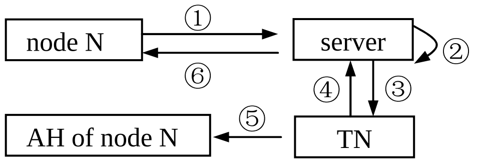


Fig.2. Node registration procedure

2.4 File Advertising

After initialization, one node becomes a member of the system, he then can advertise files, search files, and download files.

Systems will need to provide their own mechanisms for participating nodes to exchange resources, and to agree on a reasonable amount of tokens for the requested resources. In a karma-based file sharing system [1], each file is assigned a fileId through some consistent hashing mechanism. The node closest to the fileId serves as a rendezvous point for people who are offering and seeking that file. A node seeking to download a particular file acquires a list of nodes and initiates an auction by asking providers to submit a karma bid to transfer the file in request. It then selects the lowest bidder, though other alternatives, such as second price auctions are also possible.

To facilitate people to find files, we can make an extension to the method and advertise the files on the web site. Today we can search most of the resources on the web through search engine such as Google. If we post the file on the web site, we can let more people find the files. We can post the files on BBS, or personal Blog, or other web sites. The advertised information may include fileId, price and detailed information for people to read, then people can decide whether the file is what he want or not. Through the fileId, a peer can easily locate the rendezvous node described above, and the peer can get a whole list of nodes who have that file and corresponding prices. Therefore, he can decide to select from which node to download the file.

3. PAYMENT SCHEME

3.1 Payment Protocol

During each transaction, the system will increase the funds in the source node's account and decrease the downloader's funds in his account. To get a reliable payment scheme which can deter fraud and can resume from the network malfunction or computer failure, we propose a payment protocol based on distributed account scheme and transferring file in pieces. The protocol can be illustrated as Fig.3 and the specific process is explained in the following 8 steps.

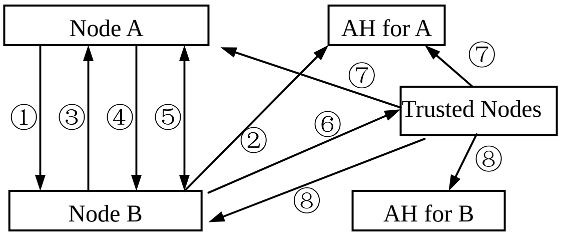


Fig.3. Transaction procedure

① Node A (a downloader) requests some file pieces from node B (the source node).

② According to some algorithm, node B decides whether upload to node A or not. If continue, node B first calculates the total amount of tokens (virtual currency) that node A should pay, then requests AH(Account Holder) of node A to freeze the corresponding amount of tokens in node A's account. Node A's AH checks whether there is enough tokens in node A's account, and then informs B. If the request to freeze is successful, node B continues to step ③.

③ Node B sends node A the file pieces, piece by piece. After finishing one piece, node B waits for A to send B an electronic check (standing for some tokens, and the structure is illustrated in Fig.1) which is signed by A.

④ Node A sends B an electronic check with his signature, and the amount in the check is for one piece of the file that B has just sent to A.

⑤ Repeat step ③ and ④, until some predefined condition (defined in step ⑥) is satisfied, then continue to ⑥.

⑥ If one of the following three conditions satisfies, node B sends requests to the trusted nodes (Trusted nodes are eligible to redeem checks and possess one part of the system's private key) to redeem the checks. 1). 30 seconds elapse since the last redemption; 2). The accumulated tokens in the unredeemed checks reach a predefined amount; 3). All pieces that node A requests have been sent out.

⑦ The trusted nodes validate the electronic checks that A has sent to B. If they are valid, the trusted nodes will ask for A's account information from A's AH and then get this information. After that, the trusted nodes decrease the corresponding amount of tokens shown in the checks from node A's account, and then send back the updated account information with the trusted nodes' signature to node A and A's AH.

⑧ Correspondingly, for B, the trusted nodes ask for his account information from his AH in order to increase the funds in B's account. After doing that, his updated account information with the trusted nodes' signature was sent back to him and his AH.

Note: in step ⑦, the trusted nodes check the validity of the electronic check in three steps:

1) Check the identity of the payee. Through this information validation, the trusted nodes can believe that node B (i.e. the payee) is a legitimate owner of the electronic check.
2) Check the signature. With the identity of the payer, the trusted nodes can acquire the payer's public key certificate, and can check the validity of the signature.
3) Check the double redeeming and the expiration. Each electronic check has a sequence number, i.e. ID. AH of the payer record the maximum sequence number of electronic check that has been redeemed between the payer and payee. If the sequence number of the check to be redeemed is less than that in AH record, the redemption will be rejected. Therefore, redeeming one check twice can be avoidable. To save the AH from maintaining this record forever, each electronic check is given an expiration time. Once the check expires, the AH no longer have to worry about it being re-redeemed and can erase the record of the sequence number.

3.2 Deal with Failures

As node A or node B may fail in the process of transaction, the transaction may become unreliable. Our protocol described above can solve the problem well.

If node A or network fails, node B will redeem checks in 30 seconds. The trusted nodes can use A's public key certificate to validate the checks issued by A, so the redemption process can continue. 1.5 minutes (i.e. three thirty-seconds) after the redemption, A's account will be unfrozen because A's AH will receive no more account update requests. The account unfreezing is for A's future transaction.

If node B fails, node A can still download the remainder of that file from other nodes. And after 1.5 minutes, node A's account will be unfrozen. Node B can redeem the checks anytime before expiration. If he is not able to do it, he cannot redeem the checks any longer. This usually causes a lot of loss to node B if the transaction is based on whole-file transferring. But according to our protocol, the loss becomes much less, i.e. at most the amount of tokens accumulated in 30 seconds transaction.

3.3 More Detail about AH and TN

What we should explain further is that the Account Holder and Trusted Nodes are all a quorum of nodes. The payee can select one trusted node, who selects other trust nodes. So step ⑦ can be described in details as Fig.4.

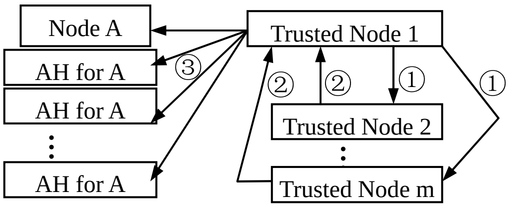


Fig.4. Detailed illustration of Account Holder and Trusted Nodes

In Fig.4, the steps are:

① The trusted node (denoted as trusted node 1) selected by the payee (i.e. node B) selects other trusted nodes (denoted as trusted node 2 to m) to sign the account information together with him. Then trusted node 1 send A's account information to trusted node 2, trusted node 3, …, and trusted node m.

② All the trusted nodes signed the account information and send it back to trusted node 1.

③ Trusted node 1 combines the partial signed account information, and send it to node A and A's Account Holder.

In this mechanism, trusted node 1 masks all the other trusted nodes. To the payer, the payee, and AH, all the trusted nodes act as one node, thus the mechanism simplifies the design and the realization.

4. DEAL WITH INFLATION AND DEFLATION

With time, the total tokens divided by the number of active users varies. It inflates when nodes use up their money and leave the system, and deflates when nodes accrue tokens and leave. If uncontrolled, when a node newly joined in the system, the value of his initially issued tokens is different. When in inflation, perhaps the new nodes cannot download files from other nodes with his initially issued tokens.

To deal with inflation and deflation, Vishnumurthy [1] proposed a method that apply the Correction Factor(ρ) to tokens owned by all nodes at the end of every epoch. The value of ρ is:

$$\rho = \frac{\mathrm{Karma}_{old} \cdot \mathrm{N}_{new}}{\mathrm{Karma}_{new} \cdot \mathrm{N}_{old}} \qquad (1)$$

where Karma_{old} is the total karma at the beginning of this epoch and N_{old} is the total active nodes at the beginning of the epoch. At the end of an epoch, each node in a bank-set transmits to all nodes a message containing: 1) the number of nodes in the bank-set that went inactive in this epoch and their unused karma balance; 2) the number of new nodes that joined the system in this epoch. When a node receives these messages from all nodes in the system, it computes the current number of nodes in the system (N_{new}) and the current total karma in the system (Karma_{new}). Using the previously stored values of Karma_{old} and N_{old}, the node computes the adjustment ρ to be applied, and applies it to accounts for which it is part of the bank-set.

The scheme may have two drawbacks. 1). Because of the distributed nature of the correction, nodes could be in different epochs at the same time. When two such nodes engage in a transaction, appropriate currency conversion should be made to maintain consistency. 2). This scheme needs $O(N^2)$ messages to be transmitted at the end of each epoch, where N is the number of nodes in the system.

We make a change to the scheme, i.e. the Correction Factor(ρ) is not applied to tokens owned by nodes, but to the value initially issued to newly joined node. So, at the end of an epoch, every node transmits a message only to the registry server, instead to all nodes. Then the registry server can calculate N_{new} and Karma_{new}, and then the Correction Factor(ρ). If at previous epoch, the initially issued tokens are T_{init}, then at this epoch, the initially issued tokens will be $\mathrm{T}_{init}* \rho$.

According to the change, we need not any longer worry about that nodes may be in different epochs at the same time. Furthermore, the number of the sent messages is only O(N), not $O(N^2)$, which can greatly save the needed bandwidth.

5. REPUTATION TO AVOID FALSE FILE

There are lots of cheats in our real life, there may be cheats on Internet too. For example, someone may claim to provide a valuable file, however the file he actually provideds is replaced by another file which is not valuable or even valueless. Or with intention or not, one node may deposit polluted files into the file sharing system [15]. To resolve this problem, we propose here an accusation center. If someone has been cheated, he can accuse the provider and may avoid the loss. To realize this, our proposed scheme is as following.

Once the downloader discovers that the file he has downloaded is polluted or replaced by someone, he can supplies the proof that he has downloaded the file from some node, and proves to be valueless to the accusation center. If the proof has been verified, the account of the cheater will be deducted by double revenue which he has earned by providing that file, and the penalty of double revenue may keep back the peers to continue doing the cheating. Furthermore, the penalty will be recorded in the cheater's reputation score. Once a peer's reputation score is under a certain value, he may be excluded from the system.

To help the downloader prove when and from where he has downloaded the file, the provider should transfer more information besides the file itself. For each piece of the file transferred, the information may include (Fig.5):

1) Piece content. It's the file content that should be transferred form the provider to the downloader. It's only a piece of the file, with the length of 1MBytes.
2) File name. It is the name of the transferred file.
3) Provider. This field can identify from where the downloader gets the file piece.
4) Downloader. This field can tell who has downloaded the file piece.
5) Time. This field can identify when the downloader gets the file piece.
6) Sig.. It is the signature generated by the provider, therefore no one else can tamper or counterfeit the information.

Piece content	File name	Provider	Downloader	Time	Sig.

Fig.5. Transferred information with a signature

The Account Holder of the downloader may keep the record of the corresponding transaction. Using the record, the accusation center can confirm that the transaction did occur, and carry out the penalty. The penalty information (i.e. reputation score) is stored in the Account Holder of the cheater. In future transaction, the Account Holder will check the reputation score, to decide whether the peer can continue the transaction or not. If the reputation score is under a certain value, the transaction cannot go on.

6. SYBIL ATTACK PROOF

In a peer-to-peer domain without the external identifiers, any node can manufacture any number of identities [17]. This is a fundamental problem in any P2P system. The use of an external identifier, such as a credit card number or unique processor id, would solve this problem at the loss of privacy. We propose a scheme that distinguishes peers by IP address. For example, the CFS cooperative storage system [18] identifies each node (in part) by a hash of its IP address. When a peer wants to register in the system, the registry server can check his IP address. The registry server may reject his register request if someone has registered from the same IP address in recent **n** days. Here **n** may not be too large, as there are internal networks and DHCP, in which lots of peers have the same IP address. However, in the near future, when all peers are allocated with a unique IPv6 address, **n** can be a larger number.

With this limitation, one node cannot easily manufacture several identities, and he cannot do whitewashing when he used up his money or his reputation is very bad.

7. CONCLUSIONS

In this paper, we propose a payment scheme that owes the following characteristics:

1) Security. First, the account information is kept by the node himself and is signed by the trusted nodes, so no one can tamper or counterfeit it. Second, the electronic check is signed by the payer and nobody else can tamper or counterfeit it unless the secret key is known. Third, we specify the payee of the electronic check, therefore the check can not be redeemed by others, resulting in no theft of the check. Fourth, the account holder of the payer can check whether the check is redeemed twice, therefore there is no need to worry about the repetitive redemption.
2) Adaptivity to P2P dynamic environment. There is no fixed broker or account holder for any node, so, if one of the trusted nodes or account holders is off-line or fails, other nodes can take his place and do his work. If the payer is off-line, the redemption can still accomplish.
3) The payment scheme can well deal with inflation, deflation, sybil attack and cheating.

Having the above features, our payment scheme and the economic framework are reliable and practical in P2P networks.

REFERENCES

[1] V. Vishnumurthy, S. Chandrakumar and E. G. Sirer. Karma: "A Secure Economic Framework for Peer-to-Peer Resource Sharing," *Proceedings of the Workshop on the Economics of Peer-to-Peer Systems*, Berkeley, California, June 2003.

[2] N. Liebau, V. Darlagiannis, A. Mauthe and R. Steinmetz. "Token-based Accounting for P2P-Systems," *Proceeding of Kommunikation in Verteilten Systemen KiVS 2005*, Feb. 2005, pp. 16-28.

[3] P. Golle, K. Leyton-Brown, I. Mironov. "Incentives for Sharing in Peer-to-Peer Network," *Proceedings of the 3rd ACM Conf. on Electronic Commerce*. New York: ACM Press, 2001, pp. 264-267.

[4] E. Adar and B. Huberman. "Free Riding on Gnutella," First Monday, Vol.10, No.5, October 2000.

[5] M. Feldman, C. Papadimitriou, J. Chuang, and I. Stoica. "Free-Riding and Whitewashing in Peer-to-Peer Systems," *Proceedings of 3rd Annual Workshop on Economics and Information Security (WEIS04)*, 2004.

[6] P. Antoniadis, C. Courcoubetis, and B. Strulo. "Incentives for Content Availability in Memory-less Peer-to-Peer File Sharing Systems," *ACM SIGecom Exchanges*, Vol.5, No.4, July 2005, pp. 11-20.

[7] P. Antoniadis and C. Courcoubetis. "Enforcing Efficient Resource Provisioning in Peer-to-Peer File Sharing Systems," *ACM. SIGOPS Operating Systems Review*, Vol.40, No.3, July 2006, pp. 67-72.

[8] Y. Desmedt and Y. Frankel: "Threshold Cryptosystems," *Proceedings of CRYPTO '89*, vol. 435 of LNCS, Springer- Verlag, 1989, pp. 307-315.

[9] J. Shneidman, and D. Parkes. "Rationality and Self-Interest in Peer to Peer Networks," *Proceedings of IPTPS 03*, Berkeley, February 2003.

[10] C. Park, and K. Kurosawa, "New ElGamal Type Threshold Signature Scheme," *IEICE Transactions on Foundational Electronic Communications and Computer Science*, 1996, E79-A (1), pp. 86-93.

[11] B.Yang, and H.Garcia-Molina, PPay: "Micropayments for Peer-to-Peer Systems," *Proceedings of the 10th ACM Conference on Computer and Communication Security*, ACM Press, New York, 2003, pp.300-310.

[12] Zou Jia, Si Tiange, Huang Liansheng, and Dai Yiqi, "A New Micro-payment Protocol Based on P2P Networks," *Proceedings of the IEEE International Conference on e-Business Engineering*, Oct 2005, pp. 449－455.

[13] G. Medvinsky, and B. C. Neuman, NetCash: "A Design for Practical Electronic Currency on the Internet," *Proceedings of the 1st ACM Conf on Computer and Communications Security*, ACM Press, Virginia, 1993, pp. 102-106.

[14] G. Tan, and S. A. Jarvis, "A Payment-based Incentive and Service Differentiation Mechanism for Peer-to-Peer Streaming Broadcast," Proceedings of the 14th International Workshop on Quality of Service (IWQoS. 2006), June 2006, pp. 41-50.

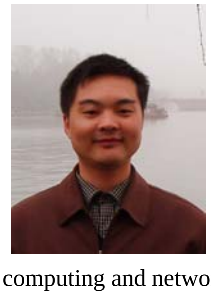
[15] J. Liang, R. Kumar, Y. Xi and K. Ross, "Pollution in P2P file sharing systems," Makki K, Knightly E, eds. Proc. of the IEEE Infocom 2005, Vol.2. Miami: IEEE Press, 2005, pp. 1174-1185.

[16] A. Rowstron, P. Druschel: Pastry: "Scalable, Distributed Object Location and Routing for Large-scale Peer-to-Peer Systems," *IFIP/ACM International Conference on Distributed Systems Platforms* (Middleware), Heidelberg, Germany, November 2001, pp. 329-350.

[17] J. Douceur. "The Sybil Attack," *Proceedings of IPTPS 02*, Cambridge,March 2002.

[18] F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, I. Stoica, "Wide-Area Cooperative Storage with CFS," *18th SOSP*, 2001, pp. 202-215.

[19] Kamvar, S.D., M.T. Schlosser, and H. Garcia-Molina. "EigenRep: Reputation Management in P2P Networks," *The 12th International World Wide Web Conference (WWW-12)*, Budapest, Hungary, 2003.



Qiubo Huang is a full-time teacher of the School of Computer Science and Engineering, Donghua University. He graduated and acquired the doctoral degree from Fudan University in 2003 with the specialty of computer network protocols and distributed systems. His research interests are in distributed parallel processing, grid computing and network security.

B2B Electronic Commerce Platform Based on Mobile Agent Technology*

Xiangzhong Feng
Department of Computer, Zhejiang Ocean University
Zhoushan, Zhejiang, 316000, P.R. China
Email: fxz@zjou.edu.cn

ABSTRACT

As a new software development paradigm, agent technology is being used widely in the development of complex systems. The autonomy and mobility of agent bring many potential advantages to the electronic commerce activities in network environment. To resolve problems of information interaction and sharing between enterprises, mobile agent is used to construct a structure of in B2B electronic commerce system, and every module of the system is designed by using such technologies as JSP, Servlet, JavaBean, JDBC and Aglet, which makes the system characterized by high expansibility, distribution with high transparency, high modularization and reusability.

Keywords: Agent, Mobile Agent, Electronic Commerce, B2B, Aglet.

1. INTRODUCTION

Electronics commerce is any method of using electronic communication and computer technology to conduct business. At present, the most common form of electronics commerce is B2C or business to consumer electronics commerce. These systems are essentially electronic shop fronts to allow businesses to sell goods and services to consumers via the Internet. B2B or business to business systems are designed for businesses to collaborate or sell goods and services to each other. G2B and G2C systems involve the government providing services to business and consumers.

Information on Internet is exploding in an unprecedented speed, with the whole network becoming more and more complicated, so every participant in electronic commerce is to confront more and more uncertainties. Enterprises expect that close relation between them will be established on the Internet and make the delays of trade will be reduced than traditional electronic commerce. Meanwhile, traders also expect the system will be more proactive and provide more individual service which is intellectually adaptive.

How to use the magnitude information on the Internet reasonably and effectively is a keen concern of enterprises. Therefore, the introduction of agent technology, one significant branch of which is the electronic commerce based on agent, brings new opportunities for electronic commerce [1, 2]. In the same case, mobile agent technology brings the new way for development of electronic commerce. Mobile agent is a new distributed calculation technology, which possesses such characteristics as mobility, autonomy and sociality. When this technology is applied, more convenience for Internet resource searching and online automatic trading will be offered so that electronic commerce will have the revolutionary advance.

As a new software development paradigm, agent technology is being used widely in the development of complex systems. Several projects of electronic commerce have been developed based on agent technology. For example, Agent–based Electronics–Marketplace (AEM)[3] is a distributed multi–agent system formed by agent which provides Electronics–Commerce services to end–users within a business context; In [4], agent–based framework for electronic commerce simulation games has been developed by using Zeus, a Java–based multi–agent system developed at British Telecom Lab; In [5], simulated environment for mobile agents is described which allows analyzing the market–based resource control system of D'Agent mobile agent system.

The reminder of this paper is organized as follows. In the next section mobile agent technology and system structure is presented. Section 3 constructs B2B electronic commerce system based on mobile agent. Summary and future works are discussed in section 4.

2. MOBILE AGENT TECHNOLOGY AND SYSTEM STRUCTURE

2.1 Technological Characteristics of Mobile Agent

Mobile agent is a new distributed calculation module. And it's known as a unique distributed exploitation framework, a new generation of software design after process-orientated and object-orientated methods [6, 7]. It is attracting more and more attention of the software technology. Here there are some good reasons for mobile agent is used in electronic commerce [8].

Reduce the network load: mobile agents allow package a conversation and dispatch it to a destination host, where the interactions can take place. Mobile agents are also useful when it comes to reducing the flow of raw data in the network. When very large volumes of data are stored at remote hosts, these data should be processed in the locality of the data rather than transferred over the network.

Overcome network latency: Mobile agent offers a solution overcome network latency, because it can be dispatched from a central controller to act locally and directly execute the controller's direction.

Execute asynchronously and autonomous: Task can be embedded into mobile agent, which can be dispatched into the network then. After being dispatched, the mobile agent becomes independent of the creating process and can operate asynchronously and autonomously.

Adapt dynamically: Mobile agent has ability to sense its execution environment and react autonomously to changes. Multiple mobile agent possess the unique ability to distribute themselves among the hosts in the network so as to maintain the optimal configuration for solving a particular problem.

Sociality: Mobile agent can share information with other entity by some communication and cooperate to fulfill some task.

* This research has been supported by science research foundation of Department of Education of Zhejiang Province for special purpose, 20050136

2.2 System Structure of Mobile Agent

System structure of mobile agent consists of mobile agent and service facilitator of mobile agent (or server of mobile Agent). The service facilitator of mobile agent offers operational environment and service interface for mobile agent and realizes the transference of agents between hosts by ATP (Agent Transfer Protocol). Mobile Agent provides service for the visitors to the service facilitator by the communication between ACL (Agent Communication Language). The structure is as Fig 1 in the following:

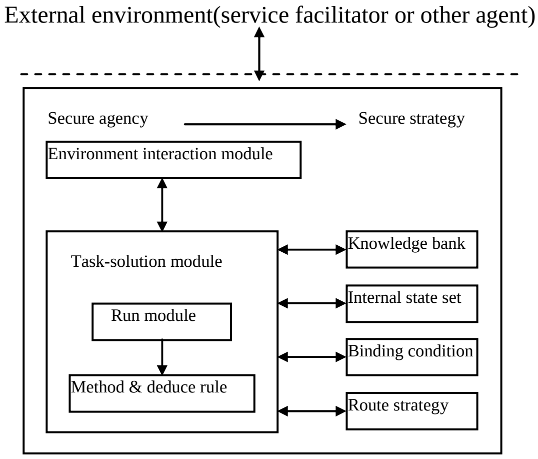


Fig.1. system structure of mobile agent

Mobile agent system structure is composed of secure agency, environment interaction module, task-solution module, knowledge bank, internal state set, binding conditions and route strategies. They are closely connected. Mobile agent first passes the secure agent, and communicates with external environment by some secure strategies, and then senses the external environment by environment interaction module. Task-solution module comprises the operation module including agent as well as reference methods and rules related to agent. Knowledge bank reserves knowledge and task-solution structure acquired by agent. Internal state set is the current situation of used agent and can affect the process of task-solution, so does the task-solution of agent. Binding conditions are to restrict agent's behavior and function. Route strategies determine the moving route of agent.

The server of mobile agent offers rudimental service for mobile agent so that it can transfer on the network and fulfill its functions in the aimed machines. The services are:

Service of lifecycle administration: service for mobile agent's creation, sending, transmission, acceptance and transaction, including the distribution of transaction environment and durable storage, etc.

Directory service: offering unique naming service to help mobile agent find needed service and form route information.

Event service: providing communication mechanism for the interaction between mobile agent and the service facilitator of mobile agent.

Security service: ID identifying and integrity checking of mobile agent and offering secure operational environment.

3. CONSTRUCT B2B ELECTRONIC COMMERCE SYSTEM BASED ON MOBILE AGENT

3.1 Composition of B2B Electronic Commerce System Based On Mobile Agent

B2B electronic commerce system based on mobile agent can accurately search related information of enterprises on the serving chain at any time and in an all-round way, which realizes the sharing of enterprises' information. This paper designs a B2B electronic commerce system according to Fig 2 in the following:

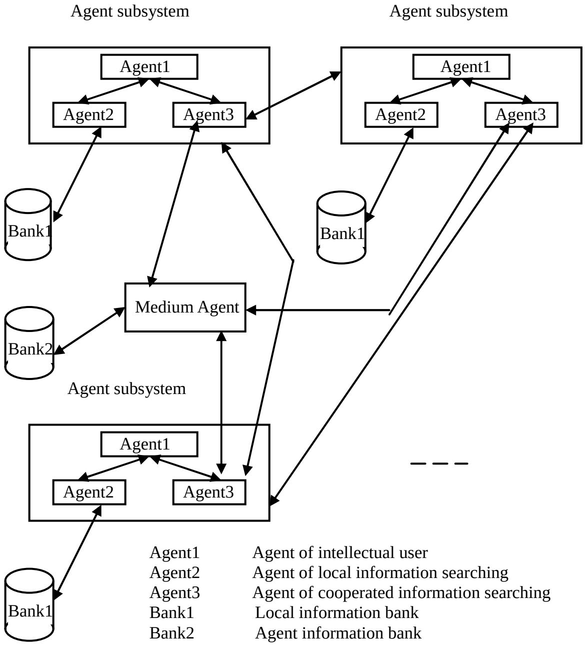

Agent subsystem
Agent subsystem
Agent1
Agent2
Agent3
Agent1
Agent2
Agent3
Bank1
Bank1
Medium Agent
Bank2
Agent subsystem
Agent1
Agent2
Agent3
Bank1
Agent1 Agent of intellectual user
Agent2 Agent of local information searching
Agent3 Agent of cooperated information searching
Bank1 Local information bank
Bank2 Agent information bank

Fig.2. electronic commerce system based on mobile agent

The system is made up of multiple agent subsystems, medium agent system, local information bank and agent information bank. Each agent subsystem is made up three agents which are agent of intellectual user, agent of local information searching and agent of cooperated information searching. The medium agent is to collect, manage, account and search all kinds of agent information. It establishes agent allies according to agents' functions and it serves as credible security identification center to make sure of the secure information mechanism between agent subsystems. Agent of intellectual user provides services for user specially and searchs for related service agents for user automatically; it's like user assistants. Agent of local information searching is to receive the request of agent of intellectual user and then to work for the different requests. Agent of cooperated information searching is mobile; it receives cooperated information searching request from agent of intellectual user, acquires the system address of aimed agent form medium agent system, transfers to the new aimed host machine and then fulfills the searching task.

The process of system is as follows:

User send requests to agent of intellectual user; agent of intellectual user requires the local information searching service; after the local information searching, cooperated information searching can be used according to user request, that is, agent of cooperated information searching requires for property information of other related agent subsystem through medium agent system; after acquiring context information of other agent subsystem, agent of cooperated information searching together with related cooperation can ask for transferring to other information nodes on which agent of local information searching is required to fulfill some operation and return the result after its fulfillment.

3.2 Implementation of B2B Electronic Commerce System Based Mobile Agent

This paper takes clothes enterprises as an instance in realizing the system by adopting JSP, Servlet, JavaBean, JDBC and Aglet [8-10]. JSP view is to display data and interact with user directly; Servlet controller is to generate dynamic pages for user and realize some function of intellectual user; JavaBean and JDBC are to fulfill the different business logics and operate in different database.

(1) Design of database

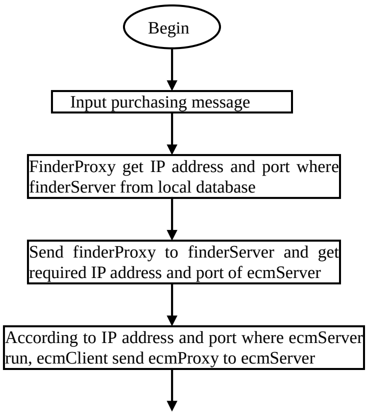
Concentrated database module or distributed database module can be options when database of a system is designed. concentrated database is centered on a database system in a physical location and user can access this database through network. Distributed database is distributed in different physical locations. In this system, the second module is adopted and SQL Server 2000 database is selected.

The system consists of database Ecmdb and Sysdb. They can be installed in several servers. Ecmdb includes user information table, product information table, order information table and logs. Composed of finder and finderconfig, Sysdb is to store the name of servers and websites and so on.

(2) Realization of function module

The system incorporates three modules: producer information, purchasing information and system administration. Producer information and system administration are realized by JSP, Servlet, JavaBean and JDBC. Specifically speaking, JSP realizes display layer; Servlet realizes controlling layer; JavaBean realizes business logic layer; JDBC realizes the connection to database. Purchasing information module adopts JSP, Servlet, avaBean, JDBC and Aglet. In this module, long-distance searching sub-module is constructed, which includes such programs as finderProxy.java, finderServer.java, ecmProxy.java, ecmClient.java and ecmServer.java. FinderProxy.java is to acquire agent information; finderServer.java is to provide service (medium service agent) for finderProxy; ecmProxy.java is to connect ecmClientand ecmServer, in which ecmClient.java works at clients terminals as well as acquires information of products and ecmServer.java works at the server terminals as well as provides ecmClients' needed information.

Mobile agent technology is applied in purchasing information module. Its information flow is illustrated as Fig 3 in the following:


Begin
Input purchasing message
FinderProxy get IP address and port where finderServer from local database
Send finderProxy to finderServer and get required IP address and port of ecmServer
According to IP address and port where ecmServer run, ecmClient send ecmProxy to ecmServer

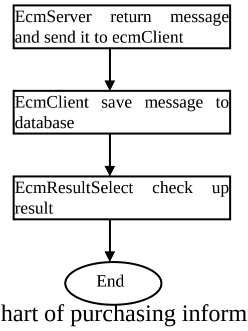

EcmServer return message and send it to ecmClient
EcmClient save message to database
EcmResultSelect check up result
End

Fig.3. flowchart of purchasing information

As show in figure 3, when users input purchasing message, finderProxy gets IP address and port of finderServer from local data bank, and dispatches it to finderServer. The finderProxy interacts with finderServer after it reaches, and gets IP address and port of ecmServer relation to users purchasing message. The ecmClient dispatches ecmProxy to ecmServer according to IP address and port of ecmServer, and interacts between ecmproxy and ecmServer. The ecmServer returns message to ecmClient and saves message to databank, users can check result from ecmResultSelect.

4. CONCLUSIONS

With its intellectual and mobile characteristics, mobile agent provides some effective solutions for electronic commerce system which can not effectively cooperate and share information resource. Mobile agent integrates software, communication and distributed technology to move in the network automatically to fulfill tasks. The article designs a B2B electronic commerce system and realizes information share and interaction between different clothes enterprises in different regions by combininig agent, JSP, Servlet and JavaBean, which makes the system characterized by high expansibility, distribution with high transparency, high modularization and reusability.

REFERENCES

[1] C Guilfoy, J Jeffcoate, H Stark. "Agent on the Web: Catalyst for e-commerce," Ovum Ltd, Tech Rep, 1997

[2] M Ma. "Agents in e-commerce. Communications of the ACM," 1999, 42(3):79-91

[3] Fortino, G., Garro, A., Russo, W., "Modelling and Analysis of Agent – Based Electronic Marketplaces." *IPSI Transactions on Ihternet Research*, 2005, 1(1):24-33

[4] Wang, Y., Tan, KL. Ren J, "A Study of Building Internet Marketplace on the Basis of Mobile Agent for Parallel Processing," World Wide Web: Internet and Web Information System, 2002, 5, 41-66.

[5] Bredin, J., Kotz, D., Rus, D, "Market-based Resource Control for Mobile Agents. Proceedings of ACM Autonomous Agents," 1998, 5

[6] D Martin,A Cheyer, D Moran " The open agent architecture: a framework for building distributed software system. Applied Articial Intelligence," 1991, 13(1):91-128

[7] P Dasgupta, N Narasimhan, L moser, et al. "Magent: Mobile agent for networked electronic trading," *IEEE Trans on Knowledge and Data Engineering*,1999, 24(6):509-523

[8] D B Lange M Oshima. "Programming and Deploying Java Mobile Agents With Aglets," Addison-Wesley, 1998

[9] Zhang Yunyong, Liu Jinde. “Mobile Agent Technology,” Beijing, *Tsinghua University Press*, 2004.

[10] Feng Xiangzhong, WANG ping. “Implementation of the Office Log System of MVC Design Pattern Base On J2EE Platform,” *Journal of Computer Applications,* 2005,25(12):2964-2965

Xiangzhong Feng is a Associate Professor of Department of Computer, Zhejiang Ocean University. He graduated from Dalian University of Technology in 1994. He was a visiting scholar of Department of Computer Science, Concordia University, Montreal, CANADA (2002.12~2003.12).

Architecture & Application of Decision-making Information System for Grouped Enterprises*

Xudong Song[1,2], Xiaobing Liu[1], Kun Zhai[1]
[1]Dalian University of Technology, [2]Dalian Jiaotong University
Dalian, Liaoning Province, China
Email: xudongsong@126.com

ABSTRACT

Basing on CIMS project of an enterprise, construction methods of decision-making information system of grouped enterprises were studied and relative structure basing on distributed data warehouse was presented in this paper. The architecture framework of multilayer distributed decision-making information system and its function were put forward by adopting distributed data warehouse, on-line analytical processing, data mining, etc. It satisfies the leaders of different levels with their decision-making requirements. Finally, the key techniques of the implementation of this system were discussed.

Keywords: Grouped Enterprise, Data Warehouse, Decision-Making Information, On-line Analytical Processing

1. INTRODUCTION

At present, with increasing severity of business competition, grouped enterprises have accumulated mass transaction data which are distributed to different departments and operation platforms. How to find valuable information for enterprises decision-making from mass transaction data has become a very important problem. In view of above-mentioned problem, this paper combines some key techniques, such as distributed data warehouse, on-line analytical processing, data mining, decision support system[1,2], and puts forward a whole structure framework of decision information system for grouped enterprises.

2. WHOLE STRUCTURE FRAMEWORK OF DECISION-MAKING INFORMATION SYSTEM FOR GROUPED ENTERPRISES

In order to improve headquarters' scrutiny ability and decision-making level of different departments, more and more grouped enterprises badly need distributed data administration and decision-making support of various levels [3]. So grouped enterprises begin to create data warehouse and build decision-making information system. Data warehouse is a subject oriented, integrate, non-volatile and time variant data set, which supports decision-making establishment process of enterprises management [4]. This paper puts forward a whole structure framework of decision-making information system for grouped enterprises, which is based on the basis of distributed data warehouse. The whole structure framework is shown in the Fig.1. This framework adopts distributed technique, builds multilayer data environment and provides multilayer decision analysis. Namely, it supports not only departments' level decision-making but also grouped headquarters' level decision-making, which can provide more useful and valuable decision information for enterprises.

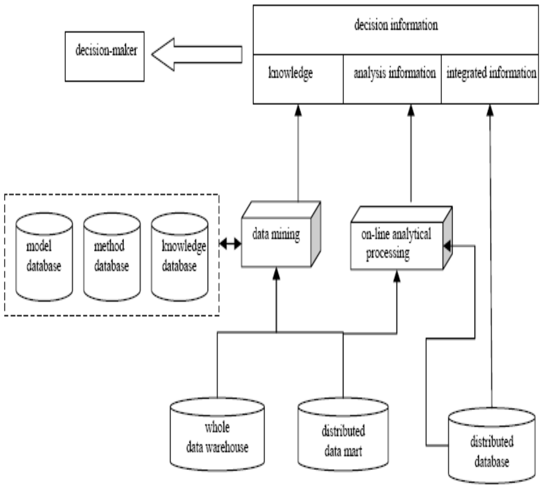


Fig.1. whole structure framework of decision information system grouped enterprises based on Data distributed data warehouse

In the structure framework, the whole grouped enterprises-oriented view consists of distributed database, which provides a uniform data platform of data store and data organization for decision-making information system. Each department has its different data and demand in the enterprises. Building data mart for each department can improve greatly departments' decision-making ability and analysis level. Distributed data mart provides directly data capacity for on-line analytical processing and data mining, which can improve greatly application capability of multilayer distribution for decision-making information system.

Model database, method database and knowledge database of decision information system are managed uniformly by database management system [5]. Model database and method database can provide instruction for diverse analysis tools. The knowledge of knowledge database can not only instruct new knowledge discovery but also obtain continually new knowledge.

On-line analytical processing and data mining based on distributed data warehouse can find automatically potential pattern from vast data and make automatically prediction analysis.

The model of model database includes not only math model but also data processing model, graph and image model, report forms model, intelligent model and so on. Diverse models can extend greatly system decision ability. Methods used in solving problems comprise the method database, which is built on the basis of models and calculates in different ways according to definite models. Its expansion capability can extend

* This work is supported by National Natural Science Foundation of China, Grant Number: 70572098

momentarily a new component in the groupware. Therefore, Method database also includes new methods and combined methods. Knowledge database has the knowledge used in solving problems, which is mainly referred to the one adopted in inferring calculation.

3. ARCHITECTURE STRUCTURE OF DECISION-MAKING INFORMATION SYSTEM FOR GROUPED ENTERPRISES

J2EE is defined as standard for developing enterprise distributed application by SUN Corporation, which provides a multilayer distributed application model and a series of developing technique standards. Its groupware technique includes EJB, JSP, SERVLET, JMS, JDBC and so on. Combining J2EE technique and distributed data warehouse theory, the architecture structure of decision information system for grouped enterprises based on Web environment is designed by this paper. It can satisfy decision-making demand for different department's decision-maker. The architecture structure of decision-making information system for grouped enterprises based on Web environment is shown in Fig.2.

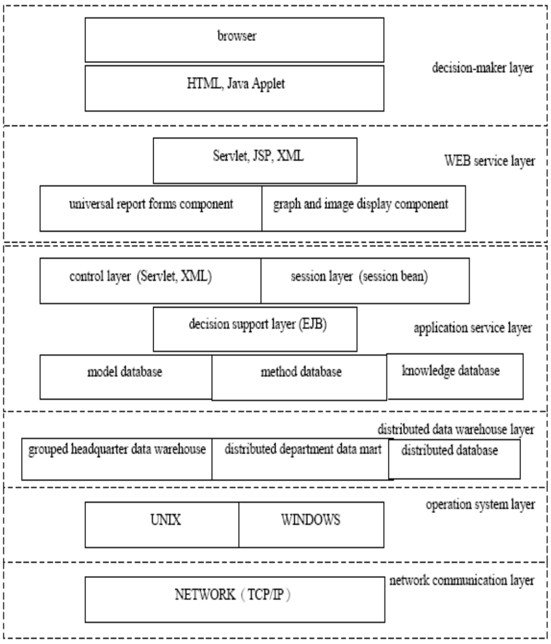

browser
HTML, Java Applet
decision-maker layer
Servlet, JSP, XML
WEB service layer
universal report forms component
graph and image display component
control layer (Servlet, XML)
session layer (session bean)
decision support layer (EJB)
application service layer
model database
method database
knowledge database
distributed data warehouse layer
grouped headquarter data warehouse
distributed department data mart
distributed database
operation system layer
UNIX
WINDOWS
network communication layer
NETWORK（TCP/IP）

Fig.2. The architecture structure of decision-making information system for grouped enterprises

The merits of system architecture structure are as follows:

1) The system has powerful cross-platform characteristic and can be maintained easily. The whole system is of powerful cross-platform due to J2EE's relative characteristic. The middle layer's application makes the system easy-maintain and extend powerfully, it is also useful for connecting with the existing system precisely.
2) The system has high efficiency and strong security. The processing logic of distributed database query and kernel business is encased to EJB in the decision-making information system. The connection pool technique is applied to the system, which also improves system's efficiency and security highly.
3) The system has transparent performance in accessing distributed data warehouse. Distributed data environment supports automatic navigation and transparent accessing function. Client application can realize dynamic transmission of client connectivity according to client ascription and operation rights, which also carries out remote operation through connectivity. The system doesn't need care position of remote data and modify any codes, which means the system can access different databases.

4. APPLICATION OF DECISION-MAKING INFORMATION SYSTEM

4.1 Analysis of decision-making information system

North-eastern Special Steel Grouped Corporation mainly produces special steel. The productions are diverse. The indent of each production is small. Therefore, the corporation is of small-scale and diverse production. Its CIMS project includes three distributed productive sites which are located in Dalian, Fushun and Qiqihaer of China and hundreds of sales sites which are located in all over China. So the project is typically regional distributed system. Through analysis of decision-making information system of corporation, the subjects are presented as follows: customer, production, supplier, equipment, company, indent, consignment, materiel, cost, stock, and sales trend.

4.2 Design of decision-making information system

Decision-making information system is partitioned three layers: general decision-making information, department decision-making analysis and grouped headquarter decision-making analysis.

General decision-making information establishes corporation report forms system. It utilizes all-around data to describe corporation business, market status and customer situation through report forms and fuzzy query. Its detailed function is shown in Fig.3.

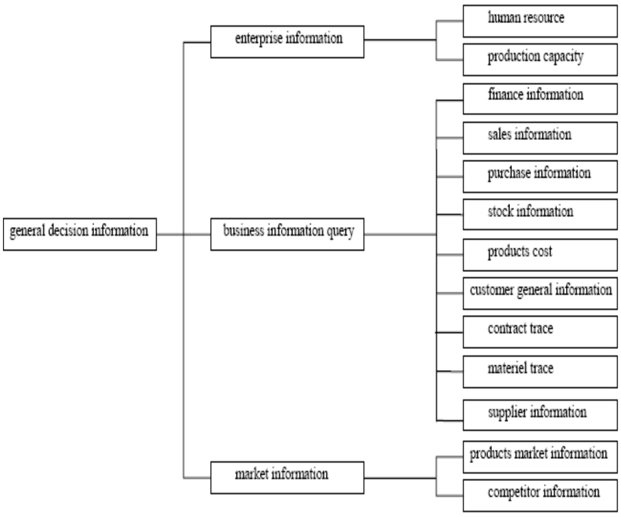


Fig.3. Function modules of general decision information

Department decision-making analysis adopts general report forms system, inference engine and graph and image display components and utilizes data mart information to analyze and subdivide business process. Through data mining based on various models and methods, and data warehouse mined by data mining, it can obtain refined data and deduce business pattern. These patterns are not apt to find market rules, customer trend and commercial patterns in report forms query. Its detailed function is shown in Fig.4.

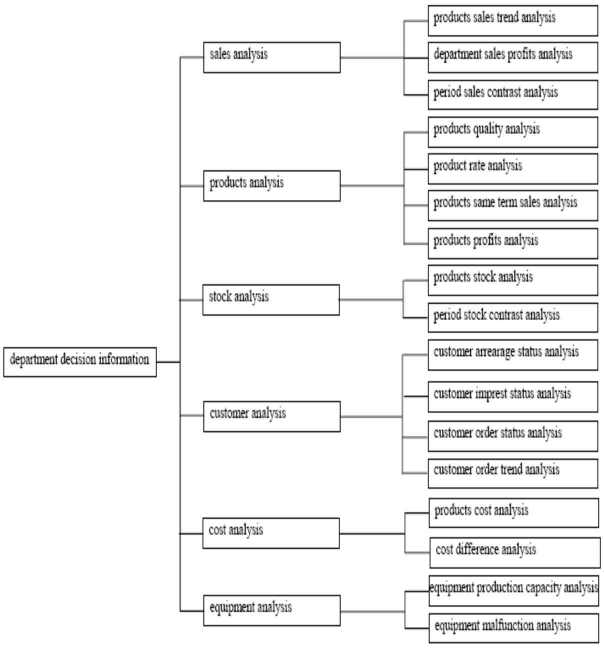

department decision information
sales analysis
products sales trend analysis
department sales profits analysis
period sales contrast analysis
products analysis
products quality analysis
product rate analysis
products same term sales analysis
products profits analysis
stock analysis
products stock analysis
period stock contrast analysis
customer analysis
customer arrearage status analysis
customer imprest status analysis
customer order status analysis
customer order trend analysis
cost analysis
products cost analysis
cost difference analysis
equipment analysis
equipment production capacity analysis
equipment malfunction analysis

Fig.4. Function modules of department decision analysis

Grouped headquarter decision-making analysis adopts general report forms system, inference engine and Graph and image display components and utilizes data warehouse information to realize parallel query combining advanced decision support technique. It can make prediction intellectually and reliably. Its detailed function is shown in Fig.5.

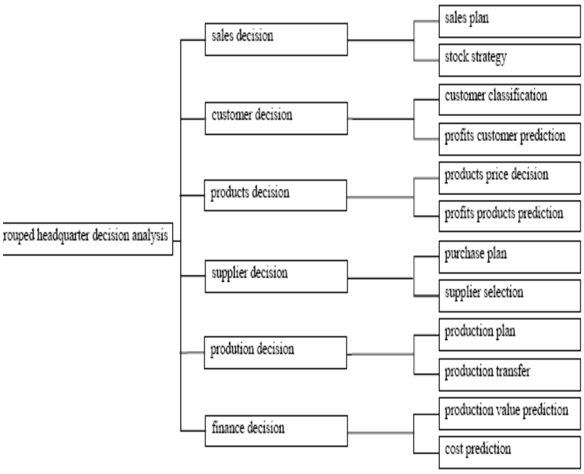


Fig.5. Function modules of grouped haadquarter decision analysis

4.3 Fulfillment of decision-making information system

Decision-making information system is partitioned three layers: general decision-making information, department decision-making analysis and grouped headquarter decision analysis.

The decision information system of corporation applies JBuilder9 as developing tools and selects Struts1.1 series component of supporting MVC (Model-View-Controller) structure framework. It also adopts general report forms system and graph and image display components to improve fulfillment. In general report forms system, all data and information of report forms query, query condition, report forms format and so on is organized and stored to database. And then through page produced dynamically by procedure, the information is displayed on client browser. The system supports stored procedure and model transfer. So it can realize complicated query effectively.

The system application server applies Weblogic 8.1 and adopts Oracle8i and Oracle Express as platform to build relational data warehouse and multidimensional data warehouse. Through Relational Access Manager provided by Oracle, the metadata is managed. And by virtue of analytical function provided by Oracle 8i, the data are analyzed and disposed. The system supports functions of roll-up, drill-down, slice, dice, pivot, graph and image display and so on. In view of data in distributed data warehouse, the system adopts various methods and predictive models to make data analysis and data mining. For example, in view of sales trend subject, the decision information system of corporation provides sales channels' trend analysis of diversified steel species, sales profits' trend analysis of diversified steel species in different time and different region, customer trade distributions' trend analysis of diversified steel species and so on through graph and report forms.

5. CONCLUSIONS

The paper applies techniques of distributed data warehouse, on-line analytical processing, and data mining to construction of decision-making information system for grouped enterprises and puts forward an architecture structure of decision information system for grouped enterprises based on distributed data warehouse. Through being practiced successfully in certain corporation, it is indicated that the decision-making information system based on this architecture structure can satisfy decision-makers' requirement of acquiring enterprises' decision information quickly and effectively. So it can bring economic and social benefits highly.

REFERENCES

[1] AKINDE M O, BOHLEN M H, JOHNSON T, *et al*. *Efficient OLAP query processing in distributed data warehouses. Information Systems*, 2003(28):111-135.

[2] GAO Hong-shen. *Decision support system: theory, method and cases*. Beijing: Tsinghua University Press, Nanning: Guangxi Science and Technology Press, 2000(in Chinese).

[3] HAN Lan-shen, SHAO Bei-en. *Building hybrid distributed data warehouse for grouped enterprises. Computer Integrated Manufacturing System-CIMS*, 2003, 9(1):80-84 (in Chinese).

[4] INMON W H. *Building the data warehouse*. John Wiley & Son, Inc.2002.

[5] LUO Shu-qiang, HE Yu-lin. "Quality management decision support system based on networked manufacturing," *Computer Integrated Manufacturing System-CIMS*, 2004, 10(2):171-175 (in Chinese).

Research on Intelligent Customer Relationship Management for Grouped Enterprises*

Xudong Song [1,2] , Xiaobing Liu [2]
[1.] Dalian Jiaotong University , [2.] Dalian University of Technology
Dalian, Liaoning Province, China
Email: xudongsong@126.com

ABSTRACT

In view of characteristics of grouped enterprises, intelligent customer relationship management for grouped enterprises was researched. The new data mart concept of intelligent customer relationship management for grouped enterprises called as dynamic data marts was proposed, and a new method of data modeling was provided. An architecture structure of intelligent customer relationship management for grouped enterprises was presented, through combining many techniques involving data warehouse, on-line analytical processing, data mining, etc. Finally, the structure of data environments and functional integrated framework of intelligent customer relationship management for grouped enterprises were expressed.

Keywords: Grouped Enterprise, Intelligent Customer Relationship Management, Data Environment, Data Model

1. INTRODUCTION

With the fast development of Ecommerce, Internet and Communication technology, the competitions for global markets are increasingly drastic. The differences of products between craft brothers are becoming smaller when some technologies mature, and the focus of competition has changed from products-centered to customers-centered namely the emphases service has been focus on some valuable customers who would be provided individuation services in order to decrease the cost of services. In order to enhance the profits and gain the most customers, customer relationship management [1] (CRM) is attaching more and more attentions by all the enterprises.

CRM origins from customers-centered business mode as a bran-new business strategetic thought, aiming to improving relationship between enterprises and customers and maximise customer values and enterprises’ profit. From the point of view of management science, CRM is a set of management ideas and business mode which is customers-centered and information enabled. From the point of view of software application, CRM is a software package realizing such management ideas.

At present, CRM has greatly progressed in both theories and applications, but the current CRM system emphasizes the customers relation management based on technology, namely it focus on Call center, auto marketing and auto sales etc, and these functions are far from satisfying those grouped enterprises. Grouped enterprises need to completely analyze and mine the information of customers through using the foundation of integrated data environment and the technology of data mining combining with the idea of knowledge management in order to provide the better decision support for the leader of grouped enterprises. In this paper, we studied grouped enterprises intelligent customer relationship management (ICRM).

2. GROUPED ENTERPRISES ICRM

In order to better share the resources and enhance the market competition capability, many industries have recombined their assets and built grouped enterprises. The grouped enterprises have the following peculiarity: the headquarters of grouped enterprises mainly is responsible for the whole sale and purchase, namely carry the unity to the sales order, the unity to the materials purchase, and it is the grouped enterprises’ decision-making center and information center; each stock company has its relative self-determination and self-governed account; stock companies distribute far away in different cities or countries.

Grouped Enterprises ICRM should have the capability of adjusting both the inner and outer resource of grouped enterprises, optimizing the increment chain of market and providing the decision-making analysis for enterprises to manage customer relationship etc through setting up the mutual benefit between enterprises and customers. Using ICRM system, grouped enterprises may find new market and new channel, enhance values and loyalty degree of customers, and set up the organize that can adapt to change in order to realize the maximal benefit. The above-cited is concretely expressed with three points:

1) Attract and keep more customers. Grouped enterprise ICRM should satisfy the needs of customers by quicker and more circumspect high quality services in order to assure to realize customers’ lifelong values.

2) Increase the benefit of enterprises, decrease the cost of enterprises, and realize the maximal value of customers. With the customers as center, ICRM should adjust and manage the resources of personals, assets and goods etc as well as corresponding sales and purchases etc, which are correlative with customers.

3) Provide the decision support service. Grouped enterprises ICRM can provide a powerful capability of data analysis and mining, which will effectively sustain the enterprises’ decision-making and production direction.

3. ARCHITECTURE STRUCTURE OF GROUPED ENTERPRISES ICRM

Grouped Enterprises ICRM should be built on the base of the integrated data environment. We put forward the ICRM architecture structure for grouped enterprises, showed in Fig 1, which includes two parts: layered data environment frame and ICRM functional frame.

* This work is supported by National Natural Science Foundation of China, Grant Number: 70471056

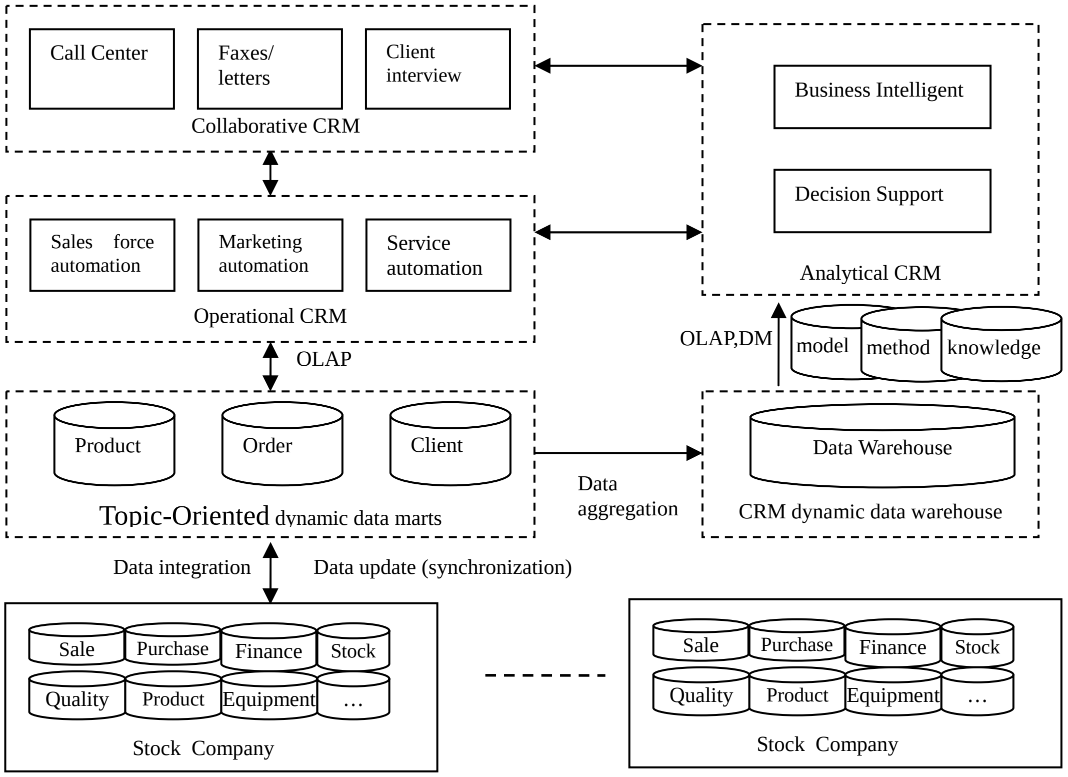


Fig.1. Architecture Structure of Grouped Enterprises ICRM

3.1 Layered Data Environment Frame

The grouped enterprises often include many stock companies, and each stock company has its own self-governed business management system. Generally, each stock company need to set up its own many local databases, which together construct the group, namely ERP/MES operational distributed databases.

In order to support the uniform purchase and uniform sale, all ERP/MES operational distributed databases need to be combined, and the products information, finance information and provider information etc need to be integrated, even incompatible information origin in enterprises, need to be organized through data extraction and transformation. Finally, it is also necessary to set up Topic-Oriented dynamic data integration environment in order to support some departmental grade applications, and set up the Topic-Oriented whole data warehouse in order to support the management and decision-making activity of group.

Here, we provide a concept of Topic-Oriented dynamic data marts, which is different from the static data of data warehouse[2]. The dynamic data marts, which combined the short-time update data and long-time history data, may be better to adapt to the quick change of market and requirement in ICRM system and satisfy quick department businesses management. Furthermore, we showed three topics: product, order and customer. Topic-Oriented dynamic data marts can provide the data environment for operational CRM. According to the individuation information got from all kinds of customers, the operational CRM may establish corresponding marketing flow, sale flow and service flow, and the information produced from these activities can be record into dynamic data marts. At the same time, the pertinent data can be copied or updated according to the mapping relation between Topic-Oriented dynamic data marts and business databases.

The core of building integrated data environment is the built of data model of CRM dynamic data marts and data warehouse. Through studying the model building of the data marts and data warehouse, we put forward a new model of Topic-Oriented data marts and data warehouse base on coupling dimensional degree. Referring original star schema or snowflake schema, this kind of extended model of data marts and data warehouse synthesize original dimensional tables of every topic into new dimensional tables called topic-coupling -dimension according to the coupling degree relative to topic, at the same time the fact table change nothing and remove the dimensional tables that haven’t straight relationship to topic. In the topic-coupling-dimension, the properties of each dimensional table have been combined and the combination of each dimensional table primary keys just constitute the primary key of topic-coupling-dimension. So, for some special query requirements, we don’t need to do more conjunction operations. In order to enhance the speed of on-line query, we only need to correlate the topic-coupling-dimension with fact table.

3.2 ICRM Functional Frame

Grouped Enterprises ICRM function frame can be divided into three parts: operational CRM, collaborative CRM and analytical CRM.

(1) Operational CRM

Operational CRM, mainly includes Sales Force Automation, Marketing Automation and Service Automation etc. It has realized smooth link and conformity between the front management and the background management. The main aim of operational CRM is that the business personnel of each department can share the customers resource in the daily work and decrease the stagnation point of information flow. The customers needn’t be bothering to single deal with the relations among all departments of enterprises, but need to look the enterprises as a whole body. Operational CRM is the most basic applied module of ICRM.

Sale automation module can exchange information with the sale management system of ERP through the dynamic data marts and forecast orders after analysis. Service automation module can exchange information with the sale and customers management system of ERP through dynamic data marts and deal with order management after analysis. The customers can obtain all kinds of services of ICRM in the front and put forward all kinds of consultation and other service requirement .

(2) Collaborative CRM

Collaborative CRM provide the function of call center etc, which provide an entity of receiving and giving off calls for customers services, market management, technology support and other special business activity and provide many kinds of channels for collecting customers information and doing a service each other with customers, which boosts up the communication capability of enterprises with customer. Through the modern communication technologies of telephone, fax and internet etc, the call center face customers as exterior and link with the whole enterprise as interior, namely link with the management, service, scheduling, product and maintenance of enterprises. Collaborative CRM may also store all kinds of information acquired from customers into the dynamic data marts, and synchronize operational distribution database through the mapping relation between dynamic data marts and operation distribution database.

(3) Analytical CRM

Deriving knowledge about customers is one of the main aims for analytical CRM. Usually, customer buying behavior is analyzed to forecast potential products, points of time, or quantities of future orders. This knowledge is used in order to provide customers with what they require at a given time and place. Furthermore, this knowledge is useful for grouped enterprises, because they can adjust their product development to market requirements. In turn, this may lead to decreased delivery times, which as pointed out above, contributes potentially to customer satisfaction[3].

As the core of ICRM, Analytical CRM combined many techniques involving data warehouse, on-line analytical processing, data mining, model base, knowledge base and method base etc. Data warehouse realizes the store and synthesis of customers information data. The primary focus of data mining is to discover knowledge, previously unknown, predict future events and automate the analysis of very large data sets[4]. Through data mining, the complex relation among customers datum and the effect on decision-making by this kind of relation can be found. Data mining is applied to find the valuable knowledge of data warehouse and forecasts the behavior of customers, credit venture and the sale trend etc. OLAP realizes the multi-dimensional data analysis. Model base realize the assemble assistant decision-making of many generalized models. Knowledge base is applied to imitate some intelligence behavior regulations, patterns and rules in the course of mankind making decision-making; method base provide system the universal methods of decision-making, the methods of optimization and software tools etc.

4. SOFTWARE ARCHITECTURE OF GROUPED ENTERPRISES ICRM

Grouped Enterprises ICRM is built on utterly heterogeneous dynamic and distributed network computing environment. To implement such a environment we need such new computing platform as J2EE which provide a highly portable and compatible platform for grouped enterprises. A multi-tier distributed enterprise application combining Internet can be established on J2EE. It is divided into three tiers: client tier, middle-tier, database tier. The middle tier includes web presentation and business logic. Users input request information with the browser installed in client computer and look over results returned by server. Web presentation accept data requested by client and invoke EJB components to implement business logic after simple disposal, then returned the results to client in HTML format. EJB components accept data from web presentation or client tier and save the disposed result in back end databases. EJB container provides rock-bottom service including session bean and entity bean. The software architecture of grouped enterprises ICRM is showed in Fig 2.

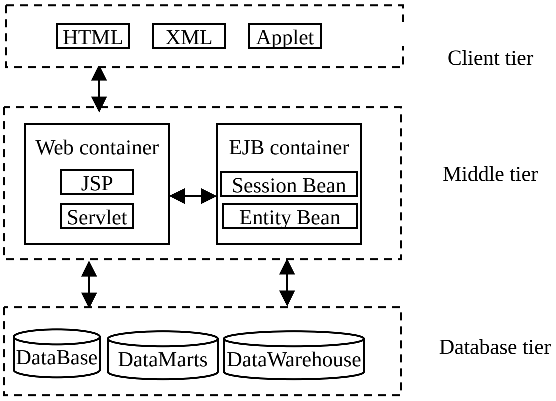


Fig. 2 . Software Architecture of Grouped Enterprises ICRM

5. CONCLUSIONS

Grouped Enterprises ICRM is the only channel enterprise and customer interact, containing important information on enterprise decision support and other application system need. This paper analyzed the object and functions of grouped enterprises intelligent customers relationship management, and presents the architecture structure of grouped enterprises ICRM which can guide and help grouped enterprises to develop and implement actual customer relationship management system in their management practice.

REFERENCES

[1] Thearling, K. Data mining and CRM: zeroing in on your best customers. DM Direct, 12(1999).

[2] INMON W H. Building the data warehouse. John Wiley & Son: Wiley Computer Publishing Inc.1998.

[3] Joerg Becker, Alexander Dreiling, Roland Holten, Michael Ribbert. Specifying information systems for business process integration – A management perspective. Information Systems and E-Business Management. Volume 1(3): 231-263(2003).

[4] Jaideep Srivastava, Jau-Hwang Wang, Ee-Peng Lim, and San-Yih Hwang. A Case for Analytical Customer Relationship Management. Lecture Notes in Computer Science. Volume 2336:15-27 (2002)

E- Education

From e-Learning to e-Research: Building Collaborative Virtual Research Environments Using Sakai*

Xiaobo Yang, Rob Allan
STFC e-Science Centre, Daresbury Laboratory,
Warrington WA4 4AD, UK
Email: {x.yang, r.j.allan}@dl.ac.uk

ABSTRACT

Aiming at providing teachers and students virtual learning environments (VLE), e-Learning systems are today widely adopted in educational systems. These e-Learning systems provide a set of services so that teachers and students can communicate through the Web easily. In this paper, we are going to describe how such an e-Learning system can be adopted to support research activities. Scientists are now facing increasingly complex challenges. To meet these challenges, there is need to support research activities using the latest information technology. Since VLE systems are becoming more and more mature, they are natural candidates for building virtual research environments (VRE) systems to support both researchers and administrators involved in research. In this paper, we will discuss how existing e-Learning systems can be converted to VRE systems to support research. In particular, a Sakai based VRE system will be discussed to present what this e-Learning based VRE system can bring to researchers and their supporting staff and to increase their productivity.

Keywords: E-Research, Virtual research environment, Portlet, JSR 168, Web Services for Remote Portlets

1. INTRODUCTION

Today, e-Learning systems are widely accepted and deployed. With the development of distributed information technology, these systems are now mainly Web-based. Besides providing teaching and learning materials, e-Learning systems provide a platform for their users, i.e., teachers and students, to communicate easily. Tools like instant messaging and online discussions are available. Web 2.0 additionally brings the possibility to provide more interactive user interfaces (UI) to web applications including e-Learning systems, so that they can provide a similar experience to desktop applications. In general, e-Learning systems are becoming more and more mature and powerful to help teaching and learning activities. In this paper, we are not going to discuss e-Learning systems themselves; but will focus on making use of such a system to help research, i.e., to realise e-Research.

Research, another key role of many universities besides teaching and learning, is today becoming increasingly complex. On one hand, researchers are facing increasingly complex scientific challenges, which normally require more people and resources for cooperation. In fact, cooperation among research groups at university and even country level is now quite common. For example, the European EGEE [1] project "brings together scientists and engineers from more than 90 institutions in 32 countries world-wide to provide a seamless Grid infrastructure for e-Science..."

On the other hand, research is not limited to scientific activities, but also involves administrative activities and cooperation with supporting staff for proposals, recruitment, project and financial management. Hence, a platform is required to provide support for both research itself and its surrounding activities. As e-Learning systems are becoming mature, they are the current focus for supporting research. In this paper, we are going to present how an e-Research system can be built up using an existing e-Learning system, Sakai.

E-Research, originating from the term e-Science, was adopted by UK JISC (the Joint Information Systems Committee) to cover all research domains not just the natural sciences. It is "concerned with technologies that support all the processes involved in research including (but not limited to) creating and sustaining research collaborations and discovering, analysing, processing, publishing, storing and sharing research data and information". JISC includes virtual research environments (VRE), grid computing, text and data mining, etc. as typical technologies for realising e-Research. JISC has funded a set of projects for investigation of various technologies to realise e-Research. For example, in VRE-1 programme, JISC funded projects ranging from lightweight grid middleware (GROWL) to VRE research in biology (IBVRE) and humanities (BVREH). For more information, see JISC VRE web site [3].

E-Research has also been described by the Department of Education, Science and Training in Australia. "The term 'e-Research' encapsulates research activities that use a spectrum of advanced ICT capabilities and embraces new research methodologies..." Again, by providing improved access to knowledge and information, e-Research will "enable researchers to perform their research more creatively, efficiently and collaboratively across long distances and disseminate their research outcomes".

Overall, e-Research as a concept aims at improving research efficiency and productivity by making use of existing and emerging technologies to cover all phases of research processes. In [9], Lawson and Butson gave a detailed review of e-Research studies in USA, UK, Australia and New Zealand. Although e-Research was noted as a vague concept, it was reported that today’s research activities can indeed benefit from e-Research with increased quality of research, savings in cost and time etc.

The rest of the paper is organised as the follows. As an implementation of e-Research, VRE systems will be discussed first followed by conversion of an existing e-Learning system to an e-Research system. This will be based on our work of building a VRE system using Sakai [5], an open-source e-Learning framework. VRE systems will be further discussed before concluding remarks are drawn.

* The Sakai VRE Demonstrator Project is funded by the UK Joint Information System Committee under the Virtual Research Environments Programme.

2. FROM E-LEARNING TO E-RESEARCH

2.1 Virtual Research Environments an E-Research Implementation

The architecture of VRE systems has been described in [12] as service based. In brief, VRE systems are constructed on top of a bundle of services either local or remote. Such a VRE system should also be able to be extended easily on demand by plugging in new services. Adoption of the service-oriented architecture (SOA) brings VRE systems interoperability plus the maximum flexibility. Various services provided by different providers can be integrated within one VRE system. This naturally requires information system integration, i.e., a new VRE system could be built up using services from other (VRE) systems. The integration here can be realised in two ways: 1) integration of traditional data-centric services, and 2) integration of presentation-based services.

Option 1) is easy to understand because web services are commonly adopted today. A VRE system can obviously act as a client of various web services, for example to provide live information of weather forecast or status of an experimental facility. Option 2) goes a further step by integrating markup fragments rather than raw data. This comes from the idea of web components known as portlets. E-Learning systems normally provide web portals as their gateways, for instance, uPortal [6], so that end-users are able to access information through these portals. Facing end-users, portals typically provide functionalities like authentication/ authorisation and customisation. Most importantly, a portal must provide its visitors with a pathway for accessing internal and external content. Portlets acting as web components can be published using web service technology as remote portlets. This can be done through the Web Services for Remote Portlets (WSRP) standard [8]. These two approaches have been discussed in detail to show how existing grid tools can be integrated in Sakai VRE [14].

While VRE systems are aiming at improving research productivity, the key is collaboration. Today, research activities require collaboration between distributed resources including researchers. Hence, universal communication is essential in a VRE system. This may be done through instant messaging, online discussion, and communication with mobile devices, etc. Moreover, VRE systems are required to manage digital contents. Data (including meta-data) is the core of today's research and is exchanged among researchers all the time. Data may be collected from experimental instruments or numerical simulation programs. Researchers will analyse these data and probably generate new data from them. Also documents are inevitably created for project management, publication and so on. All these data are required to be managed in an efficient way, which may be realised by a digital repository either outside or inside a VRE system.

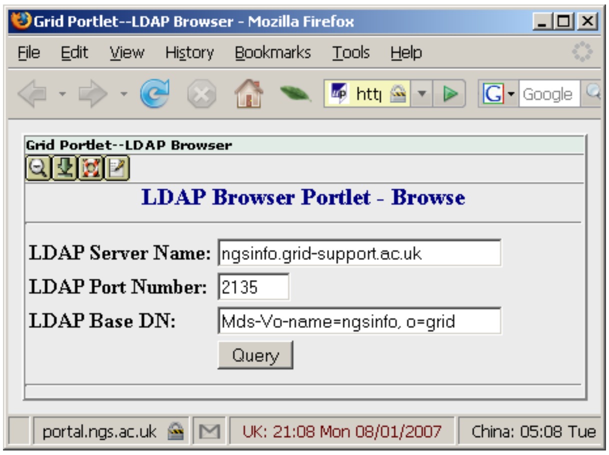
2.2 Support E-Research Using Sakai

As a collaboration and learning environment for education, Sakai [5] provides a set of tools and services for accomplishing its aim. Built on top of the Spring framework, Sakai reaps the key benefit of this lightweight framework – the ease of extension. Sakai services and tools can be added in or removed from the system on demand which makes Sakai a highly customisable platform to meet various requirements.

Sakai provides communication tools like chat room, discussion and presentation. Recently a new audio/ video conference tool has being developed at the Lancaster University. These tools aim at linking teachers and students together and can facilitate collaboration for tackling large research challenges. Sakai also provides repository tools, for example the resources tool sharing resources such as image files and project documents. At Daresbury Laboratory, we have developed a document management tool to provide support for organising conferences/ workshops [12], in particular for reviewing submitted papers.

A fundamental aspect of an SOA is the ability to loosely-couple services as re-usable components. As mentioned above, *portlets*, the basis of today's portal world, are designed as web components which generate markup fragments. Example output from a portlet called LDAP Browser Portlet is listed below. These fragments are taken from uPortal [6] within which this portlet is deployed. They are normal HTML fragments except that there are no tags like *html*, *head* or *body*. These tags are added by the portal framework to generate a full HTML page from fragments to be rendered in browsers. Fig 1 illustrates the output of the markups below inside uPortal in *detach* mode which displays only one portlet.

```
<center><b><font size="+1" color="navy">
LDAP Browser Portlet - Browse
</font></b></center>
<hr/>
<table border="0">
  <form action="/tag.2355daaece79b3c7.render.
userLayoutRootNode.target.46.uP
?uP_portlet_action=true#46" method="POST"/>
    <tr>
      <td>
        <b>LDAP Server Name:</b>
      </td>
      <td>
        <input type="text"
          name="ldap_hostname" size="32"
          value="ngsinfo.grid-support.ac.uk"/>
      </td>
    </tr>
    <tr>
      <td>
        <b>LDAP Port Number:</b>
      </td>
      <td>
        <input type="text"
          name="ldap_hostport" size="5"
          value="2135"/>
      </td>
    </tr>
    <tr>
      <td>
        <b>LDAP Base DN:</b>
      </td>
      <td>
        <input type="text"
          name="ldap_basedn" size="32"
          value="Mds-Vo-name=ngsinfo, o=grid"/>
      </td>
    </tr>
    <tr>
      <td/>
      <td>
        <input type="submit" value="Query"/>
      </td>
    </tr>
    </form>
  </table>
  <hr/>
```


Grid Portlet--LDAP Browser - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http
Google
Grid Portlet--LDAP Browser
portal.ngs.ac.uk
UK: 21:08 Mon 08/01/2007
China: 05:08 Tue

LDAP Browser Portlet - Browse

LDAP Server Name: ngsinfo.grid-support.ac.uk
LDAP Port Number: 2135
LDAP Base DN: Mds-Vo-name=ngsinfo, o=grid
Query

Fig.1. LDAP Browser Portlet displayed in a web browser (uPortal detach mode)

As you can see from Fig.1, uPortal makes use of the above markup generated by the LDAP Browser Portlet and adds additional information to it so that a HTML portlet page is created. Here, uPortal adds a title named "Grid Portlet -- LDAP Browser" for this portlet and four icons for all view modes supported by the portlet.

Although portlet portability is guaranteed by the JSR 168 Java specification [4], a better approach to portlet re-use is to deploy them once but run them anywhere. This requires another portlet specification, the Web Services for Remote Portlets (WSRP) standard [8]. Suppose a portlet container exposes its portlets through some web service interfaces, the portlet fragments can then be consumed on client side. Since we have developed a set of grid portlets to perform tasks like proxy management, job submission and file transferring for the UK National Grid Service (NGS) [13], these can be included in our VRE system using WSRP so that researchers can seamlessly access remote computing and data grid resources.

At the time this paper is being written, the latest release of Sakai, version 2.4, has Apache Pluto (a reference implementation of JSR 168 from Apache) integrated to provide native JSR 168 support. While this would be helpful for deploying loosely-coupled portlets inside Sakai, there is still a need of accessing remote portlets outside of Sakai. For example, portlets may be managed by a provider who does not allow their portlets to be distributed or deployed anywhere else but exposed through its WSRP producer. In this scenario, the aforementioned WSRP specification helps Sakai to consume those remote portlets.

We have written a WSRP4J (a WSRP 1.0 implementation from Apache) based WSRP consumer designed for Sakai and described it in detail previously [14]. For completeness of this paper, WSRP is briefly described here. In Fig.2, when a request arrives at the portal server (in our case, Sakai), a WSRP consumer will be responsible for redirecting this request to a remote WSRP producer, within which remote portlets are published. The producer passes the request from the consumer to its portlet container, from which markup fragments will be generated then sent back to the consumer. Now the consumer will collect these markups and ask the portal (Sakai here) to render it for the end-user.

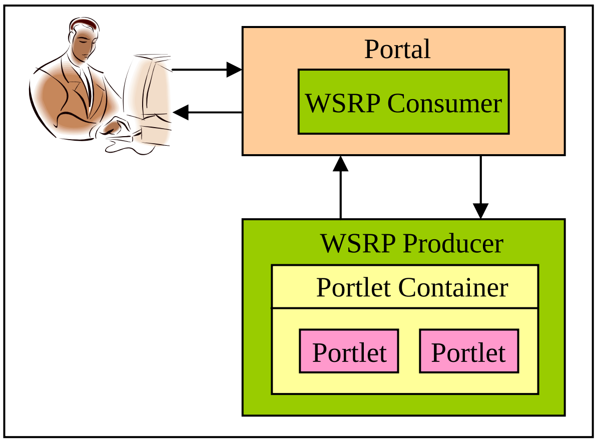
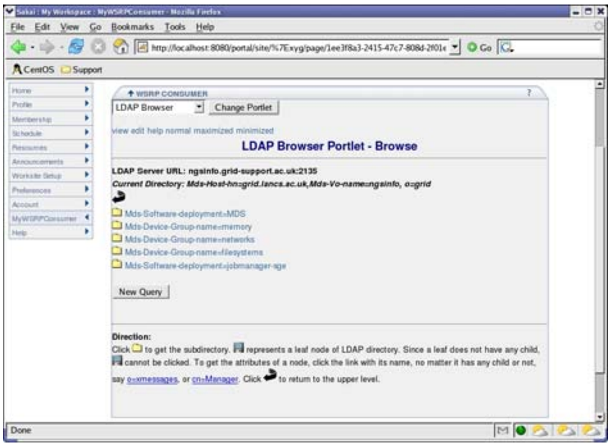


Fig.2. Communication in WSRP scenario

The same portlet as shown in Fig.1, the LDAP Browser Portlet, has been published as a remote portlet using WSRP4J. Fig.3 is a screenshot taken of this portlet running remotely inside Sakai through a WSRP consumer. Through WSRP, grid portlets developed for the UK NGS Portal [13] provide Sakai users the ability to access computational and data grid resources seamlessly within our VRE system.


Sakai : My Workspace : MyWSRPConsumer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:8080/portal/site/%7Exyg/page/1ee3f8a3-2415-47c7-808d-2f01e
Go
CentOS
Support
Home
Profile
Membership
Schedule
Resources
Announcements
Worksite Setup
Preferences
Account
MyWSRPConsumer
Help
WSRP CONSUMER
?
LDAP Browser
Change Portlet
view edit help normal maximized minimized
Done

LDAP Browser Portlet - Browse

LDAP Server URL: ngsinfo.grid-support.ac.uk:2135

Current Directory: Mds-Host-hn=grid.lancs.ac.uk,Mds-Vo-name=ngsinfo, o=grid

- Mds-Software-deployment=MDS
- Mds-Device-Group-name=memory
- Mds-Device-Group-name=networks
- Mds-Device-Group-name=filesystems
- Mds-Software-deployment=jobmanager-sge

New Query

Direction:
Click to get the subdirectory. represents a leaf node of LDAP directory. Since a leaf does not have any child, cannot be clicked. To get the attributes of a node, click the link with its name, no matter it has any child or not, say o=xmessages, or cn=Manager. Click to return to the upper level.

Fig.3. LDAP Browser Portlet consumed by Sakai as a remote portlet

Besides its core tools and services provided by the Sakai development team, Sakai is now being extended at different universities around the world so that it can meet various (e-Learning) requirements. Here, we are extending Sakai with additional tools for supporting research. For example, with the WSRP consumer we have developed, Sakai is able to provide researchers with transparent access to computational and data grid resources through portlets developed and tested for a previous project and hosted centrally.

2.3 Further Discussions

We describe extensions of Sakai, an e-Learning platform, to support research above. In particular, we talk about using WSRP to make use of remote grid portlets so that researchers can easily make use of the UK NGS grid resources. Whilst currently the main middleware used by the UK NGS is the Globus Toolkit 2.4 [2], our VRE system is flexible enough to include grid resources based on WSRF [7], the WS-Resource Framework. Investigation has been taken to study loading various grid resources (GT2-/GT3-/GT4-based) for execution of grid tasks [11]. This can be done at the portlet level. Since in our example, WSRP is adopted, there is no need to update our VRE system. Update of the remote grid portlets will bring WSRF support to our VRE system through WSRP.

A VRE system is designed to be a platform for researchers to collaborate. The concept of *Web 2.0* agrees well with VRE. It is not easy to give a definition of Web 2.0, but it is well understood that one of the core competencies is to harness *collective intelligence* [10]. The aim of collaboration among researchers is to collect ideas, data (including meta-data), knowledge, etc. so that grand challenges can be tackled. Ideally VRE systems should be connected with internal/ external knowledge and data repositories, to which all researchers should contribute.

Fig.4 describes some typical services a VRE system may provide. Basically VRE systems should provide researchers with services for efficient communication and collaboration, for example, instant messaging, blog and wiki. Moreover, as discussed above, grid can act as a core service since it is ideal for handing CPU- and storage-intensive studies. Repositories, as shown in Fig.4, would provide users knowledge which may be shared among them.

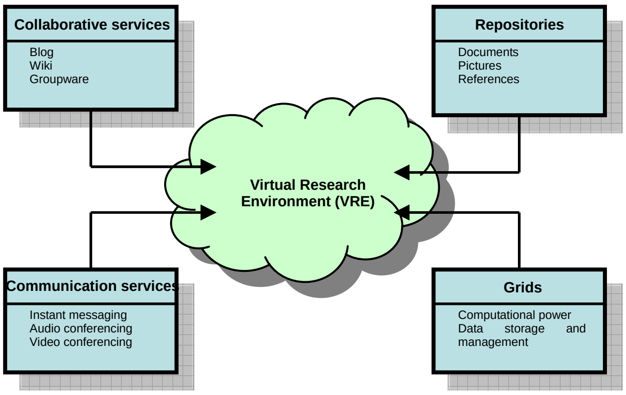


Fig.4. Typical services provided by a VRE system

Clearly there will be more services to be integrated within VRE systems to meet ever increasing demands from academia. Recently at the 7th Sakai Conference, discussions raised demands like version control, granular privacy and archiving. All these will have to be addressed to enhance VRE systems.

3. CONCLUSIONS

In this paper, we discuss the requirements of e-Research. In particular, as an implementation of e-Research, virtual research environments (VRE) have been considered. We argue that collaboration is core to VRE systems. Communication and digital contents management are essential to support today's large-scale research activities. Moreover, we give an example on how to make use of an existing e-Learning framework, Sakai, for construction of a VRE system. Integration of both data-centric and presentation-oriented services has been discussed, the latter one being accomplished by extension of Sakai with a WSRP consumer.

This work was undertaken at the STFC e-Science Centre, Daresbury Laboratory supported by UK JISC (The Joint Information Systems Committee). We would like to thank the anonymous reviewers for their constructive feedbacks for improving this paper.

REFERENCES

[1] EGEE, http://www.eu-egee.org/.
[2] Globus Toolkit 2.4 Release Manuals, http://www.globus.org/toolkit/docs/2.4/.
[3] JISC Virtual Research Environments Programme, http://www.jisc.ac.uk/programme_vre.html.
[4] JSR 168, Java Portlet Specification 1.0, http://www.jcp.org/aboutJava/communityprocess/final/jsr168/.
[5] Sakai: Collaboration and learning environment for education, http://www.sakaiproject.org/.
[6] uPortal, http://www.uportal.org/.
[7] WSRF (WS-Resource Framework), http://www.globus.org/wsrf/.
[8] WSRP (Web Services for Remote Portlets) 1.0, http://www.oasis-open.org/committees/download.php/3343/oasis-200304-wsrp-specification-1.0.pdf.
[9] I. Lawson, R. Butson, e-Research at Otago, http://docushare.otago.ac.nz/docushare/dsweb/Get/Document-3584/eResearch+at+Otago+Report.pdf.
[10] T. O'Reilly, What is Web 2.0, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.
[11] X.D. Wang, X. Yang, R. Allan,"Flexible Grid Portlets to Access Multi Globus Toolkits, International Workshop on Collaborative Virtual Research Environments (CVRE06),"in *Proc. Fifth International Conference on Grid and Cooperative Computing Workshops (GCCW2006)*,pp. 565~570, Changsha, China, Oct 2006.
[12] X. Yang, R. Allan, Sakai VRE Demonstrator Project,"Realise E-Research through Virtual Research Environments,"in *WSEAS Transactions on Computers*, Vol 6, Issue 3, pp. 539~545, Mar 2007.
[13] X. Yang, D. Chohan, X.D. Wang, R. Allan,"A Web Portal for the National Grid Service,"in *Proc. UK e-Science All Hands Meeting 2005*, available on CDROM, pp. 1156~1162, Nottingham, UK, September 2005.
[14] X. Yang, X.D. Wang, R. Allan, M. Dovey, M. Baker, R. Crouchley, A. Fish, M. Gonzalez, T. van Ark, "Integration of Existing Grid Tools in Sakai VRE,""International Workshop on Collaborative Virtual Research Environments (CVRE06),"in *Proc. Fifth International Conference on Grid and Cooperative Computing Workshops (GCCW2006)*,pp.576~582, Changsha,China,Oct 2006.

Xiaobo Yang is a software developer working at the Grid Technology Group, STFC e-Science Centre, in the United Kingdom. He graduated from the Tsinghua University in 1997 (B.Eng in Automotive Engineering) and University of Glasgow in 2003 (Ph.D. in Aerospace Engineering) accordingly. He is interested in Grid portals, collaborative virtual research environments, and service-oriented architecture.

Rob Allan leads the Grid Technology Group, STFC e-Science Centre, in the United Kingdom. His background is as a physicist and – since the mid-1980s – developer of high-performance computing applications using the latest technologies. He has managed several large HPC and e-Science projects in the United Kingdom. He is particularly interested in making the Grid widely usable.

The Implementation of Course Discussion System Using JXTA

Shadi Ibrahim, Qingping Guo
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430063, P. R. China
Email: {shadi,qpguo}@whut.edu.cn

ABSTRACT

Peer-to-Peer (P2P) systems are distributed systems in which each node runs software with equivalent functionality. Developing a system under P2P environment will increase the availability of the system services and reduce hardware commitments. JXTA is an open network-computing platform designed for Peer-to-Peer computing. Its goal is to develop basic building blocks and services to enable innovative applications for peer groups. This paper demonstrates the design and development of P2P course discussion system based on JXTA (P2P-CDS), which provides interactive learning environment through secure group communication and file sharing. Educators (i.e. students, researchers etc.) may cooperate and share their information and data in a secure manner under the P2P environment.

Keywords: Peer-to-Peer, Client-Server, Peer Group, Cooperative System, LDAP and JXTA

1. INTRODUCTION

The positive effects that cooperation has on so many important outcomes, make cooperative learning one of the most valuable tools educators have[1]. Students, teachers and researchers interact, exchange their ideas and share their data in cooperative learning systems. Yet almost all of these systems are based on client-server concept where the data storage and computation are centralized on a small number of high-end servers. As the number of active clients and available information is simultaneously increasing, servers will hardly handle this increment. Redesigning these systems in a more distributed, resource sharing fashion using the Peer-to-Peer approach is a more scalable and robust solution [2].

The term peer-to-peer refers to a class of systems and applications that employ distributed resources to perform a critical function in a decentralized manner. The resources encompass computing power, data (storage and content), network bandwidth, and presence (computers, human, and other resources). The critical function can be distributed computing, data/content sharing, communication and collaboration, or platform services[3]. In P2P systems a very large number of autonomous computing nodes (the peers) pool together their resources and rely on each other for data and services. Every node of the system acts as both client and server and provides part of the overall information available from the system.

In this paper we propose and demonstrate the implementation of a course discussion system which operates in a peer-to-peer environment so we have increased the availability of the system services and have reduced hardware commitments. The JXTA grouping service makes it possible to communicate and share data in secure manner under P2P environment. Implementing our system using the JXTA protocols makes it applicable over the internet across firewalls and NATs (Network Address Translations). Using central Lightweight Directory Access Protocol (LDAP) servers for authentication enhances the security of our system concerning the access rights to the services provided by the system (i.e. students can't create or modify any of the sub-peer groups.) as well as the identity consideration, all our system users can verify each other and this will ensure that only reliable and trusted information will be exchanged among our system users.

In our system we are using already developed application MyJXTA [4]. We extended some features of MyJXTA and defined new access rights and controlled it by using LDAP servers for authentication. All the information about the system's users is kept in the LDAP-aware directories. Our system intends to provide a secure place where users (students) can have beneficial and secure group chatting and they can share their data (documents, codes, applications etc.) in secure and trusted environment. And by using LDAP server for authentication we verify the access rights of each user in addition to identify all the system's users.

2. BACKGROUND AND RELATED WORK

2.1 JXTA Technology

The project JXTA was unveiled by SUN on April 25, 2001[5] and was intended to be a platform on which to develop a wide range of distributed computing applications. JXTA provides a set of XML based protocols to cover typical P2P functionality. As a set of standard protocols, JXTA is independent of any programming language, platform, operating system or device and underlying transports[6]. The Project JXTA protocols create a virtual network overlay on top the existing physical network infrastructure. The Project JXTA virtual network allows a peer to exchange messages with any other peer independently of its network location (firewalls, NATs or non-IP networks[7]). Messages are transparently routed, potentially traversing firewalls or NATs, and using different transport/transfer protocols (TCP/IP, HTTP) to reach the receiving peers (see Fig.1).

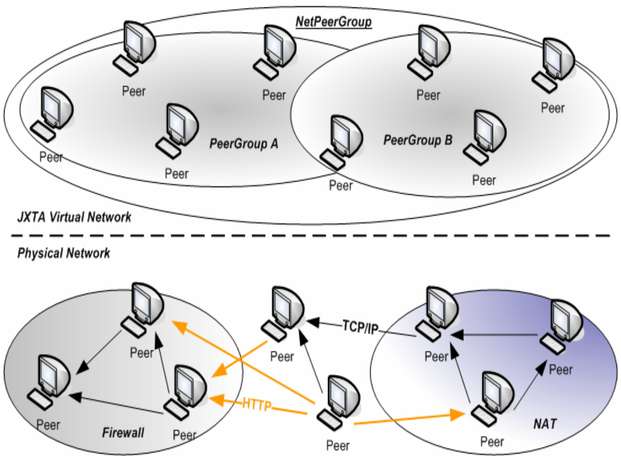


Fig.1. JXTA Virtual Network

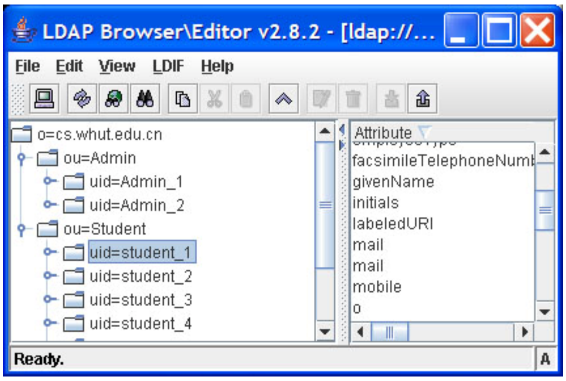
JXTA protocols are composed of six protocols[8]:

- The Endpoint Routing Protocol (ERP) is the protocol by which a peer can discover a route (sequence of hops) used to send a message to another peer.
- The Peer Resolver Protocol (PRP) is the protocol by which a peer can send a generic resolver query to one or more peers, and receive a response (or multiple responses) to the query.
- The Rendezvous Protocol (RVP) is the protocol by which peers can subscribe or be a subscriber to a propagation service.
- The Peer Discovery Protocol (PDP) is the protocol by which a peer publishes its own advertisements, and discovers advertisements from other peers (peer, peergroup, module, pipe and content).
- The Peer Information Protocol (PIP) is the protocol by which a peer may obtain status information about other peers, such as state, uptime, traffic load, capabilities, etc.
- The Pipe Binding Protocol (PBP) is the protocol by which a peer can establish a virtual communication channel or pipe between one or more peers.

2.2 LDAP

Lightweight Directory Access Protocol is a protocol that defines a directory service and the access to that service[9]. By using TCP/IP, LDAP allows clients on multiple platforms (i.e., Windows, Macintosh, and UNIX) to access centralized directory services[10].

It stores information similar to a database but contains more descriptive and attribute-based data. The data is optimized for reading. The information is arranged in a hierarchical structure (see Fig.2), which allows for separation of the data based on different criteria.



o=cs.whut.edu.cn
ou=Admin
uid=Admin_1
uid=Admin_2
ou=Student
uid=student_1
uid=student_2
uid=student_3
uid=student_4

facsimileTelephoneNumb
givenName
initials
labeledURI
mail
mail
mobile
o

Fig.2. LDAP Structure

2.3 Using LDAP Server with JXTA Based Application

Some P2P applications based on JXTA are using LDAP server to verify the security credential and to obtain the identity of system's peers. [11] Presents group membership service for JXTA extended with single or bi-directional authentication using LDAP server, this solution is different than to the already existing implementation [5],[8] of a group access authentication for JXTA using password or null authentication. In [12] the authors present the development of the previous solution to be usable with J2ME application.

[13] Proposed a P2P collaborative research network for legal academics and researchers to facilitate document sharing, and presents a prototype based on JXTA technology. Peers can freely download shared files but their upload permission is controlled by decentralized user authentication using LDAP servers.

3. P2P COURSE DISCUSSION SYSTEM

The P2P course discussion system is an educational system used to provide an interface to a shared environment by combining resource contributions from users into a large pool of resources. In a university and within a school all the P2P-CDS users have a membership in a global peer group called CDGroup (Course Discussion Group). In addition every course is modeled as separate peer group. Each peer group is a child group of the CDGroup; with its membership, users can obtain all of services provided by it.

The membership of the CDGroup is divided into two subsets: administrators and students. By default, students have access to find and discover all the child peer groups and for each joined group they may discuss their ideas, search, add and download the shared files under secure environment. Administrators correspond to the school's administrative staff, they have the ability to create and modify child peer groups.

3.1 P2P Course Discussion System Architecture

P2P-CDS actually uses already developed application MyJXTA (see Fig.3). In addition to MyJXTA we are using central LDAP server for authentication and in order to make them work together we are using Novell LDAP Classes for Java [14].

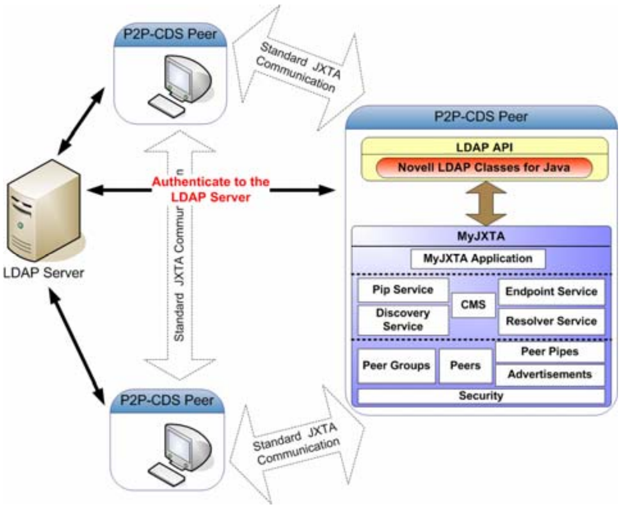


Fig.3. P2P-CDS Architecture

- MyJXTA is a JXTA Technology based collaboration application. Designed from the ground up to be easy to use, modify, extend and deploy [4]. It is an exemplary JXTA application that strives to showcase JXTA best practices for all core lib/apis via deployments of massive scale and provide a framework from which people can learn from and build upon [14].
  Features include: group chat , secure one-to-one chat , anonymous and credentialed group create/find/join/leave , group and 1to1 chat for each joined group , chat history, share search/publish/view/store, drag-and-drop content publishing , pipe search (peer's are associated with pipes) , group search , message listener/filter/directive interfaces.

- Novell LDAP Classes for Java [15]. LDAP Classes for Java enable you to write applications that access, manage, and update information stored in Novell eDirectory or other LDAP-aware directories. These classes are based on the IETF LDAP Java Application Program Interface.

3.2 Join P2P Course Discussion System

After the user successfully log into the JXTA network using the log in procedure provided by MyJXTA, the user is in the default JXTA group NetPeerGroup and also in our system default group CDGroup. To obtain the course discussion peer group services the user should be authenticated using central LDAP server. So an authentication window (See Fig.4) will be popped up to get the userID and password which are already stored in our LDAP directory server.

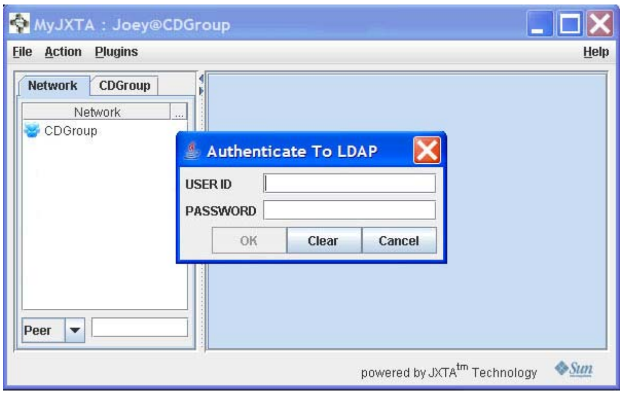


Fig.4. P2P-CDS Log in (Authenticate to The LDAP server)

Novell Java Package is used to develop the LDAP operations methods (see Fig.5). The program first establishes a connection to the LDAP server and then a "bind" operation runs to deliver UserID and user Password. Server controls the user credentials and gives access or returns an error code. If the user is successfully bound to the LDAP server the user role (administrator or student) will be returned as output. The output value determines the access rights for the CDGroup services.

```
.....
import com.novell.ldap.LDAPConnection;
import com.novell.ldap.LDAPException;
import com.novell.ldap.LDAPJSSESecureSocketFactory;
.....
// Settings for LDAP
int LDAP_Version     = 3;
int LDAP_Port        = 389;
String LDAP_Host     = "ldap.cs.whut.edu.cn";
String StaffLogin = ",ou=Staff,o=cs.whut.edu.cn";
String StudentLogin= ",ou=Students,o=cs.whut.edu.cn";

LDAPConnection conn = new LDAPConnection();
......................
 // Authenticate to the LDAP Server
 private int LDAPAuthenticate(LDAPConnection conn, int
   LDAP_Version, String LDAP_Host, int LDAP_Port, String
   My_DN, String My_Password) {
    try {
       // connect to the LDAP server
          conn.connect( LDAP_Host, LDAP_Port );
       /*   bind to the LDAP server.
       My_DN = "uid="+UserID+{StaffLogin Or StudentLogin} */
          conn.bind(LDAP_Version,My_DN,My_Password);
          ……
          return 1;
          }
    catch( LDAPException e ) {
           System.out.println( "Error: " + e.toString() );
          return -1;
          }
      }
.....
```

Fig.5. An Example of LDAP Operations

3.3 Create New Peer Group

Administrators have the ability to create and modify the child peer groups. After successfully logging into the system, the administrator is in our system default group CDGroup. By clicking create Group from the Group menu (seeFig.6) create group action occurs.

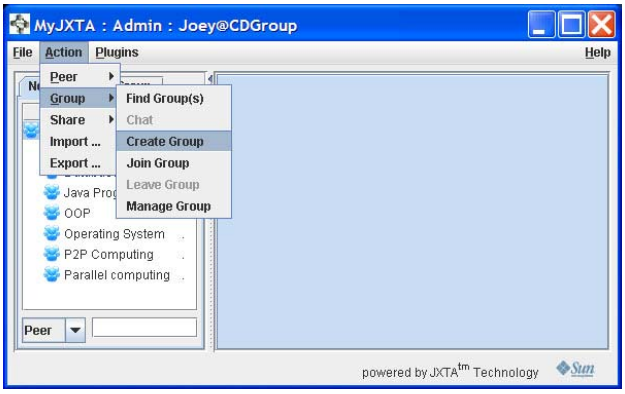


Fig.6. Create New Peer Group

To create new peer group the administrator needs to give a peer group name, password and description. After the administrator inputs these desired information a peer group advertisement will be created which contain information about group ID, module implementation advertisement, name, description and membership service. A form of an example advertisement is (see Fig.7)

```
<?xml version=”1.0” encoding=”UTF-8”?>
<!DOCTYPE jxta:PGA>
<jxta:PGA xmlns:jxta=”http://jxta.org”>
<GID>
    urn:jxta:uuid-A1491744FC9541B5A4B2D33040813A02
</GID>
<MSID>
urn:jxta:uuid-DEADBEEFDEAFBABAFEEDBABE000000010406
</MSID>
<Name>
    P2P Computing
</Name>
<Desc>
    Peer-to-peer course
</Desc>
<Svc>
    <MCID>
urn:jxta:uuid-DEADBEEFDEAFBABAFEEDBABE0000000505
    </MCID>
    <Parm type=”Param”>
        <login>
            myjxtauser :csp2p06 :
        </login>
    </Parm>
</Svc>
</jxta:PGA>
```

Fig.7. Peer Group Advertisement

3.4 Join Peer Group

In JXTA, the membership service is used to apply for peer group membership, to join a peer group, and to exit from a peer group. The membership service allows a peer to establish an identity within a peer group. If a student wants to join in any of the active child peer groups, he must achieve its guaranteed password. These identifying information are achieved offline by entering the course homepage. In order to join a peer group, students need to select a group by clicking on the group name in the Network panel (see Fig.8). A join window will be popped up to get the password of the selected group.

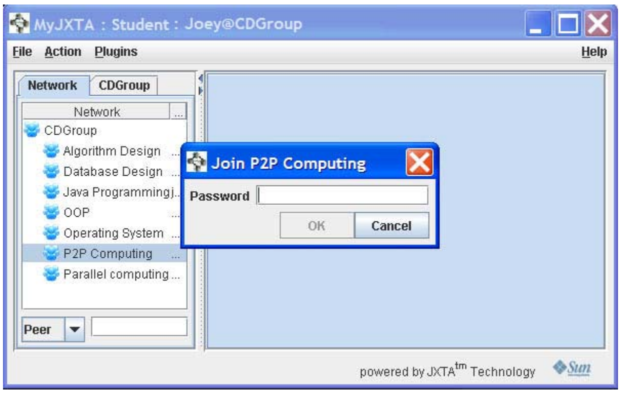


Fig.8. Joining (P2P computing) Peer Group

After successfully logging into the attributive peer group, students can communicate, search and share their data under secure environment.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have presented the design and implementation of a course discussion system based on JXTA, which provides interactive learning environment through secure group communication and file sharing. The course discussion system aims to offer significant advantage in the educational systems by giving students the ability to cooperate and share their information and data in a secure and decentralized fashion.

Our future work consists of completing the implementation of the course discussion system. Once this work is completed, we can test our implementation in our inter-university network and improve the performance evaluation of our system.

REFERENCES

[1] Johnson, D.W. & Johnson, R.T. Cooperative Learning. http://www.co-operation.org/pages/cl.html.

[2] Aberer, K., Punceva, M., Hauswirth, M. & Schmidt, R. (2002) ."Improving Data Access in P2P Systems." *IEEE Internet Computing*. Vol. 6, No. 1. pp: 58-67. http://www.computer.org/internet/ic2002/w1toc.htm.

[3] Milojicic, D.S., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne , J., Richard, B., Rollins, S. & Xu, Z.C. (2003). *Peer-to-Peer Computing.* HP Laboratories Palo Alt, HPL-2002-57 (R.1), July 3rd, 2003.

[4] MyJXTA. http://myjxta2.jxta.org.

[5] JXTA. http://www.jxta.org.

[6] Li, S. (2001). "Making P2P interoperable: The JXTA story." http://www-128.ibm.com/developerworks/java/library/j-p2pint1.html.

[7] LongWork Network. http://www.echelon.com/products/Core/protocol/Default.html.

[8] Traversat, B., Arora, A., Abdelaziz, M., Duigou, M., Haywood, C., Hugly, J.C., Pouyoul, E. & Yeager, B. (2003). Project JXTA 2.0 Super-Peer Virtual Network. http://www.jxta.org/project/www/docs/JXTA2.0protocols1.pdf.

[9] LDAP, http://en.wikipedia.org/wiki/Lightweight_Directory_Access_Protocol.

[10] "What Is LDAP?" https://kb.iu.edu/data/anih.html.

[11] Kawulok, L., Zielinski, K. & Jaeschke, M. (2004). "Trusted Group Membership Service for JXTA." *Proceedings of 4th International Conference Computaional Science* - ICCS 2004, Springer LNCS 3038, Krakow - Poland, June 6-9, 2004. pp: 218-225.

[12] Kawulok, L., Zielinski, K. & Jaeschke, M. (2005). "Trusted group membership service for JXME (JXTA4J2ME)." *Proceeding of IEEE International Conference on Wireless and Mobile Computing, Networking and Communications* 2005. (WiMob'2005), on Vol. 4, 22-24 August, 2005, Montreal, Canada. pp: 116- 121.

[13] Shi, H., Zhang, Y.C., Zhang, J.Y., Beal, E. & Moustakas, N. (2006). "Collaborative Peer-to-Peer Service for Information Sharing Using JXTA." *Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences* on Vol. 1, 20-24 June 2006, Hangzhou, China. pp: 552- 559.

[14] http://wiki.java.net/bin/view/Jxta/WhatIsMyJXTA.

[15] Novell.http://developer.novell.com/wiki/index.php/LDAP_Classes_for_Java.

A Design of Examination Server System Based on TCP/IP Protocol

Xinzhong Zhu
School of Information Science and Engineering, Zhejiang Normal University
Jinhua, Zhejiang 321004, P.R. China
Email: zxz@zjnu.cn

ABSTRACT

This paper applies Visual Basic 6.0 programming language, the Winsock control and the Client-Server model to design an examination server system based on the TCP/IP protocol. This examination system, has overcome some disadvantages of the traditional examination systems. In the practical test and usage, this system has made great achievements in convenient management and high efficiency.

Keywords: TCP/IP protocol, Client-Server, Winsock control, Communication protocol

1. INTRODUCTION

Although the traditional single plane and papery examination system has some advantages of circulating easily and freely without network environment support, it cannot manage the examination comprehensively, confirm the examinees' information effectively. Nor can it control the examinees' reexamination. Also, under this single plane system, each computer chooses one topic so randomly and separately that the topics chosen by two adjacent computers may be the same and repeat; the final records must be collected and input artificially. Along with the development of network technologies, the software network's turn is already an irresistible general trend. Thus, to design an examination system based on the Client-Server model is very influencing and meaningful. This paper tries to propose an examination system adding a certain server model to the traditional client system of single plane version .While the key of this system lies in the design of server model.

2. SYSTEM FUNCTION REQUIREMENTS

The examination works as follows:

Step 1: Start the examination server, initialize the examination environment (import the students' information, delete the related directory and examination document), and then, receive the client requests.

Step 2: Examination starts:

Step 2.1: The students login in at the client software ends. The clients send the examination data such as student numbers, computer names and IP addresses to the examination server.

Step 2.2: After the server receives the data from the PC client, it must confirm the data, described as follows: firstly, it queries the database to make sure whether the student is valid or not. If the student isn't valid, it returns the error message. If the student is valid, the server will check up the student's state (not-taking the exam, taking the exam, and finishing the exam). If the student hasn't taken the exam, it can allow the student to take the exam, and send a signal to the client. Or, it will send other signal to the client.

Step 2.3: After the client has received the message allowing the student to take the exam, the valid student can begin the examination operations in the client system. Or, it will give the relative messages.

Step 3: The student quits during the examination:

Step 3.1: The client sends the student number to the server that the student wants to quit.

Step 3.2: After the server receives the message, it will drop the student's state (taking the exam).

Step 4: During the examination, if the student quits because of the computer malfunction, he must cancel his examination state in the server so that he can login in the client examination system again to go on with his examination.

Step 5: The student finishes the examination:

Step 5.1: The client sends the information of the student's score, number, and examination document to the server.

Step 5.2: Receiving the data, the server will record students' score, and set his state in finishing the exam. It sends a signal of "Received successfully" to the client.

Step 5.3: Receiving "Received successfully", the client will upload the score files and exam result files to the server; Or, it will send the score information again. Repeat more than once, or it will indicates something goes wrong.

Step 5.4: Receiving the uploaded files, the server will store the data in proper dictionary, and send a signal of "Uploaded successfully" to the client.

Step 5.5: Receiving "Uploaded successfully", client will leave the examination system. Or the server will send files continuously. Repeat more than once, or it will indicate something goes wrong.

Step 6: Examination is completed. The server can print and output all student scores by different categories.

Additionally, examination server involves the function of examination information query, student information management, and so on.

3. SYSTEM DESIGN ANALYSES

From the perspective of system function requirements, the key of this examination system lies in the way to connect the clients, the interaction between server and client, the analysis of the exam data, and the control of the mutual exclusive operations in database. This paper mainly discusses the connecting and communication problems between the server and the client based on TCP/IP protocol.

3.1 Relative Protocols and Technologies Introduction

TCP/IP protocol is used in the communication of WAN among different networks, different chips and hosts with different operation systems. TCP is a transmission control protocol, it can provide a data flow service on the reliable and session oriented connection. With automatic error correcting technologies of confirming, flow controlling, and multiplexed synchronism, it could guarantee the reliability and order of the transmission data. IP is an Internet protocol, which means the packet switching protocol. It defines the route rules of data packet transmission between different hosts. Its basic task is to transfer the data packet along the network, and every IP data packet is independent with each other. Thus, the combination of TCP and IP constructs the whole transmission protocol.

With the complexity of TCP/IP protocol taken into consideration, if we develop this exam system directly with the

Winsock control, it will influence the comprehensive application and transmit-ability owning to the diversity of the development platform. And also, much extra burden should be born by the programmer. In order to simplify the difficulty of developing and promote development efficiency, we can accomplish this with Winsock.

Winsock, also known as windows socket, is a development interface for application programs based on TCP/IP protocol. Mainly it has two communication models: one is stream mode corresponding to TCP protocol, and the other is datagram model corresponding to UDP protocol. In this system, the stream mode is adopted. As the net application program interface, socket provides multi-types attributes and methods for programmers, while hides some bottom complicated protocols and data structures. So programmers only need to set some attributes and call some methods to implement the net communication based on TCP or UDP. In this paper, we adopt the Winsock control of Visual Basic 6.0 to implement the communication between server and client.

3.2 The Design of Communication Protocol Based on Client-Server

To implement the information exchange between client and server, we define a group message forms, and denote their meanings shown in the following table 1.

Table 1. A group message form

Object	Operation	Information form
The client	Login request	LOGIN=Test Question NO.: Computer NO.
The server	Invalid student NO.	LOGIN=INVALID
	Taking the exam	LOGIN=DOING
	Finishing the exam	LOGIN=DONE
	Login successfully	LOGIN=OK|Name|Birthday|School|Test Question NO.
The client	Stopping the request	OFF=Test Question NO.: Name
The server		OFF=OK
The client	Request to terminate the exam	END=Test Question NO..:Name:Score
The server		END=OK
The client	Uploading the files	UPL=Size|Student NO.|File Name…|Data
The server		UPL=OK

4. SYSTEM IMPLEMENT

4.1 The Communication Connection

The server program establishes the connection and intercepts the whole net. The progress is described as follows:

(1) Set an array (named by SckServer) of Winsock controls in the server frame, and design the communication model as TCP, that is, intercalate the attributes of the Winsock control as *sckTCPProtocol.*

(2) Set the *LocalPort* attributes of the server program as listening and interception port (intercept the connection demand from client). And the interception port must be integer (so long as it is never used by other TCP/IP application program. In this program, the interception port is 7001).

(3) Make the server program start to intercept the net with the listening method, and wait for connection demand from clients.

The server program starts to intercept the net program code as follows:

```
Private Sub cmdSeverStart_Click(Index as Integer)
If Index = 0 Then
    'The server is starting
    If SckServer (0).State = sckClosed Then
        'The server is listening the demand from clients
        SckServer (0).LocalPort = 7001
        SckServer (0).Listen
    End If
    CmdStart (0).Enabled = False
    cmdStart (1).Enabled = True
    ServerInfoStr = Left ("The server start to run at " & Format(Time, "hh:mm:ss") & vbCrLf & ServerInfoStr, 1000)
    'Establishing the connection
Else
    'The serer is stopping
    If Not (AdcStudentInfo. Recordset. BOF And AdcStudentInfo.Recordset.EOF) Then
        Beep
        Exit Sub
    End If
    If SckServer (0).State <> sckClosed Then
        SckServer (0).Close
    End If
    Cmd Start (1).Enabled = False
    cmd Start(0).Enabled = True
    ServerInfoStr = Left ("The server is stopping at" & Format (Time, "hh:mm:ss") & vbCrLf & ServerInfoStr, 1000)
    Conn.Close
    Set Conn = Nothing
End If
    txtInfo.Text = ServerInfoStr
End Sub
```

After the server program starts up and listens the whole net and before the clients demands connection, server needs to set the RemoteHost property of the client Winsock control as remote host, and set the RemotePort property as communication port (7001). And then, the client program starts the Connect method to demand to be connected with the server program.

The implement procedure of client program (KsClient) is described as follows:

(1) Set a Winsock control (named by SckClient) in the client frame, and its Protocol property as sckTCPProtocol.

(2) Set the remote host name, which is either the computer name or the IP address. If it is the host name, the server should translate it to the corresponding IP address; if it is the IP address of the remote host name, it will connect the server directly.

The proceeding code that the client demands to be connected is described as follows:

```
Public Sub TCP_ Initial ( )
    Max _ Connecting _Time = 1’ (Setting the max connection time)
    ServerIP = "192.168.1.8" ‘(192.168.1.8 is the server’s IP address)
    ServerPort = 7001 ‘(The port is the same as the server’s
listening port)
        ComputerName = sGetComputerName ‘(It is used to get the client computer name)
        Use Name = sGetUserName ‘(It is used to get the client user name)
   End Sub
   Public Sub TCPConnecting(ByVal SendDataItem As Byte, ByVal Connect _ Flag As Boolean)
          TCPSendDataItem = SendDataItem
          TCPConnect _ Flag = Connect _ Flag
          If KsClient.SckClient.State <> sckClosed Then KsClient.SckClient.Close
          KsClient.SckClient.Connect ServerIP, ServerPort
          KsClient.Timer.Interval = 1000 * Max_Connectting_Time
   End Sub
```

The client sends the connection request, and invokes the Connection Request event, which gets a parameter request ID. When the server program is listening, it could use Accept method to receive the client request named by request ID. (The original socket is unchanged, and a new socket is used to establish the connection.). Thus, the server program uses the *SendData* method to send data. Remark: the Accept method must regard the above requested ID as its parameter.

The Connection Request event code of the server program is as follows:

```
   Private Sub Scksever_ConnectionRequest (Index As Integer, ByVal RequestId As Long)
   Dim i As Integer
   Dim WhichSocket As Integer
   If Index <> 0 Then Exit Sub
   For i = 1 To SockNum
   'Checking up whether there are some loaded Winsock controls which haven’t been connected.
       If SckServer(i).State = sckClosed Then
            WhichSocket = i
            Exit For
        End If
   Next i
   If WhichSocket = 0 Then
   'If the number of connected clients is larger than that of socket, more sockets would be loaded.
        SockNum = SockNum + 1
        Load SckServer(SockNum)
        WhichSocket = SockNum
        If SockNum > UBound (bConflict) Then
          ReDim Preserve bConflict(SockNum)
        End If
   End If
   SckServer(WhichSocket).Accept RequestId ‘Connecting the client by the new loaded socket
   ServerInfoStr = Left (SckServer(WhichSocket).RemoteHostIP & "Socket:" & WhichSocket & "Requesting Web Server" & vbCrLf & ServerInfoStr, 1000)
   txtInfo.Text = ServerInfoStr
   Call DrawSocket
   End Sub
```

4.2 Data Transmission

After the server receives the request from the client program, the client program generates Connect event, sending the data by *SendData* method. In this procedure, the client program will send the relative information of the students (such as exam No., name, computer No.) to the server. With this information, the server will inquire about the student database to validate the student information. If the student information is valid, the server proceeding will send a message of LOGIN=OK to the client, and the test string will set the state of this student state as LOGIN=DOING, indicating the student is in the state of taking-the-exam and permit the student to continue. If the students’ information isn’t valid, the server proceeding will send LOGIN=ERR: name|test question string, which means that the student isn’t valid. If the students’ information is valid but the state is presented as LOGIN=DOING, it will give a hint that the proceeding has gone wrong. Therefore, this procedure can avoid that the student log in the exam system more times because of computer malfunction and restarting the computer.

When the student finishes the exam, the client will call the *SendData* method to send the exam results (such as score files) to the server. Receiving the data, the server will generate the Data Arrival event. And the parameter *BytesTotal* includes all of the received data byte. In this event, the *GetData* method is called to receive the data.

4.3 Inishing The Connection

When the application program has received the data, the connection must be closed to release the system resource. When one (a client or the server) closes the connection, it can use the *Close* method. And the other one receives the *Close* event, it will use *Close* method to close TCP/IP connection.

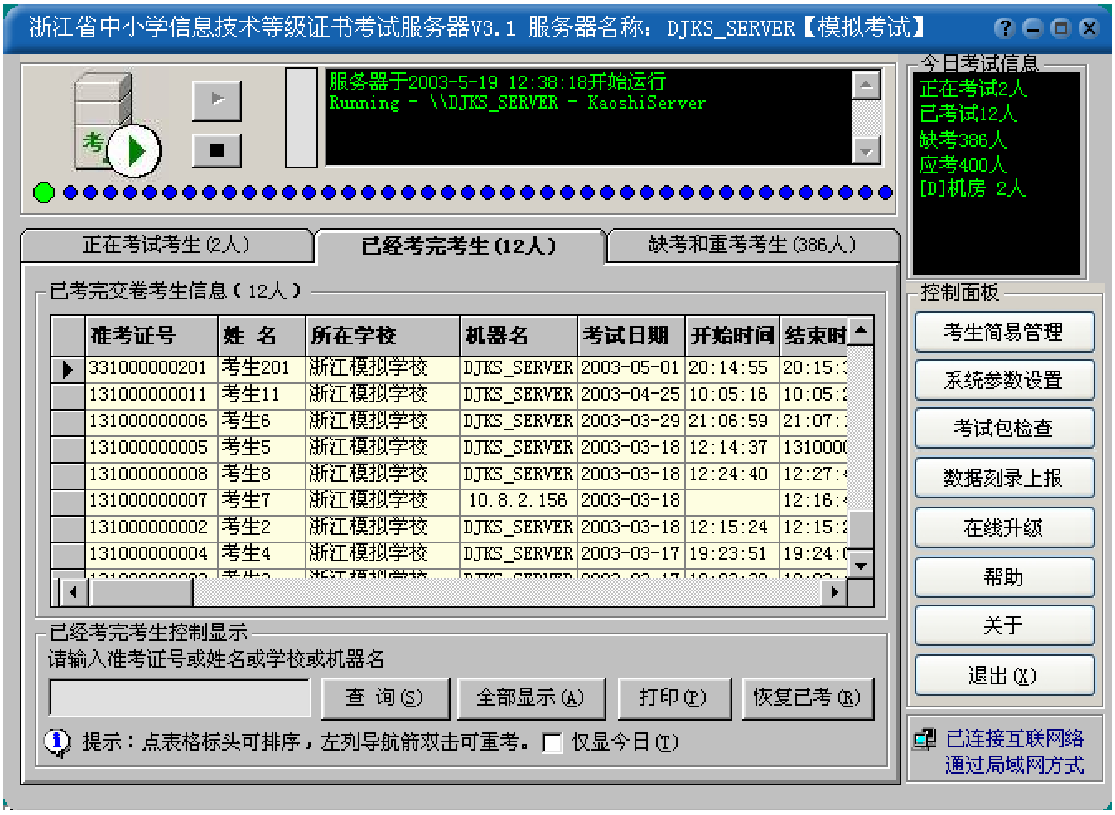
The application code of closing connection is as follows:

```
   SckClient.Close     ‘closing the connection
   Private Sub SckServer_ Close ()
         SckServer(Index).Close           'Closing the connection
         Unload SckServer(Index)          'Unloading the control
   End Sub
```

Therefore, an examination system designed based on TCP/IP protocol is mainly completed. If the system is implemented completely, it needs the database operations and programs. Because the system is a one-to-many relation of the server and the clients, in a short time, there are many clients connecting to the server. Thus, there may be many connections operating the database. So this system also needs to resolve the mutual exclusive operation problem. One way is to modify the parallel mutual exclusive operations (reading and writing, writing and writing) to serial ones. From the whole exam procedure, the number of connections operating the database at an average time is small. So we could establish a database operation instruction queue, which stores the writing operations about the database. And we design a database operation instruction processed program to execute these instructions. Because this paper only discusses how to communicate between the clients and the server based on TCP/IP, the content of data processed program wouldn’t be described in details. The software user interface of this system is showed in the following Fig.1.

5. CONCLUSIONS

The design and implement of this system could compensate for some disadvantages of the traditional exam system. And it could avoid a lot of problems occurring, such as taking the exam more than one time, repeated topics and not validating the identities of examinees rightly. It also improves the performance and efficiency of the traditional exam systems. Additionally, this system could manage the data of all the examinees uniformly and normatively.



服务器于2003-5-19 12:38:18开始运行
Running - \\DJKS_SERVER - KaoshiServer

正在考试2人
已考试12人
缺考386人
应考400人
[D]机房 2人

准考证号	姓 名	所在学校	机器名	考试日期	开始时间	结束时
331000000201	考生201	浙江模拟学校	DJKS_SERVER	2003-05-01	20:14:55	20:15:
131000000011	考生11	浙江模拟学校	DJKS_SERVER	2003-04-25	10:05:16	10:05:
131000000006	考生6	浙江模拟学校	DJKS_SERVER	2003-03-29	21:06:59	21:07:
131000000005	考生5	浙江模拟学校	DJKS_SERVER	2003-03-18	12:14:37	131000
131000000008	考生8	浙江模拟学校	DJKS_SERVER	2003-03-18	12:24:40	12:27:
131000000007	考生7	浙江模拟学校	10.8.2.156	2003-03-18		12:16:
131000000002	考生2	浙江模拟学校	DJKS_SERVER	2003-03-18	12:15:24	12:15:
131000000004	考生4	浙江模拟学校	DJKS_SERVER	2003-03-17	19:23:51	19:24:

请输入准考证号或姓名或学校或机器名

提示：点表格标头可排序，左列导航箭双击可重考。 □ 仅显今日(T)

Fig.1. The software user interface of this system

ACKNOWLEDGEMENTS

We would like to acknowledge the important contribution of B.Y. Jin and J.M. Zhao, professors in school of information science and engineering of Zhejiang Normal University. Thanks to the financial support from the 15 programming point topics project-the primary and junior high school information technique grade certification of Zhejiang Province information education examination system, It is supported by the department of education in Zhejiang province.

RRFERENCE

[1] Lin Yong, *Visual Basic Programmer Windows API Programming Handbook*, Posts & Telecom Press,2002.6.

[2] Lin Yong , Zhang Leqiang, *Visual Basic 6.0 User Programming Handbook*, Posts & Telecom Press, 2002,6.

[3] MICROSOFT, VISUAL BASIC 6.0 Programmer Guide, Beijing Hope Electron Press, 1998.

[4] Wang Gang , Lin Zhi,, *TCP/IP Programming Based on Windows*, Tsinghua University Press,2002,3.

Xinzhong Zhu, associate professor. His research interests are in image processing, pattern recognition for biometrics, computer simulation, manufacturing informatization, and software engineering.

An e-Learning System based on Domain Ontology

Xin Qi, Qianxing Xiong, Yuqiang Li
College of Computer Science and Technology, Wuhan University of Technology,
Wuhan 430063, China
Email: qixin@whut.edu.cn

ABSTRACT

In this paper we will give an overview of an e-Learning System based on domain ontology and discuss some of Semantic Web technologies. It is primarily based on ontology-based descriptions of content, context and structure of the learning materials and thus provides flexible and personalized access to e-Learning materials. At last, we illustrate how the techniques can be put into programming practice using the modern Semantic Web development tool Protégé.

Keywords: Ontology, Domain Model, Semantic Web, E-Learning, ProtÉGÉ

1. INTRODUCTION

E-Learning, or online learning, stands for all forms of Internet-enabled and/or computer supported learning. It refers to the use of computer and computer network technologies to create, deliver, manage and support learning, usually independent of specific locations or times [1]. E-Learning can involve complete online courses, where all aspects of learning, from learner enrollment to tuition and support take place online.

Many modern e-Learning systems consist of object-oriented components, implemented in mainstream programming languages like Java or C#. However, the promise of reusability of object-oriented models is often not fulfilled. In many cases, domain models contain hard-coded dependencies with the specific application. Especially once the model is encoded in a programming language, much of the knowledge that went into the initial design is lost. Another typical problem with such systems is interoperability. The result of this e-Learning system development reality is that much time is wasted with unnecessary duplicate work.

2. SEMANTIC WEB & E-LEARNING

The vision behind the Semantic Web is to make web content machine-readable so that it can be more easily analyzed by software agents and shared among Web Services. In order to exploit the benefits of Semantic Web technology in the context of e-Learning system development, the design patterns and strategies need to be understood to seamlessly integrate these technologies. While we are beginning to understand the implications of Semantic Web technology in e-Learning system development, many promising candidate solutions are beginning to emerge, including domain model, metadata, software architecture, programming APIs and code generators.

2.1 Ontology as Domain Model

Domain models can describe the relevant concepts and data structures from an application domain and encode knowledge that is useful to drive an application's behavior. Modern software development tools with support for the UML and code generation allow for developers to synchronize and verify technical implementation with user requirements using domain models.

Domain Ontology typically consists of definitions of concepts relevant for the domain, their relations, and axioms about these concepts and relationships. The World Wide Web Consortium (W3C) is recommending a number of Web-based ontology languages that can be used to formalize domain models. RDF (Resource Description Framework) Schema and OWL (Web Ontology language) can be used to describe classes, attributes and relationships similar to object-oriented model languages.

At the core are the languages RDF Schema and OWL, OWL being optimized to represent structural knowledge at a high level of abstraction. Domain models encoded in OWL can be uploaded on the Web and shared among multiple applications. OWL is supported by an unambiguous dialect of formal logic called Description Logic [2]. This formal underpinning makes it possible to exploit intelligent reasoning services such as automatic classification and consistency checking. These services can be used at build-time and therefore facilitate the construction of reusable, well-tested domain models. Reasoning services can also be used at runtime for various purposes. For example, this makes it possible to define classes dynamically, to re-classify instances at runtime and to perform complex logical queries. In addition to their foundation on logics, OWL and RDF Schema operate on similar structures to object-oriented languages, and therefore can be effectively integrated with traditional software components.

2.2 Ontology-based Metadata for E-Learning

Metadata is the Internet-age term for information that librarians traditionally have used to classify books and other print documents. Metadata tagging enables organizations to describe, index, and search their resources and this is essential for reusing them. Different communities have developed their own standardized metadata vocabularies to meet their specific needs. In the e-Learning community three metadata standards are emerging to describe e-Learning resources: IEEE LOM (http://ltsc.ieee.org/doc/wgl2/LOM3.6.html), ARIADNE (http://ariadne.unil.ch/Metadata/) and IMS (http://www.imsproject.org/metadata/imsmdylp2/imsmd_infoylp2.html). Those meta-models define how learning materials can be described in an interoperable way. However, most of those metadata standards lack a formal semantics.

The role of domain ontology is to formally describe shared meaning of the used vocabulary (set of symbols). From the student point of view the most important criterions for searching learning materials are: what the learning material is about (content) and in which form this topic is presented (context). However, while learning material does not appear in isolation, another dimension (structure) is needed to encompass a set of learning materials in a learning course. The shared-understanding problem in learning material occurs on several aspects, such as content, context and structure.

2.3 Course Ontology

The backbone of the e-Learning system is the course ontology presented partially in the Table 1. The ontology definition contains an is-a hierarchy of relevant domain concepts, relations between these concepts, further properties of concepts (attributes with value ranges), and the derivation rules to infer new knowledge [3]. The leftmost column shows the concepts of the domain organized in the is-a hierarchy. For example, “PhDStudent” is a subconcept of the concept “Student“. Attributes and relations of concepts are inherited by subconcepts. Multiple inheritance is supported as a concept may fit into different branches of the taxonomy. Attributes and relations of the concepts appear in the middle column in the Table 1. Relations refer to other concepts, like “has Author” denoting a relation between the concept “Document“ and the concept “Author“. The rightmost column shows some rules of the course ontology. For example, the fourth rule in Table 1 asserts that whenever a document D2 is known to have a child document D1 then D2 has D1 as its parent document. This kind of rules completes the knowledge and frees a knowledge provider to provide the same information at different places reducing the development as well as the maintenance efforts. The ontology representation language is F-Logic . Roughly, the statements ConceptX::ParentX and ConceptX[relationXY=>>ConceptY] could be read as ConceptX is a subconcept of the concept ParentX and ConceptX is in the relation relationXY with ConceptY.

The course ontology consists of content, context and structure ontology, mentioned in the previous section. The content ontology is visible in the description of domain terms like “Protocol”, “Service”, “Topology”. The relation “hasTopic” and the first two rules are also a part of the content ontology. The first rule determines the transitive property of the “hasTopic” relation. For example, based on the first rule and on the facts that “e-Learning hasTopic TeleTeaching” and that “TeleTeaching hasTopic WebBasedLearning”, the fact “e-Learning hasTopic WebBasedLearning” is concluded. The second rule ensures that whenever a document with the content “e-Learning” is searched for, then the documents about “TeleTeaching” and “WebBasedLearning” are also found.
The context ontology is based on the pedagogical model. Concepts like “Introduction”, “Explanation”, “Example” are used to describe several types of contexts for the learning materials.

The most important part of the structure ontology are the relations between learning materials (“preDocument”, “nextDocument”, “IsBasedOn”, “IsBasisFor”) and corresponding rules. The learning materials are organized in a tree structure. The relations “preDocument” and “nextDocument” describe a sequence of the documents at the same level in the structure tree of the learning materials. The relations “parentDocument” and “firstChildDocument” correspond to the references between two successive structure levels. The rules in the structure ontology enable a flexible semantic navigation through the learning materials organized into a course. For example, the rule “FORALL D1, D2 D1:Document[prevDocument->>D2]<-> D2:Document[nextDocument->>D1].” enables to go through the learning materials in two direction (forward or backward), even though only one“path” is defined. The concepts “Course”, “Module” and “Atom” are also part of the structure ontology. They are used to indicate the complexity of the learning materials. The simplest type of the learning materials is an “Atom”. It is a learning material that doesn’t contain any other learning material. The “Module” consists of several atoms organized in a sequence and a “Course” is a sequence of modules or other courses. In this way a course is a tree structure of learning materials on different granularity levels. Complex structures can be derived automatically from more elementary ones by exploiting the last rule in Table 1.

Table1 Partial ontology in the e-Learning scenario

Concept	Relation	Rule

Object [ ].
Document :: Object.
...
Content :: Object.
Protocol :: Content.
Service :: Content.
Topology :: Content.
Bustopology :: Topology.
Circletopology::Topology.
...
Context::Object.
Introduction:: Context.
Explanation:: Context.
Example:: Context.
Figure::Example.
...
Structure::Object.
Course:: Structure.
Module:: Structure.
Atom:: Structure.
...
Person::Object.
Author :: Person.
Student :: Person.
PhDStudent :: Student.
...
	Document [
	name=>>String;
	title=>>String;
	path=>>String;
	hasAuthor=>>Author;
	content=>>Content;
	context=>> Context;
	structure=>> Structure;
	...
	prevDocument =>> Document;
	nextDocument =>> Document;
	firstchildDocument =>> Document;
	parentDocument =>> Document;
	relatedDocuments =>> Document;
	...
	IsBasedOn=>>Document;
	IsBasisFor=>>Document;
	...].
	Content[
	hasTopic=>>Content].
		FORALL A, B, C
		A[hasTopic->>C] <- A:Content and A[hasTopic ->>B] and B:Content and B[hasTopic ->> C] and C: Content.
		
		FORALL D, C1, C2
		D:Document[content->>C1] <- C1:Content and C2:Content and D:Document[content->>C2] and C1[hasTopic->>C2].
		
		FORALL D1, D2
		D2:Document[prevDocument->>D1] <- EXISTS E1, E2, C C:Content and D2:Document[context->>E2] and E2:Example and D1[context->>E1] and E1:Explanation and D1[content->>C] and D2[content->>C].
		
		FORALL D1, D2
		D1:Document[parentDocument->>D2]
		<-D2:Document[firstchildDocument->>D1].
		
		FORALL D1, D2
		D1:Document[prevDocument->>D2]
		<->D2:Document[nextDocument->>D1].
		
		FORALL D, S
		D:Document[structure->>S:Course] <-
		Exists D1, S1 D1:Document and (S1:Course or S1:Module) and D1[structure->>S1] and D1[parentDocument->>D].
		…

3. ARCHITECTURE OF AN E-LEARNING SEMANTIC WEB APPLICATION

Domain ontologies such as those described in the previous section can be exploited by different Semantic Web applications. Fig.1 illustrates the software architecture of an application that finds appropriate courses for a student. The functionality of this application is made available to software agents through a Web Service interface, and to end-users through a conventional Web browser interface [4]. Input to these services is in both cases a collection of data objects about a student (e.g., age, grade preferences, majors, budget). The output is a list of suitable e-Learning courses together with a list of suggested learning material. These input and output data structures are formally represented in terms of OWL ontologies, so that external agents can correctly use the service.

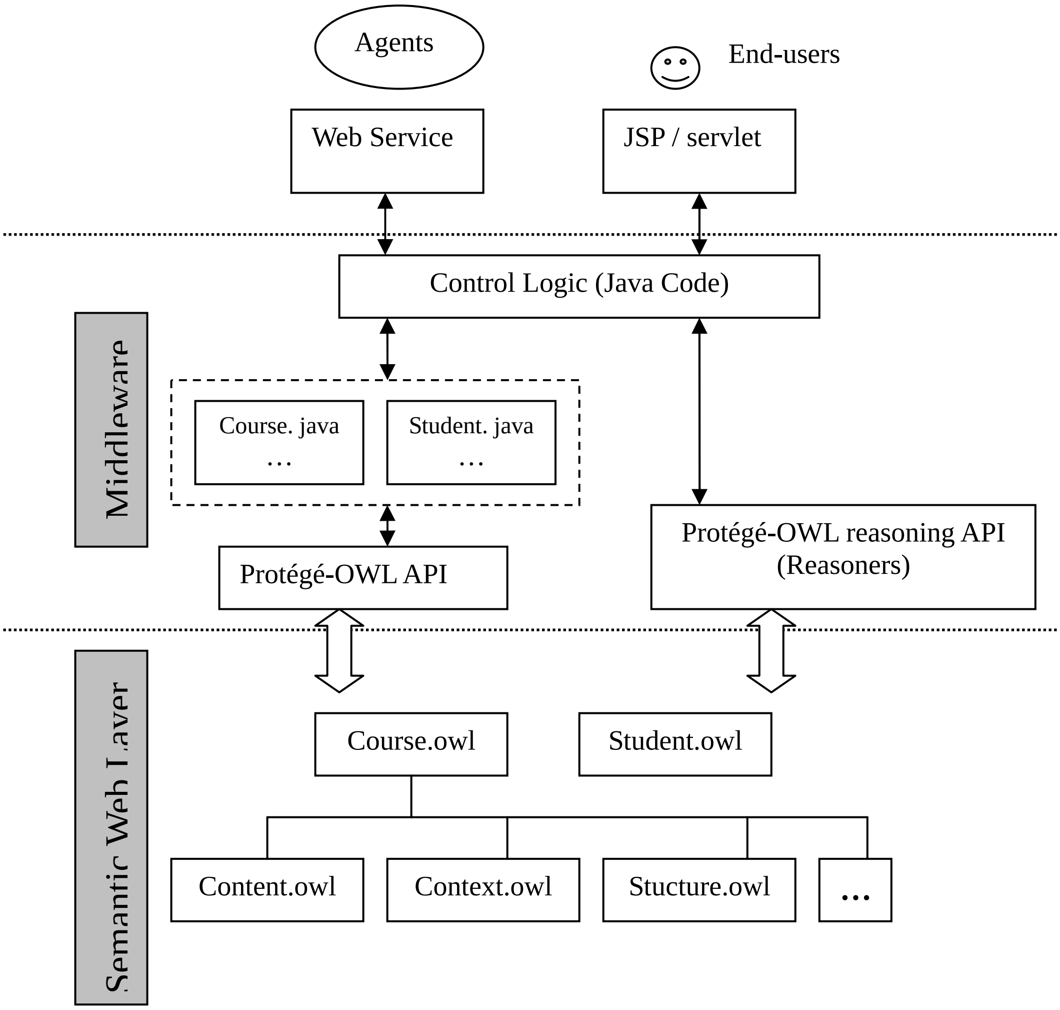


Fig.1. Architecture of a Semantic Web Application

Much of the application logic itself is implemented in a conventional object-oriented language such as Java. For example, the system must manage databases, sessions, and the user interface. The application needs to represent the objects that are exchanged between the application and other services or the user interface as Java objects.

In addition to the rather simple input/output data structures, ontologies are also used to represent the background knowledge that is needed by the application to fulfill its task. There are some core ontologies that define the basic structure of this knowledge by means of base classes. These base classes can be extended and instantiated by external ontology providers on the Semantic Web. While the base classes can and must be hard-wired into the executable system, the knowledge encoded in the external ontologies can only be used by generic reasoning engines or rule execution engines.

4. PROGRAMMING WITH DOMAIN ONTOLOGY

The programmatic access of domain ontology and manipulation of knowledge bases using ontology APIs requires special knowledge by the developers. Therefore an intuitive approach for object-oriented developers is desirable. This can be achieved by ontology tools that generate an API from the ontology, e.g. by mapping concepts of the ontology to classes in an object-oriented language. The generated domain object model can then be used managing models, inferencing, and querying. Tools supporting those features are already available today.

Modern ontology development tools such as Protégé with the OWL Plug-in allow users to exploit these ontologies conveniently, and provide intelligent guidance to find mistakes similar to a debugger in a programming environment.

The Protégé-OWL API is an open-source Java library for the Web Ontology Language (OWL) and RDF(S). The API provides classes and methods to load and save OWL files, to query and manipulate OWL data models, and to perform reasoning based on Description Logic engines [5].
In domain model of e-Learning, Course orders associate a student with a list of learning materials such as Fig.2. An object-oriented design such as the following UML class diagram may be come up after some thinking.

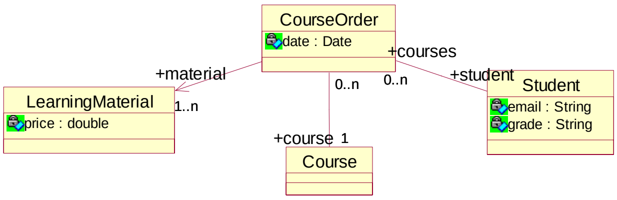


Fig.2. A simple domain model in UML syntax.

Assuming we want to build a Java application around this model, we need to access the objects in the ontology and the run-time objects, e.g., the individual courses and students. To get a feeling of how to use the Protégé-OWL API [6], the following example Java code snippet (a method that calculates the sum of all courses of a given student) has been provided:

```
public static float getPurchasesSum(RDFIndividual student) {
OWLModel owlModel = student.getOWLModel();
double sum = 0;
RDFProperty coursesProperty =
owlModel.getRDFProperty(“courses”);
RDFProperty materialProperty =
owlModel.getRDFProperty(“material”);
RDFProperty priceProperty =
owlModel.getRDFProperty("price");
Iterator courses =
student.listPropertyValues(coursesProperty);
while(courses.hasNext()) {
RDFIndividual course = (RDFIndividual) courses.next();
RDFIndividual material =
(RDFIndividual)purchase.getPropertyValue(materialProperty)
;
Double price =
(Double)material.getPropertyValue(priceProperty);
sum += price.doubleValue();
}
return sum;
}
```

5. CONCLUSIONS

As a consequence of these discussions, most Semantic Web applications, such as e-Learning system, will have a similar architecture around core ontology, external ontology, control components, and (user) interfaces. Software Development based on Semantic Web Technology also follows a very similar approach, but applies these ideas in an extreme way: domain models are not only used for code generation, but they are used as executable artifacts at run-time. A goal in future work should be to further leverage the role of declarative domain models in executable systems.

Another purpose of this paper was to clarify possibilities of using the Domain Ontology as a backbone for e-Learning. Primarily, the objectives are to facilitate the contribution of and the efficient access to information. But, in general, a Semantic Web-based learning process could be a relevant (problem-dependent), a personalized (user-customised) and an active (context-sensitive) process. These are prerequisites for realizing efficient learning. This new view enables us to go a step further and consider or interpret the learning process as a process of managing knowledge in the right place, at the right time, in the right manner in order to satisfy business objectives.

REFERENCES

[1] Derntl, M. and K.A. Hummel. “Modeling context-aware e-learning scenarios.” *3rd IEEE International Conference on Pervasive Computing and Communications. Kauai Island: IEEE Computer Society*. 2005.

[2] Saartje Brockmans, Andreas Eberhart, Raphael Volz. “Visual modeling of OWL DL ontologies using UML” "[A]. In S.A. McIlraith et al., *Proceedings of the Third International Semantic Web Conference*[C], Hiroshima, Japan, 2004, pp. 198-213. Springer, November 2004.

[3] Ljiljana Stojanovic, Steffen Staab. “eLearning based on the Semantic Web” [A]. In: *WebNet2001 - World Conference on the WWW and Internet*[C], Orlando, Florida, USA, 2001

[4] Holger Knublauch. “Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with Protégé/OWL” [A]. In: *1st International Workshop on the Model Driven Semantic Web (MDSW2004) [C]. Monterey*, USA, 2004

[5] *The Protégé Owl API*, http://protege.stanford.edu/plugins/owl/api/guide.html

[6] “A Semantic Web Primer for Object-Oriented Software Developers.” In: W3C Working Group Note 9 March 2006, http://www.w3.org/TR/2006/NOTE-sw-oosd-primer-20060309/

Xin Qi is a lecturer of Wuhan University of Technology. He graduated from Wuhan University of Technology in 2000; obtain his master's degree in computer application technology from Wuhan University of Technology in 2003. Presently he is specializing in doctor's degree in Wuhan University of Technology. His research interests are in intelligent technology, software engineering and object-oriented technology. Email: qixin@whut.edu.cn

Qianxing Xiong is a Professor and doctoral supervisor in the Wuhan University of Technology.

Yuqiang Li is a lecturer of Wuhan University of Technology.

Frameworks of Computer-Mediated-Communication in E-education

Changzheng Liu[1], Guiyun Ye[2]
[1]College of Computer Science and Technology, Harbin University of Science and Technology
Harbin, Hei Longjiang, 150080, P.R.China
[2]College of Electrical and Information Engineering, Heilongjiang Institute of Science and Technology
Harbin, Hei Longjiang, 150027, P.R.China
[1]Email: fox@hrbust.edu.cn, [2]Email: yeguiyun@yahoo.com.cn

ABSTRACT

Computer-Mediated-Communication (CMC) is fast becoming a big part of our daily lives. More and more people are increasingly using the computer to communicate and interact with each other. The internet and its advantages of connectivity, enable CMC to be used from a plethora of applications. Most common uses of CMC include email communication, discussion forums as well as real time chat rooms and audio/video conferencing. By communicating through computers and over the internet, online communities emerge. Discussion boards and other CMC applications offer a huge amount of information and the analysis of this data assists in understanding these online communities and the social networks that form around them. There have been various frameworks by different researchers aimed at analyzing CMC. This paper's main objective is to provide a complete overview of the models and frameworks available that are being used for analyzing CMC in e-Learning environments. The significance of the proposed presentation is that it aims to provide the reader with up-to-date information regarding these methods. Advantages and disadvantages of each of the CMC analysis methods are presented and suggestions for future research directions are made. Finally, these suggestions are applied to characteristic scenario in e-Learning.

Keywords: Computer-Mediated-Communication, E-education, and E-Learning, Social Network Analysis.

1. INTRODUCTION

The focus of this study is to introduce the reader to the concept of Computer-Mediated Communication (CMC) and Online Communities. Furthermore, we discuss the various types of CMC analysis that can take place. The purpose of each framework is described along with its strengths and weaknesses. The paper begins with a literature review of CMC and Online Communities, and continues with the evaluation of the existing frameworks. Finally, we draw conclusions based on the advent of new technologies and platforms that are available, as to whether or not these frameworks are up-to-date in analyzing CMC as it exists today.

1.1 CMC

It is by now no secret how vital the Internet was, is, and will continue to be in our lives. One of the most important characteristics of this medium is the opportunities it offers for human-human communication through computers and networks. As Metcalfe (1992) points out, communication is the internet's most important asset and e-mail is the most influential aspect. E-mail is just one of the many modes of communication that can occur through the use of computers.

Jones (1995) points out that through communication services like the Internet, Usenet and bulletin board services that are electronically-distributed, almost instantaneous, written communication has for many people supplanted the postal service, telephone, even fax machine. All these applications where the computer is used to mediate communication are called Computer-Mediated Communication or CMC.

"Computer-Mediated Communication (CMC) is the process by which people create, exchange, and perceive information using networked telecommunications systems (or non-networked computers) that facilitate encoding, transmitting, and decoding messages. Studies of CMC can view this process from a variety of interdisciplinary theoretical perspectives by focusing on some combination of people, technology, processes, or effects. Some of these perspectives include the social, cognitive/psychological, linguistic, cultural, technical, or political aspects; and/or draw on fields such as human communication, rhetoric and composition, media studies, human-computer interaction, journalism, telecommunications, computer science, technical communication or information studies" (December, 1997, pp.1).

1.2 Online Communities

Through the use of CMC applications, online communities emerge. As Korzeny pointed out even as early as 1978, the new social communities that are built from CMC, are formed around interests and not physical proximity (Korzeny, 1978). Another point to note, is that CMC and the Internet give people around the world the opportunity to communicate with others who share their interests, as unpopular as these interests may be, which does not happen in the 'real' world where the smaller a particular scene is, the less likely it will exist. This is due mainly to the internet's connectivity and plethora of information available and posted by anyone anywhere in the world.

The term online community is multidisciplinary in its nature, means different things to different people, and is slippery to define (Preece, 2000). The relevance of certain attributes in the descriptions of online communities, like the need to respect the feelings and property of others, is debated (Preece, 2000).

Online communities are also referred to as cyber societies, cyber communities, web groups, virtual communities, web communities, virtual social networks and e-communities among several others.

For purposes of a general understanding of what virtual communities are, we present Rheingold's definition. "Virtual communities are social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace" (Rheingold, 1993, pp.5).

＊ This work is supported by the Natural Sciences Foundation of Heilongjiang Province under Grant QC04C44, Doctor Foundation mgb05006, Natural Sciences Foundation of China mgz06011 and the Foundation of office of Education of Heilongjiang Province under Grant 11511070.

There are many reasons that bring people together in online groups. These include hobbies, ethnicity, education, beliefs and just about any other topic or area of interest. Wallace (1999) points out that meeting in online communities eliminates prejudging based on someone's appearance, and thus people with similar attitudes and ideas are attracted to each other. People are using the internet to make friends, colleagues, lovers, as well as enemies (Suler, 2004).

Preece, Rogers and Sharp (2002) state that an online community consists of people, a shared purpose, policies and computer systems while identifying the following member roles: Moderators and mediators: who guide discussions/serve as arbiters; Professional commentators: who give opinions/guide discussions; Provocateurs: who provoke; General Participants: who contribute to discussions; Lurkers: who silently observe.

2. CMC ANALYSIS FRAMEWORKS

As mentioned earlier, the Internet plays a vital role in socially connecting people worldwide. The virtual communities that emerge have complex structures, social dynamics and patterns of interaction that must be better understood. Through the use of CMC we are provided with a richness of information and pools of valuable data ready to be analysed.

There are various aspects and attributes of CMC that can be studied. Three important and widely used types of CMC analysis are Content Analysis, Human-Human Interaction Analysis and Human-Computer Interaction Analysis.

2.1 Content Analysis

Content analysis is an approach to understanding the processes that participants engage in as they input messages (McLoughlin, 1996). There have been several frameworks created for studying the content of messages exchanged in CMC. Examples include work from Archer, Garrison, Anderson & Rourke (2001) and McCreary's (1990) behavioral model which identifies different roles and uses these roles as the units of analysis. Furthermore, in Gunawardena, Lowe, and Anderson's (1997) model for examining the social construction of knowledge in computer conferencing, five phases of interaction analysis are identified and these are:

(1) Sharing/Comparing of Information;
(2) The Discovery and Exploration of Dissonance or Inconsistency among Ideas, Concepts or Statements;
(3) Negotiation of Meaning/Co-Construction of Knowledge; (IV) Testing and Modification of Proposed Synthesis or Co-Construction;
(4) Agreement Statement(s)/Applications of Newly Constructed Meaning.

Henri (1992) has also developed a content analysis model for cognitive skills and is used to analyze the process of learning within the student's messages. Mason's work (1991) provides descriptive methodologies using both quantitative and qualitative analysis.

2.2 Human-Human Interaction Analysis

Over the years there have been several models by different researchers for analyzing interaction. It is important to note that the type of interaction studied in this case is interpersonal interaction, more specifically the human-human interaction that takes place through the use of CMC. Examples of Interaction Analysis models include but are not limited to Bale's Interaction Process analysis (Bales, 1950; Bales & Strodbeck, 1951), the SIDE model (Spears & Lea, 1992), a four-part model of cyber-interactivity (McMillan, 2002) and Vrasidas's (2001) framework for studying human-human interaction in Computer-Mediate Online Environments. We have found the technique called Social Network Analysis (SNA) to be more suitable for analyzing CMC in e-Learning and explain it in more detail.

"Social Network Analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships" (Krebs, 2004, pp.1). Preece (2000) adds that it provides a philosophy and set of techniques for understanding how people and groups relate to each other, and has been used extensively by sociologists (Wellman, 1982; Wellman 1992), communication researchers (Rice, 1994; Rice et al., 1990) and others. Analysts use SNA to determine if a network is tightly bounded diversified or constricted, to find its density and clustering, and to study how the behavior of netwok members is affected by their positions and connections (Garton, Haythornhwaite & Wellman, 1997; Wellman, 1997; Hanneman, 2001; Scott, 2000; Knoke & Kuklinski, 1982). Network researchers have developed a set of theoretical perspectives of network analysis. Some of these are (Bargotti, 2002): Focus on relationships between actors than the attributes of actors Sense of interdependence: a molecular rather atomistic view Structure affects substantive outcomes Emergent effects. "The aim of social network analysis is to describe why people communicate individually or in groups" (Preece, 2000, pp. 183), while the goals of SNA are (Dekker, 2002): to visualize relationships/communication between people and/or groups using diagrams to study the factors which influence relationships and the correlations between them. to draw out implications of the relational data, including bottlenecks to make recommendations to improve communication and workflow in an organization. Network analysis is concerned about dyadic attributes between pairs of actors (like kinship, roles, and actions), while social science is concerned with monadic attributes of the actor (like age, sex, and income). As SNA is useful in collecting important actor relationship data, HCI techniques can be used to supplement some of its limitations.

Ego-centered analysis – Focuses on the individual as opposed to the whole network, and only a random sample of network population is normally involved (Zaphiris, Zacharia, & Rajasekaran, 2003). The data collected can be analyzed using standard computer packages for statistical analysis like SAS and SPSS (Garton, Haythornthwaite, & Wellman, 1997). Whole network analysis – The whole population of the network is surveyed and this facilitates conceptualization of the complete network (Zaphiris et al., 2003). The data collected can be analyzed using microcomputer programs like UCINET and Krackplot (Garton et al., 1997). There are several characteristics of social networks many of which will be investigated when we use SNA in our unified model. The following are important units of analysis and concepts (Garton et al., 1997; Wellman, 1982; Hanneman, 2001; Zaphiris et al, 2003; Wellman, 1992).

2.3 Human-Computer Interaction Analysis

A working definition of Human-Computer Interaction (HCI) as provided by ACM SIGCHI (2002, pp.8) is: "Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them". The focus is on the interaction between one or more humans and one or more computational machines (ACM SIGCHI, 2002) HCI is a multidisciplinary subject which draws on areas such as computer science, sociology, cognitive psychology and so on (Schneiderman, 1998). The concept of HCI consists of many tools and techniques that are used for information gathering and evaluation. The data collected in conjunction with data collected from other frameworks assists in assessing the online communities of courses and learning more about the users while collecting their feedback. Methods for CMC data analysis include: Questionnaires, Interviews, Personas and Log Analysis.

3. METHODOLOGY

For our case study we used a synthesis of quantitative (SNA) and qualitative (questionnaires) methods and applied them to a Computer Aided Language Learning (CALL) course. Data was collected directly from the discussion board of "Learn Greek Online" (LGO). LGO is a student centered e-Learning course for learning Modern Greek and was built through participatory design and distributed constructionism (Zaphiris & Zacharia, 2001). In an ego-centered approach to SNA, we have carried out analysis on the first 50 actors (in this case the students of the course) of the discussion forum for Lesson 1 in the Greek 101 (Elementary) course of LGO and tabulated these interactions in the form of a network matrix.

To carry out the social network analysis we used an SNA tool called "NetMiner for Windows (http://www.NetMiner.com)" which enabled us to obtain centrality measures for our actors. The "in and out degree centrality" was measured by counting the number of interaction partners per each individual in the form of discussion threads (for example if an individual posts a message to 3 other actors then his/her out-degree centrality is 3, whereas if an individual receives posts from 5 other actors then his/her in-degree is 5).

Due to the complexity of the interactions in the LGO discussion we had to make several assumptions in our analysis: Posts that received 0 replies were excluded from the analysis. This was necessary in order to obtain meaningful visualizations of interaction. Open posts were assumed to be directed to everyone who replied. Replies were directed to all the existing actors of the specific discussion thread unless the reply or post was specifically directed to a particular actor.

In addition to the analysis of the discussion board interactions we also collected subjective data through the form of a survey. More specifically, the students were asked to complete an Attitudes Towards Thinking and Learning Survey (ATTLS). The ATTLS measure the extent to which a person is a 'connected knower' (CK) or a 'separate knower' (SK). People with higher CK scores tend to find learning more enjoyable, and are often more cooperative, congenial and more willing to build on the ideas of others, while those with higher SK scores tend to take a more critical and argumentative stance to learning (Galotti, Clinchy, Ainsworth, Lavin, & Mansfield, 1999).

4. RESULTS

Each node represents one student (to protect the privacy and anonymity of our students their names have been replaced by a student number). The position of a node in the sociogram is representative of the centrality of that actor (the more central the actor the more active). Students S12, S7, S4, S30 (with out-degree scores ranging from 0.571 to 0.265) are at the centre of the sociogram and possess the highest out degree and in-degree scores. This is an indication that these students are also the most active members of this discussion board posting and receiving the largest number of postings. In contrast participants in the outer circle (e.g. S8, S9, S14 etc.) are the least active with the smallest out-degree and in degree scores (all with 0.02 out-degree scores). In addition, a clique analysis was done and it shows that 15 different cliques (the majority of which are overlapping) of at least 3 actors each have been developed in this discussion board. As part of this study we look in more detail at the results from two of our actors. S12, who is the most central actor in our SNA analysis i.e. with the highest our-degree score, and S9, an actor with the smallest out-degree score. It is worth noting that both members joined the discussion board at around the same time. First, through a close look at the clique data (Table 2) we can see that S12 is a member of 10 out of the 15 cliques whereas S9 is not a member of any. An indication of the high interactivity of S12 versus the low interactivity of S9.Actor S12, answered all 20 questions of the ATTLS with a score of at least 3 (on a 1-5 liker scale) whereas S9 had answers ranging from 1 to 5. The overall score of S12 is 86 whereas that of S9 is 60. A clear dichotomy of opinions occurred on 5 of the 20 questions of the ATTLS. S12 answered all 5 with a score of 5 (strongly agree) whereas S9 answered them with a score of 1 (strongly disagree). i.e. S12 strongly agrees that S/He is more likely to try to understand someone else's opinion than to try to evaluate it. S/He often find herself/himself arguing with the authors of books read, trying to logically figure out why they're wrong. S/He finds that he/she can strengthen his/her own position through arguing with someone who disagrees with them. S/He feels that the best way achieve his/her own identity is to interact with a variety of other people. S/He likes playing devil's advocate -arguing the opposite of what someone is saying.

5. CONCLUSIONS

It is apparent from our research that most existing frameworks make either a qualitative or quantitative analysis of CMC, but rarely do we see a mixture of these techniques. Also, some models can only be used on only synchronous or asynchronous communication, but not both. Our opinion is that it is important that a unified framework is developed, for the complete evaluation of all aspects of online communication. As new teaching methods and different learning activities emerge, new types of interaction and evaluation are necessary. The analysis of CMC should take all these updates into consideration, and incorporate them into future CMC analysis models.

This paper has demonstrated the application of Social Network Analysis (SNA) in a computer aided language learning course of Modern Greek. Furthermore, an Attitudes Towards Thinking and Learning Survey (ATTLS) was carried out. Both of the methods used had the same results. More specifically, the results of the SNA showed certain students to be more central in the discussions and these findings were matched by the results of the ATTLS which identified the same individuals as the 'connected knower's'. In the future we plan to extend this study with incorporations of more methods towards a unified framework.

REFERENCES

[1] ACM SIGCHI. (1992). *Curricula for Human-Computer Interaction. New York*, NY: the Association For Computing Machinery.

[2] Archer, W., Garrison, R.D., Anderson, T. & Rourke, L. (2001). "A framework for analysing critical thinking in computer conferences," *European Conference on Computer-Supported Collaborative Learning*, Maastricht, Nerthelands.

[3] Bales, R.F. (1950). "A set of categories for the analysis of small group interaction," *American Sociological Review*, 15, 257-263.

[4] Bales, R.F & Strodbeck, F.L (1951). "Phases in group problem-solving," *Journal of Abnormal and Social Psychology*, 46, 485-495.

[5] Beidernikl, G. and Paier, D. (2003): "Network analysis as a tool for assessing employment policy," In *Proceedings of the Evidence-Based Policies and Indicator Systems Conference 03*. London, July 2003

[6] Borgatti, S. (2000). "What is Social Network Analysis," Retrieved on November 9, 2004 from

[7] http://www.analytictech.com/networks/whatis.htm

[8] December, J. (1997). "Notes on defining of computer-mediated communication," *Computer-Mediated Communication Magazine*, (3):1

[9] Dekker, A.H. (2002). "A Category-Theoretic Approach to Social Network Analysis," *Proceedings of Computing: The Australasian Theory Symposium (CATS)* Melbourne, Australia, 28 Jan to 1 Feb 2002

[10] Fahy, P. J. (2003). "Indicators of support in online interaction," *International Review of Research in Open and Distance Learning*. 4(1).

[11] Ferris, P. (1997) "What is CMC? An Overview of Scholarly Definitions," *Computer-Mediated Communication Magazine*, (4): 1

[12] Galotti, K. M., Clinchy, B. M., Ainsworth, K., Lavin, B., & Mansfield, A. F. (1999). "A New Way of Assessing Ways of Knowing: The Attitudes Towards Thinking and Learning Survey (ATTLS)," *Sex Roles*, 40(9/10), 745-766.

[13] Garton, L., Haythorthwaite, C., & Wellman, B. (1997). "Studying On-line Social Networks," In Jones, S. (Eds.), Doing Internet Research. Thousand Oaks CA: Sage.

[14] Gunawardena, C., Lowe, C., and Anderson, T. (1997). "Analysis of a Global Online Debate and the Development of an Interaction Analysis Model for Examining Social Construction of Knowledge in Computer Conferencing," *Journal of Educational Computing Research*, 17 (4), pp397-431.

[15] Hanneman, R. A. (2001). "Introduction to Social Netwok Methods," Retrieved on November 9, 2004 from

[16] http://faculty.ucr.edu/~hanneman/SOC157/TEXT/TextIndex.html

[17] Henri, F. (1992). "Computer Conferencing and Content Analysis," In A. R. Kaye (Ed), Collaborative learning through computer conferencing: The Najaden Papers, 117-136. Berlin: Springer-Verlag.

[18] Jones, S. (1995). "Computer-Mediated Communication and Community: Introduction," *Computer-Mediated Communication Magazine*, 2 (3): 38.

[19] Knoke, D., & Kuklinski, J.H. (1982). Network Analysis. Sage University Paper Series on Quantitative Applications in Social Sciences. Serious no. 07-001. Beverly Hills and London: Sage Pulications.

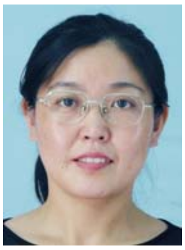
[20] Krebs, V. (2004). "An Introduction to Social Network Analysis," Retrieved November 9, 2004 from

[21] http://www.orgnet.com/sna.html

[22] Korzenny, F. (1978). "A theory of electronic propinquity: Mediated communication in organizations," *Communication Research*, 5, 3-23

[23] Martinez, A., Dimitriades, Y., Rubia, B., Gomez, E., de la Fuente, P. (2003). "Combining qualitative evaluation and social network analysis for the study of classroom social interactions," *Computers & Education* 41 (2003), pp353-368

Changzheng Liu is a vice Professor of Computer Science and Technology College, Harbin University of Science and Technology. He graduated from Harbin Engineering University in 1993; was a postdoctor of Harbin Medical University (2004~2006). He is secretary-general of Hei Longjiang Biomedical Engineering Society. He has published over 20 Journal papers. His research interests are in distributed parallel processing, Visualization in Scientific Computing.



Guiyun Ye is a vice Professor of College of Electrical and Information Engineering, Heilongjiang Institute of Science and Technology, She graduated from Harbin Engineering University in 1986. She has published over 30 Journal papers. Her research interests are in distributed parallel processing, Visualization in Scientific Computing.

The Design and Implementation of Network Teaching Platform Based on .NET

Dongfei Liu, Wei Lu
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei Province 430070, China
Email: dfliu_62@yahoo.com.cn, reed.84@163.com

ABSTRACT

ASP.NET is a kind of technique for Web application develops based on Windows platform. This paper introduced the theory of ASP.NET, and then discussed the design and implementation of a network education platform based on .net.

Keywords: ASP.NET, Network Courseware, Tree-like Catalogue, Dynamic Management

1. INTRODUCTION

Along with the swift and violent development of the internet and the computer software and hardware technique, it brings huge developable space for education system. The method of remote network teaching breaks the limit of traditional study way in time, space, environment and so on, becomes the powerful supplement of traditional teaching way. But the versatility is bad, it’s difficult to renew and maintain, as well as the repeated development may cause resource waste, all of these are the shortcomings of the existing network teaching platform. If we can make a general platform, which is convenient to add and alter courseware, so that we can avoid repeated development, with that we not only save manpower, physical resource resources, but also it’s more convenient to make and update network courseware.

2. ASP.NET TECHNIQUE

The tools that are used to make network courseware include Microsoft FrontPage, Macromedia Dreamweaver, Macromedia Flash and so on. The technologies to develop B/S mode network courseware are ASP, ASP.NET, JSP, PHP and so on.

.NET is a platform of XML Web Service supplied by Microsoft. XML Web Service allows communicate and share data through Internet between application procedures, without considering the operating system and programming language of the application procedure. The core module of .NET includes a group of blocks which are used to set up Internet operating system, the basic structure and tool (including Visual Studio.NET, .NET enterprise server, .NET frame and Windows.NET) which are used to set up and manage the new generational service, can use the software of .NET.

ASP.NET is the extremely essential technology in .NET, is a new generation of general language programming frame which is put out by Microsoft. This is a programming frame that based on the public language running package. ASP.NET also provides a Web application procedure model, which is made up of a group of controllers and a basic structure. It can form Web application procedure on server, its function is formidable. Microsoft provides the develop environment Visual Studio.NET, which is what you see, what you obtain, supports many kinds of languages, and visible. It makes design, develop, compile and running concentrate in together, greatly speed up the development efficiency of ASP.NET procedures. ASP.NET uses structured pages, divides logic codes and display codes through Code Behind technique. C# is a part of Microsoft Visual Studio**.**NET, it's a programming language that simple, advanced, safe and object oriented. It also has powerful general executable engine and abundant classes. It provides great convenience to develop procedure.

ADO**.**NET is a new module to exchange data in **.**NET frame. It integrates lots of classes which are used to deal with database. These classes represent these container objects which have database function (index, view, sort etc.). In process of making this platform, we can make full use of the characteristics and function of ADO.NET in .NET.

3. NETWORK TEACHING PLATFORM SYSTEM STRUCTURE

Let's take *computer network* as an example, design and implement a network teaching platform based on .Net. It is a structured and distributed application system which has three layers in logic: Web performance layer, logic layer and data layer. Namely input query condition in Web performance layer, then receive in logic layer and call for data layer. Data layer inquire in database and return the record which meet to request to the data layer, then logic layer quote the data that receive in data layer. At last Web performance layer quote logic layer, receive the data and display them on Web page. The following picture (Fig.1) shows the three layers structured mode:

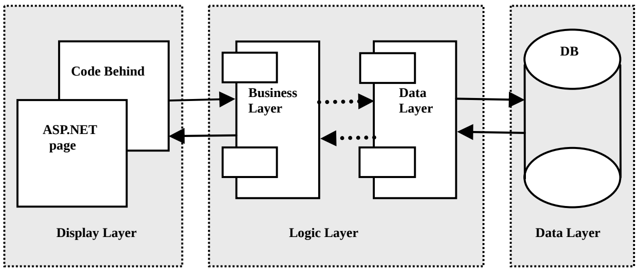


Fig.1. network teaching system structure

3.1 System Function Design

Under the three layers structured mode of .Net, we carry on the development of seven subsystems: announcement system, study system, video classroom, exercise system, communication system between teacher and student, test system and management system (Fig.2). The information of each user land each system controls with the Session object.

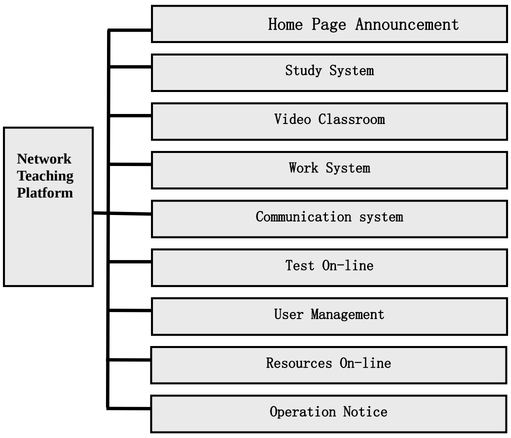


Fig.2. Function module of network teaching platform system

3.2 System Function Introduction

The function of *computer network* network courseware's each module is described as follows:

Home page announcement: The home page announcement is the first page of computer network courseware, we can get the correlated information about the teaching of computer network through the home page announcement, and teachers manage the entrance and information after identity authentication through the announcement.

Study system: It contains all contents of *computer network*, we set the following modules for study conveniently: on-line study －course outline－important concept－study instruction－video classroom－preparation－difficulty's explanation－study key－experiment's instructions, they provide convenience to prepare a lesson, study and review.

Video classroom: Explain the basic concept of computer network through video frequency.

Work system: Student may deliver the work, browse the arranged and corrected work through the work system; Teacher may manage the students' work through the work system, correct the students' paper, and manage the topic of work and so on.

Communication system between teacher and student: Student may leave problems about computer network in correlated block (the questions that can't classify may leave in "other" block); teachers may explain each classified questions regularly.

Test on-line: In on-line test, we have unit self-test, the on-line test, result inquiry module and so on, it enables students to be allowed to consolidate and check knowledge they studied in time; Meanwhile we set make out questions, correct papers, manage results, manage tests module and so on, it facilitates teacher's work enormously.

Register of manager: It's mainly used to manage the information of teachers.

Resources on-line: Links of computer network courseware.

User management: The identity authentication to teachers.

3.3 Database Design

This platform uses relational database, selects SQL Server2000 database administration software. Using Enterprise Manager and Query Analyzer and so on which SQL Server supplies and they are powerful, also extremely convenient to design, develop, arrange and manage database. It's visible to develop, arrange and manage database using SQL Server Enterprise Manager.

Main data information in database includes: examination questions, answer of exercise, the information of manager, user and teacher, information about teaching announcement, information about arranged work, information about each test and so on. The database design is mainly define tables related to teaching management and relationships between tables. The following is a part of tables and segments in database design:

Admin: Id, name, passwd, addtime
Save the information of administration
Admin_teacher: Id, name, passwd, realname, addtime, tel
Save the information of teacher
Notice_information: Id, title, content, addtime
Save the information about teaching announcement
Work_dispose: Id, title, content, addtime
Save the information about arranged work
Exa: Examid, examdate, papered, examteacher, exampwd
Save the information about each test
Studentin: studentid, studentname, studentclass, studentsex
The information of student
Question: Quesid, type, difficulty, chapter, theme, option1, option2, option3, Option4, answer, note
Examination paper topic and answer

4. THE IMPORTANT QUESTIONS AND SOLUTION METHODS

To realize the function of dynamic management is to realize the versatility of study system, test system, communication system and work system. The present method is needed to read out the essential information from database. The test system and work system can conveniently realize dynamic management just through ASP.NET dynamic controller, input questions about new subject and delete questions that we had before. Therefore, it's the key point to realize the dynamic management of study system and communication system for the versatility of the whole platform.

4.1 The Dynamic Management of Study System

The key point of dynamic management function of study system is realize tree-like catalogue, that's to say display the contents of curriculum and dynamic management, the controller IE WebControl which is supplied by .Net and is the union of TreeView and database can do this. Controller mainly contains three objects: TreeView, TreeNode and TreeNode Type, which are separately used to program production tree, treenode and define the type of node.

According to the request of browsing the curriculum content, we divide the content into 3 levels, the realization process is: in database design, we save the title of chapter and section in different tables. Reading data in each table circulated and taking them as root node, two level nodes and so on.

A part code programmed by C# is shown as follows:

```
private void InitTree(TreeNodeCollection Nds,string parentId)
{
        TreeNode tmpNd;
        int myCount=ds.Tables["study_chapter"].Rows.Coun;
        for(int i=0;i<myCount;i++)
        {
                tmpNd=new TreeNode();
                tmpNd.ID=ds.Tables["study_chapter"].Rows[i][1].T
        oString();
                tmpNd.Text=ds.Tables["study_chapter"].Rows[i][1]
        .ToString()+ds.Tables["study_chapter"].Rows[i][2].ToStri
        ng();
                tmpNd.NavigateUrl="content\\"+tmpNd.ID+".htm"
        ;
                tmpNd.Target="main";
                Nds.Add(tmpNd);
                string t="chapterid";
                string table="study_section";
                addchild(tmpNd,t,tmpNd.ID,table,3,2);
        }
}
```

We can realize the content's dynamic management through the interaction with database, and realize the design make use of DataAdapter object and DataSet of ADO.Net. SqlDataAdapter get the information from database which we need, DataAdapter insert the data into DataSet, renew data source after finish the data operation. So administrator can delete or add the title of chapter and section, also can add new curriculum.

The controller DataGrid is used to display data in form. Each line of DataGrid represents a record in data source. We bind three DataGrid controllers separately with _chapter, study_section and study_title, then get the information of chapter title and section title, and display them on pages.

4.2 The Dynamic Management of Communication System

The communication system that we have now is about only one curriculum or the course or chapter is fixed, so if we want to add or delete one section, we must alter the correlated codes.

The versatility of this system is base on different requests of each user who use this system, everyone can add a new curriculum's hard problem Q/A area or a new chapter's Q/A area, also can delete the Q/A area we have now. We design a table chapter to record the name of curriculum and chapter in database, system bind the content of chapter with DropDownList controller, users go into correlated Q/A area through choose the choices in DropDownList after their login. So administrator can operate without alter codes or database.

The related code is shown as follows:

```
protected conn myconn=new conn();
protected void DropDownList1_content()//the method of
binding DropDownList1's data
{
        string selCmd="select * from chapter";
        DropDownList1.DataSource=myconn.GetDS(selCmd,"ch
        apter").Tables["chapter"].DefaultView;
        DropDownList1.DataValueField=myconn.GetDS(selCmd
        ,"chapter").Tables["chapter"].Columns[0].ColumnName;
        DropDownList1.DataTextField=myconn.GetDS(selCmd,
        "chapter").Tables["chapter"].Columns[0].ColumnName;
        DropDownList1.DataBind();//binding finished
}
```

5. CONCLUSIONS

This paper takes computer network as an example, introduces the design and implementation of network teaching platform based on .NET, and emphatically discusses the implementation of study system and the dynamic management of communication. This platform has some versatility, it's convenient to manage and easy to operate. Greatly reduced the development cycle of developing network courseware, and avoided each kind of resources' waste of repeated development, meanwhile gives us some new ideas to develop network courseware with new technical. Of course, it has some disadvantages in application, but we will consummate the functions in the future utilization.

REFERENCES

[1] Fu Lei, *the course of ASP.NET program* [M], Bei Jing: Bei Jing hope electric press, 2002.

[2] Liu Zhenyan, Liu Huimin, Wang Huan, *Basic and enhance of ASP.NET database development* [M], Bei Jing: Tsinghua university press, 2004.

[3] Jing Yu group, *Network version of multimedia courseware manufactures* [M], Bei Jing:China machine press, 2003.

[4] Li Yingwei, Yao Suxia, Jing Li, *ASP.NET database high-level programming(C#)* [M], Bei Jing: Tsinghua university press, 2004.

[5] TOM BARNABY, *Distributed .NET Programming in C#*[M], Berkeley：Apress L．P．，2003.

[6] Zhang Nanping, Wang Wei, Xia Hongxia, “B/S application system development frame based on .NET platform [J],” Wuhan *university of technology newspaper* (information and management section), 2004.26(1): 42~44.

[7] Wang Changda, “The design and implementation of remote B/S system based on WWW [J],” *computer application,* 2001.6.48~49.

Dongfei Liu is an Associate Professor of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He is a tutor of Master and majors in network database.

Wei Lu is a Master of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. She majors in network database.

A Distance Education System Based on P2P

Xiaoling Fu, Bosheng Dong
Department of Information, North China University of Technology
Beijing, 100041, China
Email: xlfu927@163.com

ABSTRACT

Recent implementations of Distance Education System have disadvantages, such as lacking interactivity, inefficient communication, heavy load at servers, etc... This paper proposes an implementation of information interactive Distance Education System based on P2P model. A product named FangDa Instant Communication System is based on this model and it has been applied in high schools and elementary schools, and has proved model's feasible functionalities.

Keywords: Peer-to-Peer, Distance Education System.

1. INTRODUCTION

The Internet provides the learners some new methods to obtain knowledge from the educators: plenty available teaching resources, information search tools, and various communication tools, with which learners can easily do the dependent and independent learning. A distance education system should have significant features, like content, digitization, plentiful resources, convenient platform, and independent learning to catch up the demand from education expansion and network improvement.

Most of the current distance education systems are B/S or C/S models, with which the education process is performed by downloading courseware and test questions ,questing and answering through BBS，handing in and out assignments with e-mail and so on. The problems of these systems include lacking interactivity and individuality, poor instantaneity, heavy loads on servers, and so on. The convenience and plentiful resources on the Internet can not be fully used.

This paper studies the advantages of P2P (peer to peer) network and the features of distance education system, proposes and implements an online distance education system based on p2p model. The system is composed of three parts, resources library, teaching support and teaching management, which not only provides an independent online study environment but also well solves the problems from previous implementations.

2. BRIEF INTRODUCTION OF P2P

P2P is so called peer-to-peer network technique, which is a protocol in network architecture. In this mode users share computer resources and services by virtually direct shifting. IBM gave a definition to P2P as following: the P2P system consists of some conjoint computers and at least has one of the following features:

1) The system relies on marginal devices' active cooperation not central servers, and every member of this system benefits from other members not from server.
2) The members of the system play roles of both server and client.
3) The users of the system can realize the existence of each other, and they compose a virtual or actual community.

A P2P network has two typical topological structures, namely pure P2P network and mixed P2P network. The pure P2P network has no intermediate node, and every peer has the same status and responsibility. This structure is easy to extend but it lacks security and the location of the nodes is complex. The mixed P2P network has no uniform index server while some nodes play the role of server or the role of Super-Peer node to provide special functions. These nodes store the basic status of other nodes and information of content source. The files can be searched and located through these nodes. This structure is easy for both future expansion and location.

P2P is different from the traditional communication techniques. It has a great feature that without the limit of server users who can shift files, download dada, share resources and cooperate with each other to accomplish peer-to-peer accessing. So the system's communication efficiency has been enhanced and the load of server has been reduced. Concurrently, the P2P mode can provide chance to make use of the idle resources of lots of peer nodes so a great amount of processing capacity and memory power can be generated. It also has the obvious advantage of data distribution and server load balancing. Because this technique can be used to reconstruct the distributed system, it is regarded as one of the most potential useful network techniques.

Recently, the application of this technique mainly includes information resources sharing, pervasive computing, collaborative work, instant communication techniques, information retrieval techniques, and the WAN storage system. Based on these features and advantages, P2P technique fits the modern distance education system completely.

3. DISTANCE EDUCATION SYSTEM BASED ON P2P NETWORK

3.1 System Architecture

In this paper an online interactive distance education system based on P2P communication technique is proposed. It is the extension of the real school. This system organizes the users according to the organization form and management mechanism of real school. The users are classified to principal, teacher, student, parents and tutor. The system provides multiple educational methods such as text, files, short message, broadcasting, audio, video and whiteboard to construct a convenient online learning and intercourse environment for learners. The system architecture is given in Fig .1.

The peer-nodes related each other can be formed as a virtual school. The users in the same virtual school can be teachers or students according to status setting, and they can communicate with each other, share resources and do dependant learning without a server standing by. In addition, users from different virtual school can connect with each other though the search mechanism on the system, and the students and their tutors can build up point to point connections accordingly. The model of the virtual school is shown in Fig .2

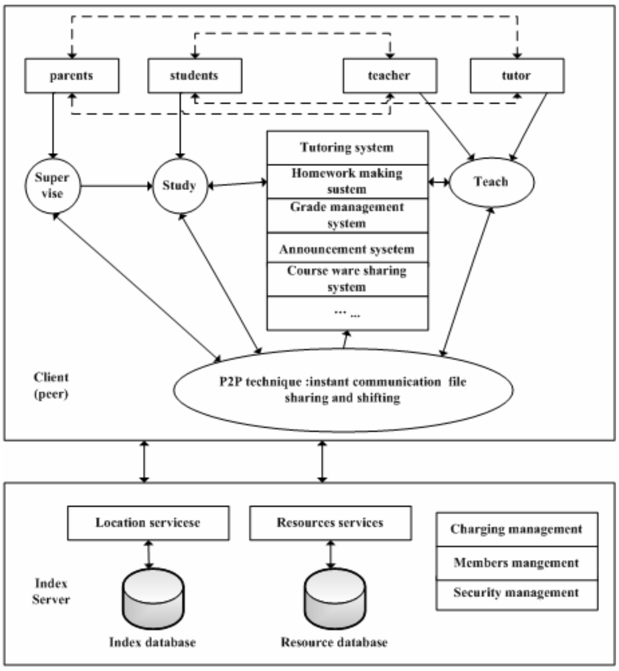


Fig.1. Online Interactive Education System

The peer-nodes related each other can be formed as a virtual school. The users in the same virtual school can be teachers or students according to status setting, and they can communicate with each other, share resources and do dependant learning without a server standing by. In addition, users from different virtual school can connect with each other though the search mechanism on the system, and the students and their tutors can build up point to point connections accordingly. The model of the virtual school is shown in Fig.2.

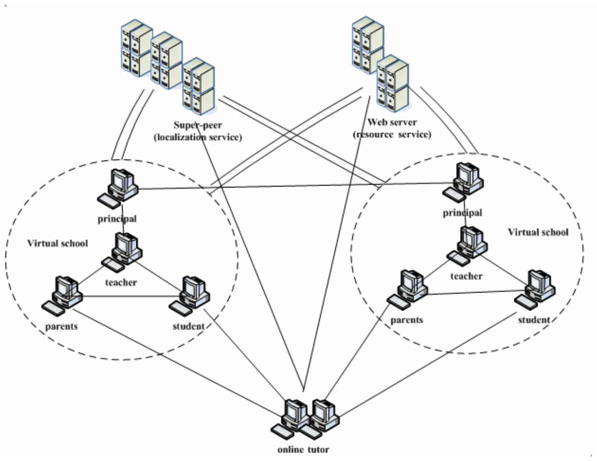


Fig.2. the model of the Virtual School

The advantage of adopting the mixed P2P structure is that the load on the central server can be decreased enormously by directing network traffic to multiple message servers, so the whole performance of the server is enhanced by reducing the response time and avoiding the possible damage from the invalidation on a single node.

3.2 User Intercourse Mode

Once the connections between users are established, users are able to send/receive messages and share files, courseware, teaching plan and test questions with each other directly. For example, the principal can send announcements to the teachers; the teachers can send messages and school report card to students or parents; the teacher can provide his teaching resources such as his courseware to his students for downloading; the teaching activities between teachers and the students can be made through text messages, short messages, audio, video and whiteboard. The students also can search for online tutors on system's database. The user intercourse mode is given in Fig.3.

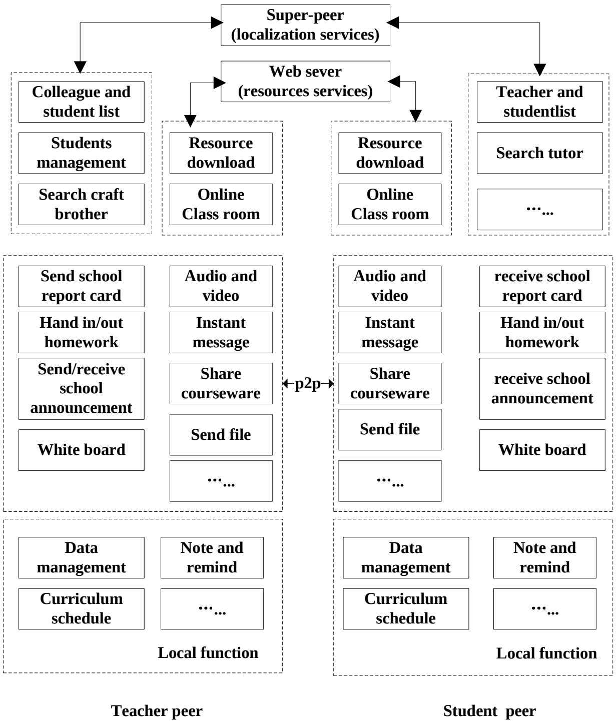


Fig.3. Intercourse Modes between Teacher and Student

4. Implementation of the Distance Education System

4.1 Client

Developed in MS Visual c++, the client runs on Windows operating system. The users can transfer messages, files, audio and video and communicate interactively with whiteboard. The client also has some local management function including event handling, personnel and data management, and etc. After login, the user queries the server about friend's address, with which the user can connect to his friend. When the P2P connection is established between two entities, the clients could communicate with each other directly with UDP. If P2P connection fails, the messages have to be transmitted through an index servers and the files be transmitted through a file servers.

4.2 Server

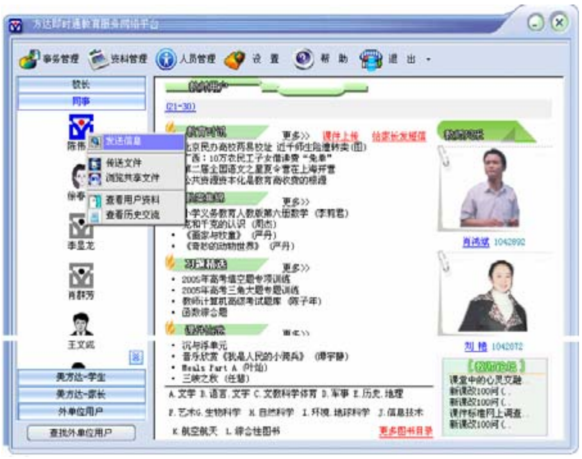
Index server (super-peer): runs on Linux operating system. It is in charge of completing the user login and identification, information transmitting, system messages broadcasting and communicating with other servers. The clients connect to it with UDP or TCP.

Resource server: is developed with web technique and it stores the shared resource. A user can search the resource through a WEB browser embedded in the client.

4.3 Communication Protocol

The data shifted among the nodes is in XML format. The sender transfers the message into XML format and the receiver parses the message into information easily after receiving it. It is convenient to communicate with other communication systems, like Jabber, if the XML format data is used.

“FangDa Instant Communication System” is an educational service platform product based on P2P Distance Education System. Its development is a subtask of the state key project, “Study of Education Technology Development in the Information Progress”, in tenth five-year plan. The working results of the system have passed assessment held by the experts from China Central Education Institute, so the product system is put into application in high schools and elementary schools, where total number of registered user has achieved 50 thousand. FangDa Instant Communication System’s teacher interface shown in Fig.4.



北京民办高校两易校址 近千师生险遭转卖(图)
广西：10万农民工子女借读费“免单”
第二届全国语文之星夏令营在上海开营
公共资源资本化是教育高收费的根源

- 小学义务教育人教版第六册数学（李莉君）
- 克和千克的认识（周杰）
- 《画家与牧童》（严丹）
- 《奇妙的动物世界》（严丹）

- 2005年高考填空题专项训练
- 2005年高考三角大题专题训练
- 教师计算机高级考试题库（陈子年）
- 函数综合题

- 沉与浮单元
- 音乐欣赏《我是人民的小骑兵》（谭宇静）
- Meals Part A（叶灿）
- 三峡之秋（任慧）

A.文学 B.语言.文字 C.文教科学体育 D.军事 E.历史.地理
F.艺术G.生物科学 H.自然科学 I.环境.地球科学 J.信息技术
K.航空航天 L.综合性图书

课堂中的心灵交融...
新课改100问(...
新课改100问(...
课件标准网上调查...
新课改100问(...

Fig.4. FangDa Instant Communication System’s Teacher Interface

5. CONCLUSIONS

The Distance Education System implemented based on the P2P technique is reported in this paper, and the product system provides an interactive environment for different education entities. With this mode, the members in the system can share resources, study dependently, and complete some simple education management tasks in P2P mode. The product system has nice instantaneity and interactivity. In addition, it reduces the possible congestions on network traffic and load on the central server caused by massy and frequent data exchange. The product system has put in application for more than one year, and has validated its function feasibility.

REFERENCES

[1] The Internet Lab., *China Network Educational Industry Report*, 2004.11

[2] Le Guangxue, Li Renhua, Zhao Changhua, Ding Lei, “Research and Application of P2P Technology,” *Computer Engineering and Application*, 2004.36:163-167

[3] Qin Lv, Pei Cao, Edith Cohen, Kai Li, Scott Shenker. “Search and Replication in Unstructured Peer-to-Peer Networks” [C]. *Proceedings of the 16th ACM International Conference on Supercomputing*. New York: ACM Press, 2002, 258-259

[4] Bryan Ford, Pyda Srisuresh, Dan Kegel. *Peer-to-Peer (P2P) communication across middle boxes* [EB/OL].

[5] http://midcom-P2P.sourceforge.net/draft-ford-midcom-P2P-01.txt. 2003

The Design and Implementation of Long-distance Experimenting System Based on Virtual Instrument and Computer Network

Wenlian Li, Yang Li, Dejun Yang
The education science and technology department of Xiangfan University, Xiangfan, Hubei 441053, China
Email: lwlian@163.com

ABSTRACT

In this paper, we discuss a method applied for designing and implementing the long-distance experimenting system based on virtual instrument and network. Making use of virtual instrument and computer network, we can construct a long-distance experimenting system that can supply a series of experiment. This system can improve the efficiency and level of experiment teaching through controlling experiment instrument across far distance achieving open long-distance experiment teaching based on Internet.

Keywords: Virtual Instrument, Long-distance Experiment, Computer Network

1. INTRODUCTION

As the rapid growth of amount of college students, many universities are facing difficult situation of experiment teaching, lacking in instruments、grounds and staff for experimenting. But the enhancement of teaching quality must be supported by the experiment teaching in a large degree. We can improve the ability for operation of students, empolder their creativity, enhance their integrated quality. Therefore we must reform our experiment teaching. One of the feasible approaches to solve this problem is building long-distance experimenting system based on virtual instrument and network.

As the rapid development of visual instrument and computer network, it is feasible to construct virtual laboratory through computer network which has advantages of simulation experiment and virtual experiment. And it can improve the facticity of experiment operation, sharing the experiment software and hardware. The long-distance experimenting system designed in this paper, combines virtual instrument and computer network. So the students can control the experiment equipment across far distance、adjust experiment parameter online、obtain real experiment results through computer network. This system reduce the limit to experiment teaching posed by time and space greatly, achieving open long-distance experiment teaching based on Internet, improving the efficiency and level of experiment teaching.

2. SYSTEM ARCHITECTURE

2.1 General Architecture

Using the five-combined-to-one virtual instrument(DSO500) with parallel circuit, designing a suit of circuits controlled by computer(USB interface circuit and electronic switch) which will be connected with a main computer(server), installing corresponding software that virtual instruments need and controlling interface circuit, connecting 6 suits of experiment instrument with electronic switch circuit, we can build a long-distance experiment system based on virtual instrument which can serve multiple(6 in the figure) experiments that can be chosen by students. This system is illustrated in Fig 1 below.

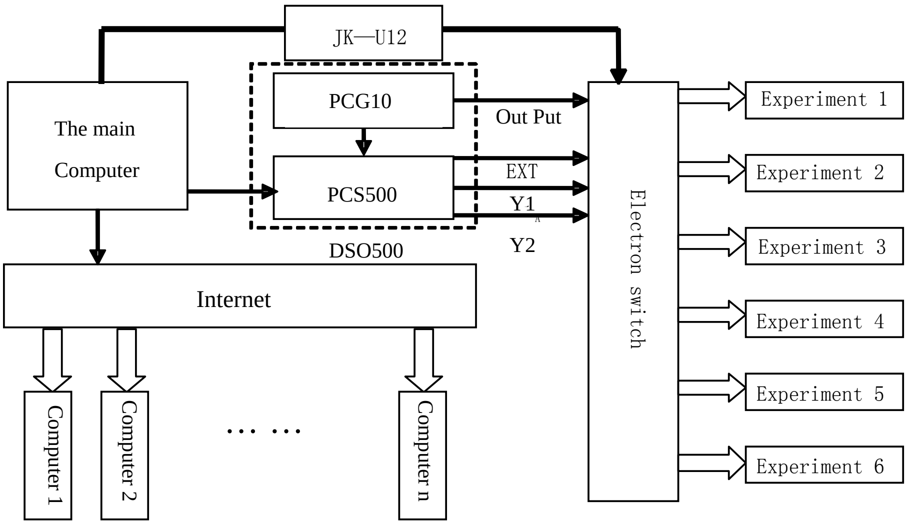


Fig.1. The framework of long-distance experiment system

2.2 The Virtual Instrument

This system uses a set of virtual instrument which can be connected with 6 experiment circuit respectively through electronic switch. The virtual instrument can use many various instruments which can meet the needs of the experiment. We choose five-combined-to-one virtual instrument (DSO500) with parallel interface which is consisted of input module PCS500 and output module PCG10. It contains multifunctional signal generator, oscilloscope, transient recorder, spectrum analyzer and bode plot, which is suitable for long-distance experiment.

2.3 The Controlling Circuit

The controlling circuit consisting of USB interface circuit, electronic switch circuit, signals input/output interface circuit, divide the input and output interface of DSO500 into 6 groups which are connected with 6 suits of experiment equipments respectively. The users can control their switch through electronic switch circuit, making one suit of the experiment instrument get connected with DSO500 to take measurements and to complete an experiment. The USB interface circuit is used for the control module called JK-U12 for data collection and with multiple functions. It connects the server with the control circuit, controlling the electronic switch. The U12 module has 20 channels for digital input and output. Its input and output can be configured discretionarily. We choose 6 channels of them as the output control, which can drive electronic switch directly.

2.4 The Design of Software

The Windows XP system provides many simple remote control functions, of which remote desktop is one kind. The users can control ling-distance computer using it through Internet. Then they can access all applications、files and other resources. The remote desktop is mainly composed of client and server. We set the main computer as the server, the computers of students as client. So we can control the long-distance server through remote desktop. The virtual instrument software Pc-Lab 2000 installed in the server is used to control and operate DSO500 for measurement. The very main purpose of remote desktop control is to take remote measurement using Pc-Lab 2000 which control DSO500 across far distance.

The network controlling software installed in the server is used for users to choose the experiments. The network controlling software, programmed with Labview designed 6 controlling buttons. These buttons correspond with the 6 output channels of JK-U12, controlling the electronic switch circuit, achieving the switch of input/output circuit with experiment equipment, fulfilling connection between virtual instruments and a certain suit of equipment. The front panel and rear panel of this network controlling software are illustrated in Fig 2 and Fig 3 respectively.

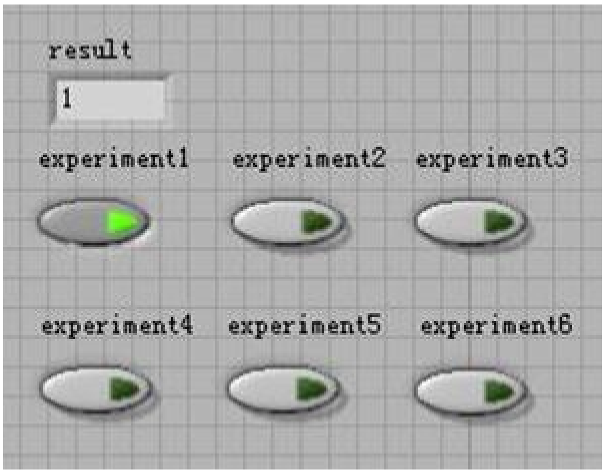


Fig.2. The front panel of the network controlling software

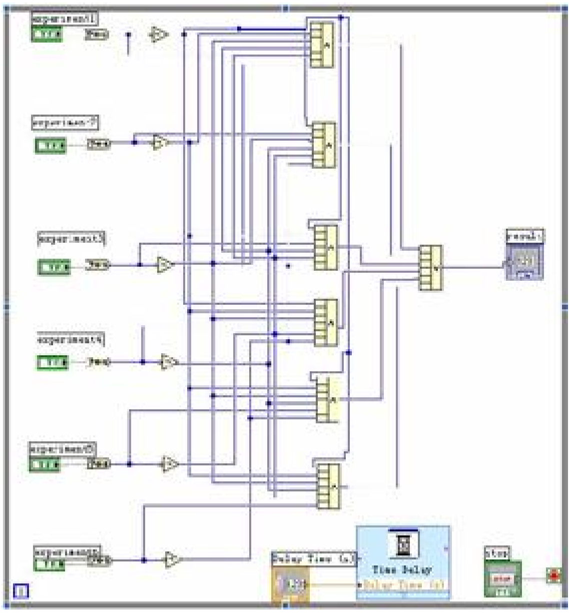


Fig.3. The rear panel of the network controlling software

3. THE EXPERIMENT OPERATING METHOD

The remote experiment system based on virtual instrument and network is suitable for experiment teaching across far distance of physics、circuit and electronic technology which need electronic measures. Here the virtual instruments take the place of traditional electronic measures, which can be controlled through computer and network easily.

The students can preview the experiment using network experiment teaching software installed in local computer. Then they can login the server by remote desktop, running the network controlling software. They can choose corresponding buttons on the panel, select the input/output circuits needed by their experiments, starting Pc-Lab 2000 installed in the server. Then they can take measurement with DSO500. The wave and data measured can be saved in server or client, for later data processing or experiment report.

When the students use the network controlling software, they can only click one control button, performing one experiment. They can choose another after completing.

In this system, the structure of every circuit and every measuring point is fixed. The circuits and parameters can not be changed. This is adverse to training students. If we change the computer controlling circuit a little, controlling the on/off of the same suit of experiment equipments using the 6 controlling channels to change the circuits and parameters, then we will get better experiment effect. The ability for experiment operation of students will get better trained.

4. CONCLUSIONS

Combining virtual instruments and computer network, achieving long-distance experiment teaching, we reduce the limit of time and space. And we can share experiment equipment and data across far distance, improving the efficiency of experiment teaching and use of experiment equipment greatly. It poses a new approach for building open laboratory.

REFERENCES

[1]. Wang hongru, “virtual instrument —new age of instrument” [J], *Foreign Electronic Measuring Technology* ,1998(1):482;

[2]. Li Wenlian, “demonstrating experiments with multimedia using virtual instrument,” *Research of Electrical Education*, 2004(3):55-57.

Wenlian Li, who was born in February, 1956, is a full professor in the education science and technology department of Xiangfan University. He graduated from Hubei University with specialty of physics. His interests are in the research and teaching of electronic technology and application of computer. He is also interested in the application of virtual instrument

Personality Mining System in Web Based Education by Using Improved Association Rules Mining Method

Mingmin Gong [1], Qi Luo [2]
[1] Department of Information Engineering, Wuhan University of Science and Technology Zhongnan Branch
Wuhan, 430223, P.R.China
[2] School of Electrical Engineering,Wuhan Institute of technology,Wuhan 430070, China
ccnu_luo2008@yahoo.com.cn

ABSTRACT

To meet the personalized needs of Web based education, an improved association mining rules was proposed in the paper. First, data cube from database was established. Then, frequent item-set that satisfies the minimum support on data cube was mined out. Furthermore, association rules of frequent item-set were generated. Finally, redundant association rules through the relative method in statistics were wiped off. The algorithm had two advantages, the first was that the execution time was short while searching for the frequent item-set; the second was that the precision of the rules was high. The algorithm was also used in personality mining system based on Web based education model. The result manifested that the algorithm was effective.

Keywords: E-Learning, Mining, Algorithm, Association Rules, Personality

1. INTRODUCATION

Nowadays, the importance of Web based education has transferred from how to solve the limit of space-time problem in traditional teaching to build up the personalized learning environment, and offer a kind of personalized knowledge service based on theories such as modern pedagogy, psychology, etc [1]. The learners are in different age level, sex, and social role, their culture, education background, attention and interest are also exist a great difference. Giving corresponding learning content and tactics to realize teaching learners according to their needs is very difficult [2]. Its basic reason lies in being difficult by obtaining the relations between the learner's personality characteristics and learning behavior patterns accurately, automatically. In this way, it is necessary to mine out the association rules between personality characteristics and learning behavior patterns. The subsequent learners’ personality characteristics are deduced from their learning behaviors by using the rules above. Basing on it, personality-learning model is set up and interesting groups are formed. Thus, personalized learning and cooperative learning are realized.

At present, many scholars have carried on a great deal of researches on association algorithms, such as such as Apriori algorithm or improved Apriori algorithm FP-tree [3]. The execution time of their algorithms was long while searching for the frequent item-set and the rule’s interest degree was low. Even a large number of redundant rules are also included [4]. Besides, the complexity of space and time are high. According to this, an improved association rule mining algorithm is proposed in the paper. It has improved the traditional algorithm, and introduced data cube, relative method in statistics and adaptive adjusting mechanism in intelligence control [5]. The algorithm had two advantages, the first was that the execution time was short while searching for the frequent item-set; the second was that the precision of the rules was high. The algorithm was also used in personality mining system based on Web based education model. The result manifested that the algorithm was effective.

2. AN IMPROVED ASSOCIATION RULES

The steps of an association rules mining based on data cube as follows:
Step1: establishing data cube from database.
Step2: mining frequent item-set which satisfies the minimum support on data cube.
Step3: generating association rules of frequent item-set.
Step4: redundant association rules are wiped off through the relative method.

2.1 Establishing Data Cube

Supposed in personality model database, we observe the data by three dimensional angles. These three dimensions separately are learning behavior patterns dimension, learner's personality characteristics dimension, and time dimension. Then data cube is obtained through OLAP operation from data base [6]. Fig.1 is data cube by 3-D.

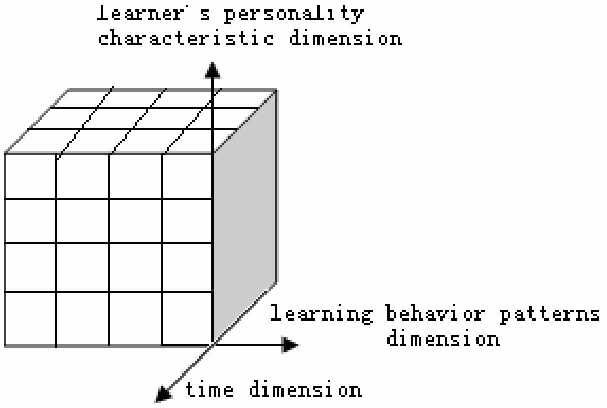

learner' s personality characteristic dimension
learning behavior patterns dimension
time dimension

Fig.1. Data cube by 3-D

Each dimension relates to one table that is called dimension table. It carries on further description to the dimension. For example, personality characteristics dimension contains name, sex, age, income, occupation, hobby, character and so on.
Above the specific angle (dimension) also has many different descriptions in detail degree. This kind of description is called dimension level. One dimension generally has the many levels. For example, the description of time dimension may describe different level from year, quarter, month, date and so on. Year, quarter, month, and date are the dimension level of time.

2.2. Mining Frequent Item-Set on Data Cube

The paper has improved Apriori algorithm. The algorithm searching for the frequent item-set on data cube is called

Personal_ Cube_Apriori algorithm. L_k Represents set of frequent k_ item-set, C_k Represents set of candidate k_ item-set

The idea of adaptive adjusting of the Personal_ Cube_Apriori algorithm as follow:

Firstly, the algorithm searches for frequent k_ item-set for each dimension. If some dimensions don't have frequent k_ item-set, it shows that the dimension level is excessively low and we should drill through above and improve the dimension level. If in some dimensions, all frequent k_ item-set is frequent k_ item-set, it shows that the dimension level is excessively high and we should drill through under and lower dimension level.

Fig .2 is shown as the flow of Personal_ Cube_Apriori

Input: n dimension data cube $(d_1, d_2 \cdots d_n, count)$, minimum support min_sup, k=1

Output: set of frequent item-set $L = L \bigcup L_k$

$L_k = gen\_frequent(C_k)$ Represents set of candidate k_ item -set generating frequent k_ item-set

$C_K = gen\_candidate(L_{k-1})$ Represents set of frequent (k-1)_ item-set joining, pruning and generating k_ item-set

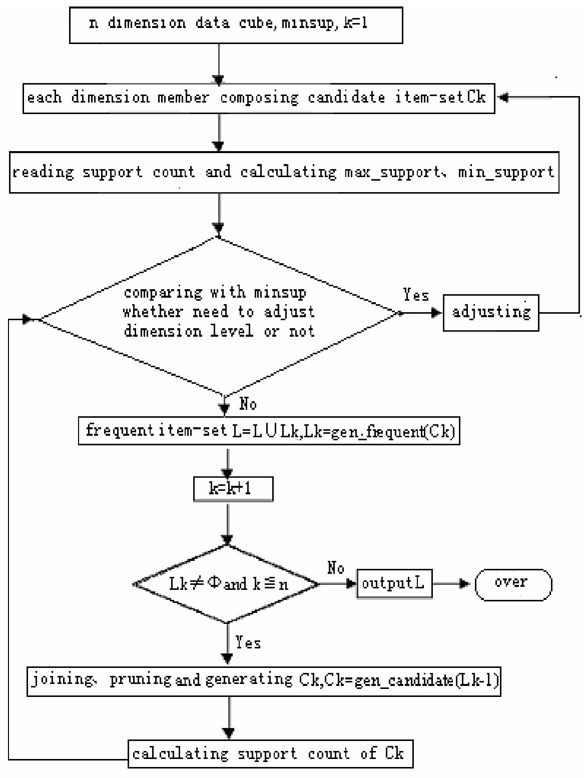

n dimension data cube, minsup, k=1
each dimension member composing candidate item-set Ck
reading support count and calculating max_support、min_support
comparing with minsup whether need to adjust dimension level or not
Yes
adjusting
No
frequent item-set L=L∪Lk,Lk=gen_frequent(Ck)
k=k+1
Lk≠Φ and k≦n
No
output L
over
Yes
joining、pruning and generating Ck,Ck=gen_candidate(Lk-1)
calculating support count of Ck

Fig.2. The flow of Personal_ Cube_Apriori

2.3. Generating Irredundant Association Rules

After mining out frequent item-set, the process of generating association rules is composed of two steps:

Step1: regarding to each frequent item-set *l*, all non- spatial subsets are generated.

Step2: regarding to each non- spatial subset of frequent item-set *l*, if

Then the rule *s*=> (*l*-*s*) is generated.

$$\frac{\sup port\_count(l)}{\sup port\_count(s)} \geq \min\_conf \tag{1}$$

min_conf represent the minimum confidence thresholds, support count(*l*) represents the number of transaction containing item-set *l*, support count(*s*) represents the number of transaction containing item-set *s*.

Lower interest degree rules are obtained in this way, so the redundant rules must be wiped off. Thus, the paper introduces the relative analysis method in statistics to wipe off the redundant rules. After mining out one association rule, we calculate the correlation among frequent item-set. Concept of correlation is formula 2

$$CORR_r = \frac{O(r)}{E(r)} \tag{2}$$

$CORR_r$ More approaches to 1, the independence among set of r item-set is better. For example, the association rule $X => Y$, *X*, *Y* are the item-set. If its correlation is worse, it must be the false strong rule.

Relative analysis method is applied in k dimension data cube. If item or dimension Y is $Y_1, Y_2, \cdots, Y_r$ *r* values, X is $X_1, X_2, \cdots X_s$ *s* values, n_{ij} represents the number of X *is* Y_i , and X *is* X_j .

$$n_i = \sum_{j=1}^{s} n_{ij} \quad i = 1, 2, \cdots, r \tag{3}$$

$$n_j = \sum_{i=1}^{r} n_{ij} \quad j = 1, 2, \cdots, s \tag{4}$$

$$n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \quad m_{ij} = \frac{n_i n_j}{n} \tag{5}$$

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{6}$$

χ^2 Is correlation degree among items, if all items are independent, χ^2 is 0. Supposed the critical value is assigned. If χ^2 is bigger than the critical value, *X* and *Y* are statistical correlation. Otherwise, they are not statistical correlation. Namely, they are independent.

3. APPLICATION

Based on the above research, we combine with the cooperation item of personalized knowledge service system in network education .The author constructs a personality mining system website. The system is also applied in network institute of ** university. The results manifest the system support personalized Web based education better the experiment is that we carry on the investigation and construct the database for personality attributes of 360 learners, 300 learners are selected to regard as the data source and other 60 learners data are selected to regard as the examination data.

The steps of experiment are shown as Fig.4

Step1: Data pretreatment and clean. First, not complete attributes values in database are filled in though rough theory. Then, according to the distribution of each attribute extreme value, sparser values is removed by minimum support min_sup =30%. Date converses to Boolean attributes according to the actual value scope of each attribute .thus, 100 attributes are obtained.

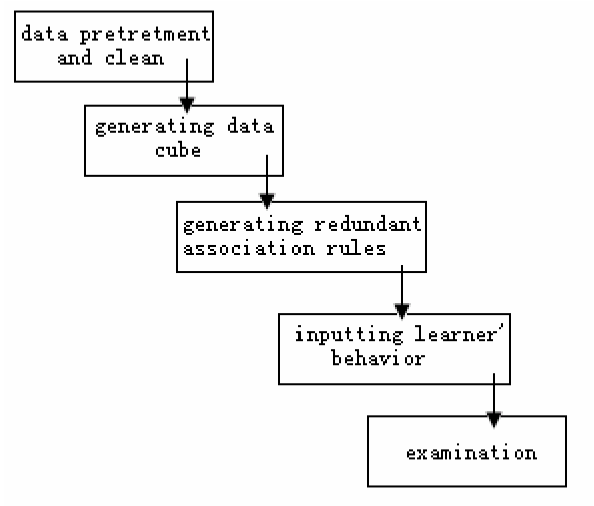

data pretretment and clean
generating data cube
generating redundant association rules
inputting learner' behavior
examination

Fig.4. The steps of experiment

Step2: Generating data cube. Data cube is generated through OLAP operation. These are three dimensions such as learning behavior patterns dimension, learner's personality characteristics dimension, and time dimension.

Step3: Generating association rules. The frequent item-set is searched out though Personality_ Cube_Apriori. Then, χ^2 is calculated and some false strong rules are removed. As a result of the application environment limited, the partial learning behavior patterns dimension and learner's personality characteristics dimension are carried on analysis though improved association rule algorithm. Learning behavior patterns selects two aspects such as learning courseware and issue information on BBS. 9766 multi-dimensional association rules are obtained. 6794 association rules are obtained though $\chi^2 < 166.7$. There are 10 former and back integrity rules of them.

Step4: Inputting the following learners' behavior matches above rules, so learning characteristics is obtained. On the basis, best learning content and learning strategy are provided to them. So, personalized learning and cooperative learning are realized

Step5: 60 Learners' data is regarded as the examination data. Compared with above data, the result manifested that the system is effective.

4. CONCLUSIONS

The article has realized Personal_Cube_Apriori algorithm with VC++ on the base of Apriori and has compared this algorithm with Apriori algorithm. The test data set is obtained by above application example. Fig. 5 is experiment result，Y axis is executing time t (ms) of generating frequent item-set and X axis is the distribution of minimum support degree, ranging from 50% to 5%.

From Fig5, when minimum support degree is lower, the executing time is longer. When minimum support degree is higher, the execution time of two algorithms is approached. In a word, when they have same minimum support degree, the execution time of Personal_Cube_Apriori is shorter than Apriori.

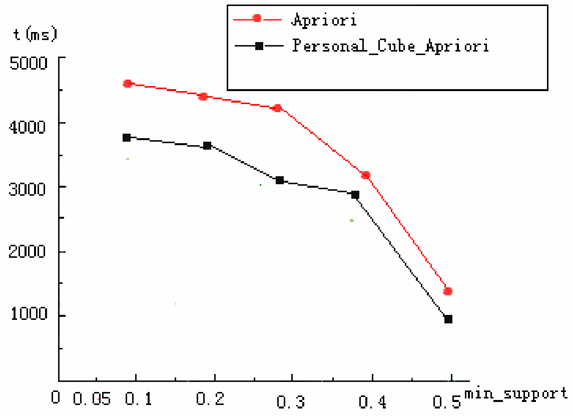


Fig.5. Comparison of two algorithms mining performance

We also carry on the experiment of rules precision, Apriori algorithm , Improved Apriori algorithm are compared with Personal_Cube_Apriori .40 association rules form each algorithm is selected as the comparative experiment data. Precision comparison of three algorithms is Fig.6

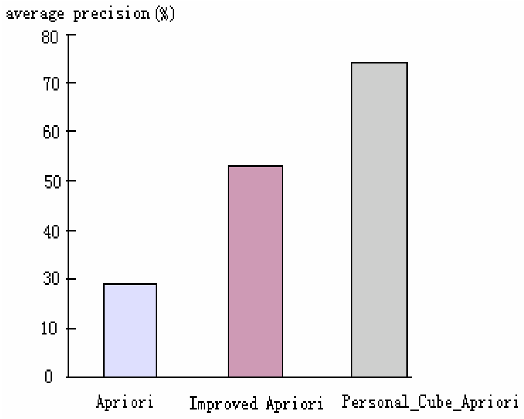


Fig.6. Precision comparison of three algorithms

In summary, an improved association algorithm is proposed in the paper. The results manifest that the algorithm is effective through above algorithm performance research. Meanwhile, the model of personality mining system is proposed. The results manifest that the system mines out the association rules between personality characteristics and learning behavior patterns. The subsequent learners' personality characteristics can deduced accurately, automatically from their learning behavior by using the rules above. I wish that this article's work could give some references to certain people.

REFERENCES

[1] Yanwen Wu, Qi, Luo "Research on Personalized Knowledge Service System in Community Web based education". Edutainment 2006 Proceedings. Lecture Notes in Computer Science, Volume 3942, 2006.4, pp.115-152.

[2] Jun Liu, Renhou Li and Qinhua Zheng, "Study on the Personality Mining Method for Learners in Network

Learning”. Journal of Xian Jiaotong University, 2004, 38(6), pp.575-576.

[3] Han J and Kamber M, “Data Mining Concept and Techniques”. Academic Prints, 2001, pp.100-103.

[4] Lina Lu and yaping Chen, “Research on algorithm Apriori of mining association rules”. MINI-MICRO SYSTEM, vol 21, 2001.9, pp.942-944.

[5] Wei Yuan, "data mining center of Statistics department of china renmin university", Statistics and information forum, 2002(1), pp.5-9.

[6] Shizhuan Yan and Zhanhuai Li, "Commercial Decision System Based on Data Warehouse and OLAP", Microelectronics & Computer, 2006(2), pp.66-67.

MingMin Gong is a University lecturer, School Information Technology, Wuhan University of Science and Technology, Zhongnan Branch. She graduated from Wuhan University in 1997; she has published one book, the calculator introduction, and continuous plait VB. Develop three applied software. Her researches interests are in calculate way research, grid computing, network security and e-commence.

Graphics, Image, Vision and Voice Processing

A Study of Semantic Retrieval System Based on Geo-ontology with Spatio-temporal Characteristic*

Jia Song[1,2], Yunqiang Zhu[1], Juanle Wang[1]
[1] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
Beijing, 100101, P. R. China
[2] Graduate University of the Chinese Academy of Sciences
Beijing, 100049, P. R. China
Email: songj@lreis.ac.cn

ABSTRACT

The development of GIS (geographic information system) and the production of a great many geographic data everyday demand a new semantic retrieval system for geographic information. This paper analyzes the requirement of geospatial and temporal extendibility and reasoning for geographic information retrieval. Then architecture of semantic retrieval based on domain ontologies is illustrated, and every component in the architecture is introduced one by one. The goal of this paper is to explore the application of ontologies in the field of geographical information retrieval. How to create geospatial ontology and temporal ontology are respectively discussed in this paper. And a case of geospatial ontology of China is described. Finally, the future work and a prototype application are mentioned in conclusion, in which we can see that the semantic retrieval system will play a key role in future.

Keywords: Semantic Retrieval, Geo-Ontology, Geospatial Ontology, Temporal Ontology, Topological Relationship, OWL

1. INTRODUCTION

With the development of Earth Observation and GIS (geographic information system) techniques, a large amount of geographic information is currently being stored and delivered over the internet. So how to effectively retrieve the geographic information that people wish to find presents many challenges for search techniques. Different with generic information retrieval, the terms about location and time of geographic information are often employed in geographic information retrieval. Because geographic information itself is characteristic of spatio-temporal conceptualization and Region conceptualization is the essential viewpoint of geography [1]. But the effectiveness of geographic information retrieval is not far satisfied for lacking geospatial and temporal extendibility and reasoning. For example, when people enter “land use in Yangtze River valley” via search engine for finding the information related to land use which refer to Yangtze River valley, only the information containing “land use Yangtze River valley” terms can be found, which means that the information such as land use in Hubei province or Yangtze Delta both of which locate in Yangtze River valley cannot be found. So search engines are still primitive, and they are not capable of extending terms according to some reasoning mechanisms. Especially spatio-temporal reasoning is quite useful in geosciences domain. Therefore lots of scholars on both computers and geosciences propose the notion and approaches of semantic retrieval based on ontologies. The definition of ontologies is presented in [2-4]. The need for semantic geospatial web is described in [5]. This paper is concerned with semantic retrieval for geographic information based on geo-ontology and mainly focuses on the semantics related to geospatial and temporal Characteristics. Architecture of semantic retrieval based on domain ontologies is firstly introduced in section 2. Then how to create the geo-ontology which is applied in semantic retrieval is discussed. And geospatial and temporal ontologies of China are described respectively in section 3. And a summary along with future directions of this research is given in the end.

2. ARCHITECTURE OF SEMANTIC RETRIEVAL BASED ON DOMAIN ONTOLOGIES

The essential to the semantic search engine is that the domain ontology library and the components which are associated with it and are integrated into the generic search engine. The architecture of semantic retrieval system is shown in Fig.1. The whole system consists of two subsystems for two respective processes associated with retrieval. One is for the indexing process; the other is for the searching process.

2.1The Subsystem for Indexing Process

The subsystem for indexing process mainly consists of resources access interface; language analyzers; an index creator; semantic annotator.

Language analyzers are shared in searching process. The resources access interface is a set of uniform interfaces to access various resources storages (Database, Xml document, Html document, Word document, etc.). The system can support to index diverse data resources if corresponding implementations are provided.

Language analysis is the process of converting field text into its most fundamental indexed representation, terms. These terms are used to determine what documents match a query during searches. An analyzer is an encapsulation of the analysis process and it tokenizes text by performing any number of operations on it, which could include extracting words, discarding punctuation, removing accents from characters, lowercasing, removing common words, reducing words to a root form (stemming), or changing words into the basic form (lemmatization). During the analysis process, different languages have different sets of stop words and unique stemming algorithms. Take Chinese language for example, it generally use ideograms rather than an alphabet to represent words. These pictorial words are not separated by whitespace. So a word splitter, which splits words on the basis of a Chinese

* China Scientific Data Sharing Program: Earth System Science Data Sharing Network (2005DKA32300)

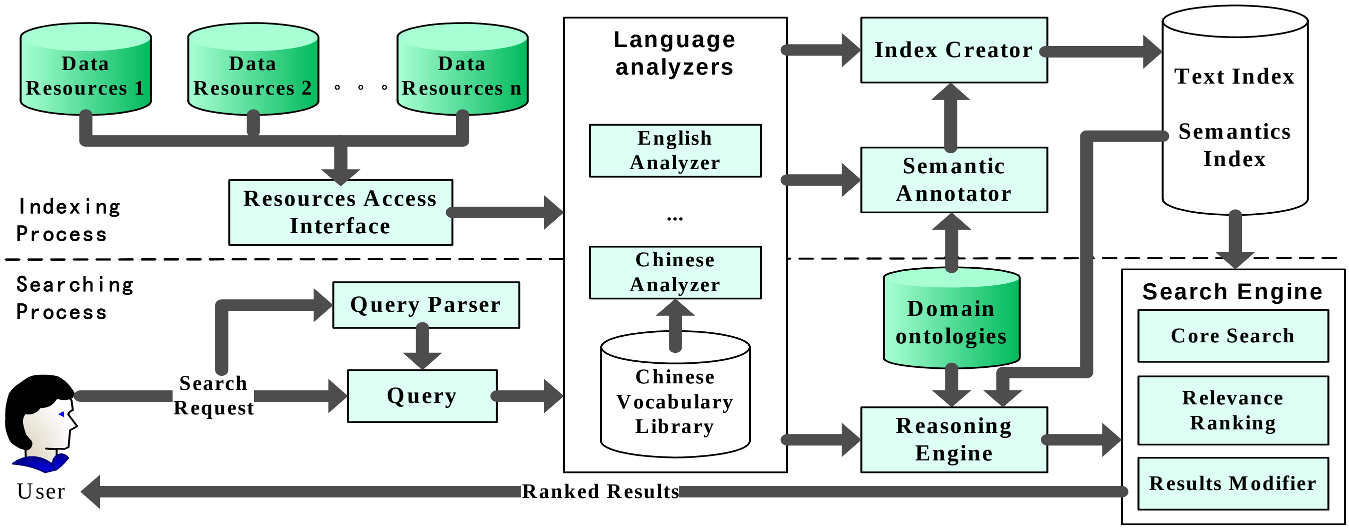


Fig.1. Semantic Retrieval System Architecture

vocabulary library, is indispensable.

After the input has been analyzed, it's turn to be added to the index by the index creator. Data structure of the index is well known as an inverted index, which makes efficient use of disk space while allowing quick keyword lookups. What makes this structure inverted is that it uses tokens extracted from input documents as lookup keys instead of treating documents as the central entities.

The semantic annotator plays a key role in indexing process for semantic retrieval supported. The semantic annotator fetches semantic identifier (ontology class or ontology instance) from domain ontologies on the basis of understanding the input analyzed so that resource entities are mapped into conceptual space or semantic network. The output of the semantic annotator is also indexed via the index creator, namely semantics indices.

2.2 The Subsystem for Searching Process

The subsystem for searching process mainly consists of query; query parser; reasoning engine; search engine.

The query is the medium between the search request from users and the components associated with search. Query encapsulates the computer-readable parameters or structure of query request. For example, a field name and a text-value pair can be a simplest query.

Although the query can be powerful, it is not reasonable that all queries should be explicitly written in code. The query parser using a human-readable textual query representation is implemented based on a rich query language, i.e., the query parser is designed for human-entered text, not for program-generated text. For example, an expression entered by the user could be as readable as this: "+pubdate: [20060101 TO 20061231] Java AND (Jakarta OR Apache)". This query searches for all books about Java that also include Jakarta or Apache in their contents and were published in 2006.

The reasoning engine, which is based on the domain ontologies and semantics indices, is a key semantic processing component. We argue that semantic retrieval has two levels: basic semantic retrieval and expanded semantic retrieval (conceptual retrieval). Basic semantic retrieval is based on semantic annotation and only relies on semantics indices without reasoning. The precise is improved via the semantic annotation. But the enhancement of recall needs ontology reasoning based on ontologies and semantics indices.

The search engine mainly consists of core search, relevance ranking and results modifier. The core search, which searches index created in the indexing process, is the key of search engine. The output of the core search requires to be ordered by relevance via relevance ranking. Besides relevance based on term matching, semantic relevance algorithm is dispensable for semantic retrieval. Finally the search results may require highlighting the text based on a query for more friendly to user, which is achieved by results modifier.

3. A GEOGRAPHIC SPATIO-TEMPORAL ONTOLOGY FOR SEMANTIC RETRIEVAL

The architecture mentioned above is a generic framework. The semantic retrieval applied in kinds of domains can be easily achieved if corresponding domain ontologies and reasoning algorithms provided on the basis of this architecture. In this section we will discuss how to construct geo-ontologies for semantic retrieval and mainly aims at location or region query expansion and temporal query expansion. Some of representative regional divisions on the geography of China are served as expert knowledge for geo-ontologies [6-8].

3.1 Geospatial Ontology of China

Geospatial ontology of China is on the basis of spatial relationship theory in GIS, which includes topological relationship, directional relationship, proximal relationship. Topological relationship plays the key role in Geospatial ontology of China. There are several models associated with topological relationship. They are RCC (Region Connection Calculus) [9], 4IM (4-intersection Method) and 9IM (9-intersection Method) [10-11], CBM (Calculus-based Method) [12] and so on.

One of ontology languages, OWL (Web Ontology Language), should be introduced briefly before we discuss a framework of geospatial ontology. OWL is based on DLs (Description

Logics), which describes the world in terms of “properties” or “constraint” that specific “individuals” have to satisfy. OWL captures classes and associated properties. The properties in OWL have two disjoint property types, object properties and datatype properties. Object properties are used to link classes together, and Datatype properties link classes to XML Schema datatypes [13]. OWL has been passed as a W3C recommendation for defining and instantiating web ontologies [14]. Besides OWL, another specification, GML (Geographic Markup Language), has to be referred to here. GML is established by the Open GIS Consortium (OGC) [15] as a standard language for encoding and sharing geographic information and provides a variety of kinds of objects for describing geography including features, coordinate reference systems, geometry, topology, time and so on [16]. On the basis of GML, a simple framework of classes and properties in the owl geo-ontologies is shown in Fig.2. And by the aid of three types of object properties: transitive, symmetric and inverse, the topological relationships in Fig.2 are stored as object properties. Some examples of topological relationships are listed as follows.

```
<owl:ObjectProperty rdf:ID = " Topological -Relationship">
        <rdfs:domain rdf:resource= "#Geometry"/>
        <rdfs:range rdf:resource = "# Geometry "/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID = "PartOf">
        <rdfs:subPropertyOf
        rdf:resource="#Topological-Relationship "/>
        <rdf:type rdf:resource= "&owl;TransitiveProperty"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID = "Contains">
        <owl:inverseOf rdf:resource = "#PartOf"/>
</owl:ObjectProperty>
```

Based on classes and properties in geo-ontology framework, the organization framework of regional divisions of China is shown in Fig.3, and one of representative regional divisions, administrative division of China, is shown in Fig.4.

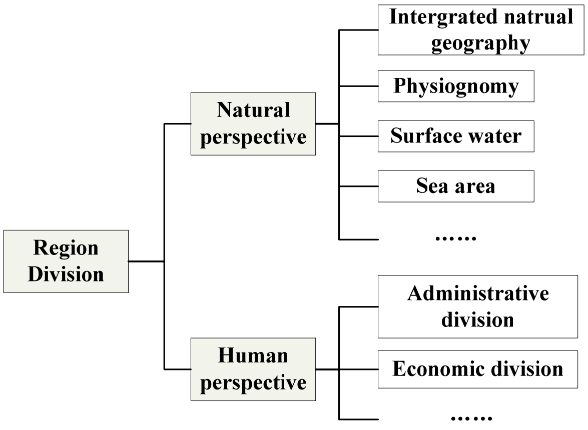


Fig.3. Framework of Regional Divisions of China

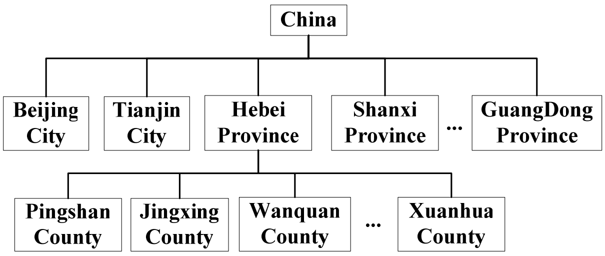
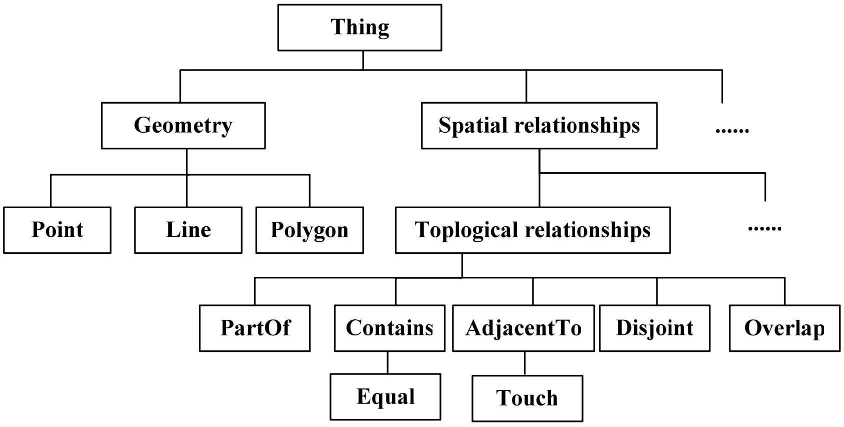


Fig.4. Administrative Division of China



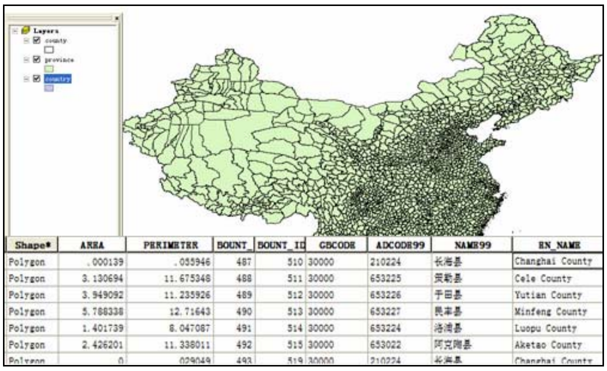
Property	Type	Domain	Range	Subproperty of
Spatial Relationship	Standard	Geometry	Geometry	
PartOf	Transitive	Geometry	Geometry	Spatial Relationship
Contains	Inverse(PartOf), transitive	Geometry	Geometry	Spatial Relationship
AdjacentTo	Symmetric	Geometry	Geometry	Spatial Relationship
Touch	Symmetric	Geometry	Geometry	AdjacentTo
Overlap	Symmetric	Geometry	Geometry	Spatial Relationship
Equals	Symmetric	Geometry	Geometry	Contains

Fig.2. Framework of Classes in Geo-Ontology

Here one of the key problems, geospatial granularity, should be considered. We define the level of county as the minimum geospatial unit. Other regional divisions of China except administrative division require being in contact with county in administrative division via spatial relationship directly or indirectly in that many of geographic data, especially statistical data of society and economy, are organized by administrative unit and more data can be found via reasoning from non-administrative regions to administrative regions.

Besides geospatial granularity, it is necessary to explore an approach to automatically constructing geospatial ontology of China. The approach which we are discussing is based on ArcGIS, which provides a scalable framework for implementing a GIS for a single user or for many users on desktops, in servers, over the Web, and in the field [17]. Take example for administrative unit of China in geospatial ontology. The basic requirements are: three layers of vector data: national boundaries of China, province boundaries of China, county boundaries of China; all of the layers with accordant geographic references and scales; available attribute data such as name, short name and code of regions. As shown in Fig.5, all levels of administrative units are represented by polygons. A containment relationship exists between provinces and counties, where one province is made up of possible many counties. Counties are normally adjacent to other counties and similarly provinces are adjacent to other provinces. The geospatial ontology for administrative unit of China can be automatically constructed using ArcObjects APIs (Application Programming Interface) of ArcGIS if only the basic requirements above are satisfied. The test code of querying counties via a certain province is listed as follows.

```
public Function QueryCounties(prov As IGeometry) As
IFeatureCursor
 ' //get the focus map layer.
   Dim layer_ As IFeatureLayer
   Set layer_ = MxDoc.FocusMap.Layer(0)
 ' //construct a SpatialFilter object for calling spatial query
   Dim query_ As ISpatialFilter
   Set query_ = New SpatialFilter
' //define spatial relationship is 'Contains'
   Set query_.Geometry = prov
   query_.SpatialRel = esriSpatialRelContains
' //execute the query
   Set QueryCounties = layer_.Search(query_, False)
End Function
```



Shape*	AREA	PERIMETER	BOUNT_	BOUNT_ID	GBCODE	ADCODE99	NAME99	EN_NAME
Polygon	.000139	.053946	487	510	30000	210224	长海县	Changhai County
Polygon	3.130694	11.675348	488	511	30000	653225	策勒县	Cele County
Polygon	3.949092	11.235926	489	512	30000	653226	于田县	Yutian County
Polygon	5.788338	12.71643	490	513	30000	653227	民丰县	Minfeng County
Polygon	1.401739	6.047087	491	514	30000	653224	洛浦县	Luopu County
Polygon	2.426201	11.338011	492	515	30000	653022	阿克陶县	Aketao County
Polygon	0	.029049	493	519	30000	210224	长海县	Changhai County

Fig.5. Administrative Unit Layers

3.2 Temporal Ontology

The basis of time ontology is the definition of the time domain. The extant definitions associated with how facts can interact with time are described in [18-20] and the definitions associated with temporal granularity are described in [21-23]. The notion that time-varying data may be modeled as an event or a state is applied in geographic information [24]. An event occurs at a point of time, i.e., an event has no duration. A state has duration, e.g., a storm occurred from 5:06 PM to 5:42 PM.

Temporal granularity should be considered in event model. We think the minute can be enough finely granular to be minimum unit of temporal ontology in geographic information. So the basic expression of event ontology is "year-month-day hour: minute". Others such as decade and century need to be mapped to the expression of event. For example, 1990s means 1990-01-01 00:00 to 1999-12-31 23:59. When people enter "China land use in 1990s", the result such as "China land use in 1995" can be found. Moreover Geological Time Scales [25] shown in Fig.6 is required in geosciences, and Chinese Chronology [26] shown in Fig.7 is also indispensable to geographic information related to China. The temporal ontology describing them should be developed.

EON	ERA	PERIOD	EPOCH	DATES
Phanerozoic	Cenozoic	Quaternary	Holocene	0-2
			Pleistocene	
		Tertiary	Pliocene	2-5
			Miocene	5-24
			Oligocene	24-37
			Eocene	37-58
			Paleocene	58-66
	Mesozoic	Cretaceous		66-144
		Jurassic		144-208
		Triassic		208-245
	Paleozoic	Permian		245-286
		Carboniferous		286-320
				320-360
		Devonian		360-408
		Silurian		408-438
		Ordovician		438-505
		Cambrian		505-570
Proterozoic	Also known as Precambrian			570-2,500
Archean				2,500-3,800
Hadean				3,800-4,600

Fig.6. Geological Time Scales (Dates are in millions of years)

DYNASTY		DATES
Xia Dynasty		21 cent. - 16 cent. B.C.
Shang Dynasty		16 cent. B.C.- 1066 B.C.
Zhou Dynasty	Western Zhou	1066 B.C.- 771 B.C.
	Eastern Zhou	770 B.C.- 221 B.C.
Qin Dynasty		221 B.C.- 206 B.C.
Han Dynasty	Western Han	206 B.C. - 8
	Eastern Han	25 - 220
Three Kingdoms	Wei	220 - 265
	Shu Han	221 - 263
	Wu	222 - 280
……		……
Yuan Dynasty		1279 - 1368
Ming Dynasty		1368 - 1644
Qing Dynasty		1644 - 1911
Republic of China		1912 - 1949
The people's Republic of China		1949 -

Fig.7. Chinese Chronology

4. CONCLUSIONS

The paper describes the one of applications of ontologies: semantic retrieval. It is obvious that domain ontologies play a major role in semantic retrieval systems. Some issues in the construction of geospatial ontology and temporal ontology have been discussed. The research shows that the effectiveness of the geographic information search will be greatly improved based on corresponding ontologies and reasoning engine. Our future work is concerned with various issues related to the actual implementation of geo-ontology with development of tools for automatically constructing geo-ontology. The development of approaches to reasoning over the geo-ontology will also be involved. A prototype application of the semantic retrieval system is "Earth System Science Data Sharing Network", which is one of the key projects in China Scientific Data Sharing Program, it can be seen that the semantic search system will play important role for geographic information on discovering and retrieving in R&D (Research and Development) Infrastructure and Facility Development in China.

REFERENCES

[1] Lin Chao, Some Remarks on the Nature of Geography, Scientia Geographica Sinica, 1981, 1(2), pp.97~104

[2] Gruber, T., A translation approach to portable ontology specifications. Knowledge Acquisition 1993, 5 (2),pp. 199~220.

[3] GRUBER T R, Toward principles for the design of ontologies used for knowledge sharing, International Journal of Human and Computer Studies, 1995, 43 (5), pp.907 - 928.

[4] GUARINO N, Formal Ontology in Information Systems, Amsterdam: IOS Press, 1998, pp.3 ~15.

[5] M.J. Egenhofer, Toward the semantic geospatial web, Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, ACM Press, New York, 2002, pp. 1~7.

[6] State Quality Supervision Bureau, Codes for the administrative divisions of the people's republic of China, GB/T 2260-2002, 2002.

[7] Natural geography of China, Science Press, 1985, pp.193~195.

[8] Wei Sun, The theoretical study and positive analysis on content system of regional planning in China, Doctoral dissertation, CAS, 2005

[9] Randell D, Cui Z, et al., A spatial logic based on regions and connection, Nebel B, Rich C, Swartout W, eds. Proc. of the Knowledge Representation and Reasoning, 1992, pp.165~176

[10] Egenhofer,M., Franzousa R., Point-set topological spatial relations, International Journal of Geographical Information System, 1991, 5(2), pp.161~174.

[11] Egenhofer, M.J., Mark D. M., Modelling conceptual neighborhoods of topological line-region relation, Int. J. GIS, 9(5), pp.555~565.

[12] Clementini E., Di Felice, et al., A small set of formal topological relationships suitable for end-user interaction, D.Abel and B.C. Ooi(eds.) Advances in spatial databases:Proc. 3rd Intl, Symosium on spatial Databases(SSD'93), 1993, pp.277~295.

[13] Alia I. Abdelmoty, Philip D. Smart, et al, A critical evaluation of ontology languages for geographic information retrieval on the Internet, Journal of Visual Languages and Computing & Computing, 16 (2005) , pp.331~358.

[14] F. van Harmelen, J. Hendler, et al, Owl web ontology language reference, http://www.w3.org/TR/owl-ref/.

[15] Open Geospatial Consortium, http://www.opengeospatial .org/.

[16] Simon Cox, Paul Daisey, et al, OpenGIS Geography Markup Language (GML) Implementation Specification, 2004.2.

[17] ESRI, ArcGIS Desktop Help, 2004.

[18] R.T. Snodgrass, I. Ahn, Temporal databases, IEEE Comput. 19 (9) (1986) , pp.35–42.

[19] R.T. Snodgrass, M.H. Bohlen, et al, Adding Transaction Time to SQL/Temporal, ISOANSI SQL/Temporal Change Proposal, ANSI X3H2-96-152r ISO/IEC JTC1/SC21/WG3 DBL MCI-143,1996.

[20] C. Bettini, C.E. Dyreson, et al, A Glossary of Time Granularity Concepts, Temporal Databases: Research and Practice, Springer, Berlin, 1998, pp. 406–413.

[21] C. Bettini, S. Jajodia, S.X. Wang, Time Granularities in Databases, Data Mining and Temporal Reasoning, Springer, Berlin, 2000.

[22] C.E. Dyreson, W.S. Evans, H. Lin, R.T. Snodgrass, Efficiently supporting temporal granularities, IEEE Trans. Knowl. Data Eng. 12 (4) (2000), pp.568–587.

[23] C.S. Jensen, R.T. Snodgrass, Semantics of Time-Varying Attributes and their use for Temporal Database Design, Time Center Technical Report, Tucson, Arizona TR-1,January 29, 1997.

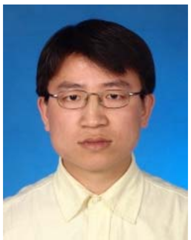
[24] C.S. Jensen, R.T. Snodgrass, Semantics of time-varying information, Inform. Syst. 21 (4) (1996), pp.311–352.

[25] Geological Time Scales, http://www.geosociety.org/science/timescale/timescl.pdf

[26] Chinese chronology, http://www.chinapage.com/dyna1.html

Jia Song is a doctoral candidate in Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences and majors in GIS (Geographic Information System). Recently, he has been working at the software development of Earth System Science Data Sharing Network, which is one of the projects in China Scientific Data Sharing Program. His research interests are spatial data mining and knowledge discovery, GIS software development.

Yunqiang Zhu is an Assistant Researcher of Institute of Geographic Sciences and Natural Resources Research, CAS (Chinese Academy of Sciences). He received his Ph.D degree from CAS in 2006. His main research interests include GIS development, 3S application in land and water resources management, and Geo-data sharing which involves multi-data integration, data sharing policy and standards, network platform development, etc.



Juanle Wang, born in 1976 in Luoyang, Henan Province, P. R. China. He is a Research Assistant in Institute of Geographic Sciences and Natural Resources Research, CAS. He received his Ph. D from CAS in 2005. His main research interests are geoscience data sharing and related standards and methods, he did a lot of research works into the metadata implementation in data sharing, he worked in the National Scientific Data Sharing Project Plan, he conducted postal doctoral research in 2005~2007, his postal doctoral research area is coastal area and near sea region data resources integration and sharing.

Two-Dimensional Photonic Band Gap Calculation Using the FDTD Method in the PC Cluster System *

Min Li[1], Jian Zhou[2], Yi Li[3], Anchun Xiao[4]
[1] School of Computer Science and Technology
[2] State Key Laboratory of Advanced Technology for Materials Synthesis and Processing
Wuhan University of Technology, Wuhan, 430070, P. R.China
[3] Computer Teaching Department of Basic Courses Institute, Zhejiang Wanli University, Ningbo 315101, china
[4] School of Science, Hubei University of Technology, Wuhan 430068, China
E-mail:lim7213@whut.edu.cn

ABSTRACT

This paper focuses on the band gap phenomena of two-dimensional photonic crystals using the finite difference time domain (FDTD) method in a PC cluster system. In this paper, the PC cluster system and parallel FDTD algorithm are proposed to simulate numerical calculation of photonic band gap(PBG). In the experiment, the two-dimensional PBG which consists of a square array of circular Alumina rob is numerically simulated by the MPI program based on the grid computing. The result of experiment shows that the parallel FDTD algorithm is correct and reasonable, and this algorithm can be developed for three-dimensional PBG numerical simulation.

Keywords: The Finite Difference Time Domain (FDTD), Photonic Band Gap (PBG), MPI, Grid Computing, PC Cluster

1. INTRODUCTION

A photonic crystal is a periodic arrangement of dielectric materials that an electromagnetic (EM) wave of a specific frequency range is forbidden to pass through. Because of the band gap phenomenon, a photonic crystal becomes waveguide which can guide light with great efficiency either along a straight path even around a sharp corner, minimizing guiding loss due to material absorption. There are many engineering applications, such as in semiconductor lasers, optical fibers, single mode waveguides, etc. The photonic band gap (PBG) is the major feature of a photonic crystal, which is important to identify and design periodic structure [1]. It was first proposed in the area of optics, a lot of works both theoretical and experimental have been performed.

The numerical simulation is an effective solution to analyze the material property and design the geometric shape. With regard to the simulation of band gaps in photonic structures, the finite difference time domain (FDTD) method was applied to analyze and interpret the experimental data. The FDTD approach was introduced by Yee in 1966 [2], which had been proven to be a powerful technique in solving the Maxwell's equations. The technique has been extensively applied to solve a broad variety of electromagnetic problems, because of its great flexibility in modeling arbitrarily shaped structures with complex media and broad frequency band computation without any matrix operations [3]. It requires great demand of computational time and computer storage, which can be only achieved by traditional supercomputer.

However, the traditional supercomputer no longer satisfies the demands of large scientific computing in that the hardware has met the bottleneck of design and the price of a supercomputer is too expensive to be afforded in general research. So grid computing is becoming the most popular high- efficiency computation machine in scientific and technologic field, which is commonly regarded as a means for creating inexpensive "super-computer" that can deal with heterogeneous and distributed computing. It can integrate the heterogeneous and distributed computing resources into an organized computing pool [4]. Nowadays, because of much better price per performance and the availability of gigabits network and communication protocols, building a cluster of PCs to act as a single large-scale parallel computer becomes more and more popular. MPI (Message Passing Interface) is the most popular communication interface used in PC clusters as well as other cluster-type parallel environment. A new version of MPI based on P2P technology is developed to exploit the maximum performance in heterogeneous environment for distributed computing.

In this paper, a parallel FDTD method is developed in a PC cluster system to handle and accelerate the calculation speed of band gap phenomenon of the photonic crystal, which can be formulated to study the band gap phenomenon in different geometric shapes, material properties, and design patterns. MPI was adopted to implement communication among different nodes in the parallel environment. Firstly, the paper proposes a general discussion on the MPI architecture and the parallel environment. The paper also presents a flexible parallel algorithm for the FDTD implementation and a simulation experiment that indicates the performance of these parallel FDTD implementations, and also discusses the result of the application experiments and our ongoing work.

2. PARALLEL COMPUTING IN THE PC CLUSTER SYSTEM

2.1 The Parallel Architecture

Most parallel architectures are grouped into three categories [5]: SIMD (single-instruction multiple-data), MIMD (multiple-instruction multiple-data), and SPMD (single-program multiple-data).The first architecture, which achieved some success, was SIMD. All processors in this architecture execute the same instruction and are synchronized in time, under the direction of the sequencer. Although programming is relatively easy, it is very difficult to make a control network of sufficient capacity to provide signal synchronization at a frequency higher than 10 MHz. Moreover, if each processor does not execute exactly the same instruction as the others, the performance degrades quickly.

The second architecture (MIMD) is very general. Each processor performs its own instructions. In this case, synchronization is controlled by the developer, who uses the messages passing through the interconnection network to

* This research is supported by the National Science Foundation of Hubei Province, P. R. China (No. 2007ABA156, 2007ABA136)

achieve synchronization.

The SPMD architecture (see Fig.1) is close to the MIMD concept. There is only one program for all the processors, but each one operates independently of the others. Again, synchronization needs to be insured by the developer. This architecture fits well with the FDTD algorithm. The three-dimensional space is divided into several subspaces, and each one is associated with a processor. All the processors execute exactly the same FDTD program, but each one operates on its own subspace.

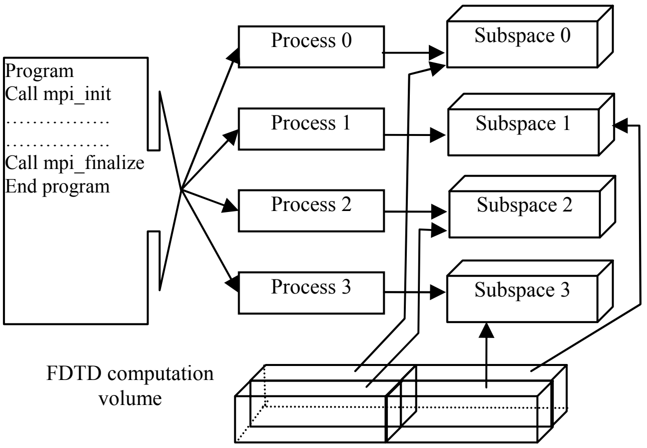


Fig.1. A SPMD architecture

With these architectures, there are two memory types: shared-memory and distributed memory.

(1) The shared-memory approach. Shared-memory architectures allow processor communication through variables stored in a shared address space. All grogram units access data from a central memory and, at any moment, the data can be accessed and eventually changed by any processor node. Every interaction among processor nodes is performed through the shared memory.

(2) The distributed memory approach. With distributed memory, each processor has its own memory. Access to the memory of the other processors is not direct: memory access between processors is performed via the interconnection network.

2.2 MPI Structure

MPI is a library specification to support message passing on whatever platform is suited for SPMD to parallel and distributed computing. MPICH is a portable implementation of the MPI specification. MPICH derives its portability from its interfaces and layered architecture [6]. At the top is the MPI interface as defined by the MPI standards. Directly beneath this interface is the MPICH layer, which implements the MPI interface. Much of the code in an MPI implementation is independent of the networking device or process management system. This code, which includes error checking and various manipulations of the opaque objects, is implemented directly at the MPICH layer. All other functionality is passed off to lower layers by means of the Abstract Device Interface (ADI).

The ADI is a simpler interface than MPI proper and focuses on moving data between the MPI layer and the network subsystem [6]. Those interested in implementing MPI for a particular platform need only define the routines in the ADI in order to obtain a full implementation. Existing implementations of this device interface for various MPPs(Massively Parallel Processors), SMPs(Symmetric Multiprocessors), and networks provide complete MPI functionality in a wide variety of environments. As a result a user can run MPI program across multiple computers at different sites using the same command that would be used on a parallel computer.

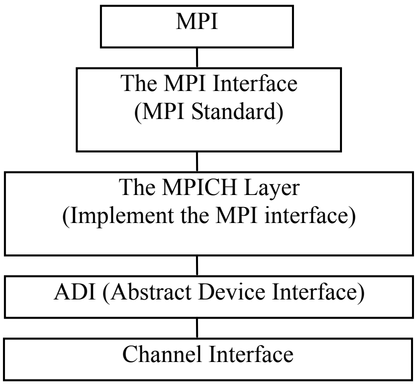


Fig.2. MPI structure

2.3 MPI and Communication

MPI treats processes (not processors), which are grouped inside a communicator. The communicator defines the communication context. A process has a local memory and an execution unit. Each processor may run more than one process. The executable program must be installed and compiled on each platform before executing. When program is executed, the user indicates the number of processes to be used from the operating-system command line. During execution, some MPI procedures provide useful information to the program, such as the number of processes used and their ID numbers. At the beginning of the program, the first MPI instructions concern the providing of those data returned by the procedure "MPI_Comm_size" and "MPI_Comm_rank:"

```
call   MPI_Init (ieer)
call   MPI_Comm_size(MPI_Comm_world,
       numberofprocess, ieer)
call   MPI_Comm_rank(MP1_Comm_world, processID,
       ieer)
```

where

ieer : error code return
MPI_Comm_world : MPI default communicator
numberofprocess : returned number of processes
processID : returned process ID number

The fundamental communication is the transmission of a message by one process and the reception of it by another process. Required arguments for the message passing are the buffer address, the data type, the number of data elements, the ID number of either the receiver or the emitter, a tag to identify the message being passed, and the communicator name to describe the exchange context.

In the MPI library, several types of communication are defined, which used as optimizing the communications, securing them, synchronizing the transmitter and the receiver. The communication type used also depends on the parallel machine used, because all possible instructions cannot be anticipated, and performance may therefore need to be optimized. The transmission can be blocking (the emitter process stops until the addressee process has received the complete message), or non-blocking (blocking only during the temporary copy to the buffer). The reception can be non-blocking in certain situations. In this case, a great deal of attention is necessary in order to avoid data conflicts. More complex communications exist, such as collective exchanges, which are useful for matrix computations.

3. THE PARALLEL IMPLEMENT OF THE FDTD ALGORITHM

3.1 FDTD Algorithm

In the FDTD method, Yee partitioned space into elementary cells (see Fig.3). To rectangular coordinates, the coordinate axes can be discretized with steps ⊿*x*, ⊿*y*, ⊿*z*. Time step is usually indicated with ⊿*t*. The generic space point can be identified with notation *(i ⊿ x, j ⊿ y, k ⊿ z)*, or , more synthetically, *(i , j, k)*. In such a cell, each H-field is surrounded by four E-field components. This interconnected E and H-field lattice, in conjunction with a central finite-difference discretization of Maxwell's equations.

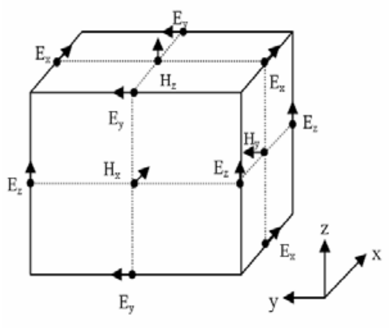


Fig.3. Yee's cell

The temporal and 3D spatial discretizations adopted in the FDTD algorithm are implemented at their best by using a leapfrog integration scheme to solve Maxwell's equations. In the leapfrog scheme, at time step *t=n+1/2*, in each mesh point *(i,j,k)*, each $H^{n+1/2}$ components is computed as a function of the previous value $H^{n+1/2}$ in the same point, plus a function of *E* components at time *t=n* in the mesh points belonging to the neighborhood of *(i,j,k)*. In a similar way, each *E* component is computed, at time step *t=n+1*, in each mesh point *(i,j,k)*, as a function of the same component at previous time step E^n plus a function of the *H* components at time *t=n+1/2* in the mesh points belonging to the neighborhood of *(i,j,k)*. The exact expressions for the computation are similar to the following, which is used to compute the E_x component:

When a numerical approach based on a spatial mesh and a time discretization is used, a stability analysis is needed. The strategy was classical Von Neumann approach, called Courant condition [8]. In accordance with such condition, the time step △t must be chosen so that:

$$\Delta t \leq \frac{1}{c\sqrt{\frac{1}{(\Delta x)^2}+\frac{1}{(\Delta y)^2}+\frac{1}{(\Delta Z)^2}}}$$

Because the actual medium is considered to be infinite, when numerical computing is conducted, it must be truncated to a finite medium domain, which brings up the unwanted reflected wave if the truncated boundaries are not properly treated. To eliminate or reduce the reflection of waves from such an artificial boundary, we use a highly effective perfectly matched layer (PML) ABCs [7]. It is proved that this algorithm is highly stable. Besides PML ABCs, the periodic boundary condition (PBC) is needed. Because of the periodicity of photonic crystals, Bloch's theorem is applied to fulfill the boundary condition. According to the Bloch's theorem, the displacement and stress components in the periodic structure can be expressed in a periodic function as follows in Fig.4.

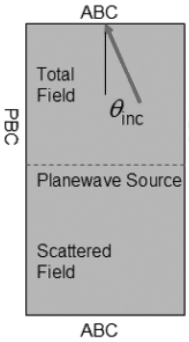


Fig.4. Partition of computational area

3.2 Parallelizing FDTD Algorithm

The first step of the parallel algorithm is the equal distribution of the two-dimensional problem space among the processes. With the MPI library, a Cartesian topology is defined on the two-dimensional volume, in order to facilitate the distribution and the communication between neighboring processes. According to a two-dimensional topology applied on the x-y plane, the computation volume is divided into nine subspaces (0 to 8). No division is done along the z axis. This topology is suitable for thin structures, such as microstrip circuits and antennas.

Each process can be addressed by its Cartesian coordinates or by an ID number. In the two-dimensional topology, each process has a subspace localized by an ID number and its coordinates. The two-dimensional Cartesian topology is created by the following procedure:

Call MPI_Cart_create (MPI_Comm_world, Ndim, dim (1 : 2) ,period (1: 2) ,Organ, Comm_2D, ierr)

where

MPI_Comm_world	:	default communicator for all processes
Ndim = 2	:	topology dimension
dim (1 :2)	:	number of processes in each of the two directions
period(l:2)=false	:	periodicity in the two directions
Organ=false	:	imposed organization of the processes
Comm_2D	:	new communicator for the processes of the topology

Conditioned branches inside the code can run different fragments of the program, depending on the process to which they belong. This is done via an identification number assigned by MPI to each process and return by a useful function available in MPI (namely *MPI_Comm_rank*).

Eight communication instructions are executed at each time iteration. They concern the E_y , E_z, H_y and H_z components for the y-z interface, and the E_x, E_z, H_x and H_z components for the x-z interface.

Now, the global parallel algorithm is written from the various different points can be described mainly the parallelism steps:

1. MPI initialization
   Determination of the process number and their ID number
2. Reading of simulation parameters
3. Creation of two-dimensional Cartesian topology
4. Creation of the derived data types for communication purpose
5. Start time iterations (time-stepping)
   Computation of the E-field components
   Communication of the E-field components at the

```
        subspace boundaries
    Computation of the H-field components
    Communication of the H-field components at the
        subspace boundaries
6. End (MPI-FINALIZE)
```

4. EXPERIMENT OF 2D PBG CALCULATION

Photonic crystals, also known as photonic band gap (PBG) materials, are artificially engineered dielectric materials that exhibit a frequency regime over which propagation of light is strictly forbidden [8].

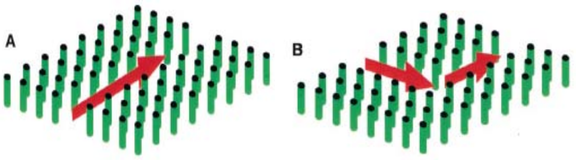


Fig.5. Schematic diagram of a photonic crystal

A linear defect in a photonic crystal can give rise to a band of defect states within the gap and act as a waveguide (see Fig. 5A). Light in the photonic crystal is confined to and guided along the one-dimensional channel because the gap forbids light from escaping into the bulk crystal. A waveguide bend (Fig.5B) can then steer light around a sharp 90° corner. A simple scattering theory predicts the existence of reflection nodes where 100% transmission efficiency can be achieved through the bend. Waveguiding of electromagnetic (EM) waves is demonstrated by a PBG line defect and, more importantly, observed near perfect transmission of EM waves around a sharp corner in a photonic crystal.

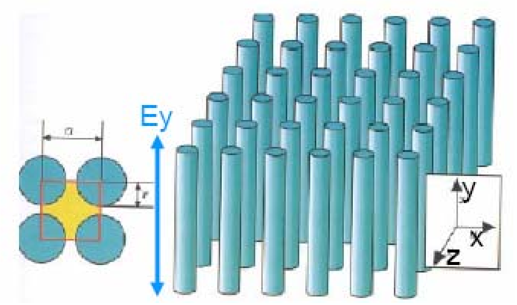


Fig.6. Band structure of 2D photonic crystal

According to the parallel architecture and the optimized MPI program described above, a test computing environment with two PC which connected with 100Mbps Ethernet is built. The two PC are 2.8GHz/512MB, which have Redhat Linux 9.0 MPICH-G2 (mpich 1.2.7) and Fortran 90. The purpose of the experiment is to run conventional MPI programs across multiple parallel computers within the same machine room for distributed execution of a large computational electromagnetics code.

The two-dimensional (2D) photonic crystal used to construct straight waveguides and waveguide bends consisted of a square array of circular Alumina rods in a dielectric constant ε as 8.9 and in radius *r as* 0.20*a*, where *a* is the lattice constant of the square array. In the experiment, the dielectric constant of the background material was fixed to be ε=1 (air), the lattice constant was chosen to be 0.59 μm (infrared band). In this case, a 12-layer PML medium is employed as the ABCs for all side.

The spatial step:

$$\Delta x = \Delta y = \frac{a}{N} = \frac{0.59}{41} = 0.014\mu m$$

The time step:

$$\Delta t = \frac{1}{c\sqrt{\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}}} = 3.4 \times 10^{-17} s$$

The Gaussian excitation:

$$E_i(t) = -\cos\omega t \exp\left[-\frac{4\pi(t-t_0)^2}{\tau^2}\right]$$

For such a 2D photonic crystal, large photonic band gap exists for light polarized parallel to the rods that extends from a frequency of 1.63×10^{14}Hz (0.32*c/a*) to 2.24×10^{14}Hz (0.44*c/a*). Where, *c* is the speed of light.

Numerical simulation result shows that the band gap would have been centered at a wavelength λ=1.55 μm, and the guided-mode bandwidth Δλ would have extended over a range of 430 nm.

5. CONCLUSIONS

In this paper, a parallel implementation of the FDTD algorithm is discussed. First, it was constructed a grid parallel computing environment based on MPICH-G2, then adopted the message passing interface (MPI) to write the parallel FDTD program in order to accelerate the whole computational speed in the PC cluster system. It was also performed a detailed numerical analysis of photonic band structure of 2D square lattices consisting of circular Alumina rods in air by parallel FDTD method. By using a heterogeneous FDTD method, the formulation can be employed to study the band gap phenomenon in different geometric shapes, material properties, and design patterns. The result of parallel FDTD method is verified by the serial FDTD algorithm. The result shows that the parallel algorithm is correct and reasonable, which can be developed to solve the 3D PBG numerical simulation

The ongoing work is to distribute applications across computers located at different sites, performance the dynamic process management for a wider class of grid computations in which either application requirements or resource availability changes dynamically over time.

REFERENCES

[1] L. F. Marsal, T. Trifonov, “Larger absolute photonic band gap in two-dimensional air-silicon structures,” *Physica E*, Vol.16, 2003, pp. 580-585.

[2] K. S. Yee, “Numerical solution of initial boundary value problems involving Maxwell’s equations in isotropic media,” *IEEE Transactions on Antennas and Propagation* Vol.14, No.3, 1996, pp.302-307.

[3] Ming-Sze Tong, Yilong lu, Yinchao Chen, “Design and analysis of planar printed microwave and PBG filters using an FDTD method,” *Microelectronics Journal*, Vol.25, 2004, pp. 777-781.

[4] Qi Huang, Jianbo Yi, Shi Jing, “Development of an MPI for power system distributed parallel computing in wide area network with p2p technology,” *IEEE PARELEC2006*.

[5] C. Guiffaut, K. Mahadjoubi, “A parallel FDTD Algorithm

Using the MPI Library," *IEEE Antennas and Propagation Magazine*, Vol.43, No.2, 2001, pp.94-103.

[6] N. T. Karonis, B. Toonen, I. Foster, "MPICH-G2:A grid-enabled implementation of the message passing interface," *Journal of Parallel and Distributed Computing* ,Vol.63,No.5,2003,pp.3-10.

[7] L Tarricone, A Esposito, L Tarricone. *Grid Computing for Electromagnetics.* Artech House, 2004.

[8] S. Y. Lin, E. Chow, V. Hietala, "Experiment Demonstration of Guiding and Bending of Electromagnetic Waves in a Photonic Crystal," *Science*, Vol.282, 1998,pp.274-276.

[9] Du Juan, Zhang Yijun, Zhu Shouzheng, "Parallelization of the FDTD Algorithm over PMICH-G2," *APMC2005 Proceedings*.

[10] G. Festa, S. Nielsen, "PML Absorbing Boundaries Bulletin of the Seismological Society of America," Vol.93, No.2, 2003, pp.891-903.

[11] Deo Prakash Vidyarthi, Anil Kumar Tripathi, Biplab Kumer Sarker, "Cluster-Based Multiple Task Allocation in Distributed Computing System," *IEEE Proceedings of the 18th International Parallel and Distributed Processing Symposium*, 2004.

A Remote Visualization System Based on GOS2 *

Guihua shan[1,2], Jun Liu[1], Yunhai Wang[1,2], Xuebin Chi[1], Zhonghua Lu[1]
[1.]Computer Network Information Center, Chinese Academy of Sciences
Beijing, 100080, China
[2.] Graduate Univ., Chinese Academy of Sciences
Email:{sgh,liujun,wyh,chi,zhlu}@sccas.cn

ABSTRACT

Remote visualization service is an indispensable part for super computing center and any computing grid. We present a grid-enabled remote visualization system architecture-ScVisGrid. It provides a web-based interface for end users to access visualization resources in SCCAS, launch visualization task in interactive mode. With component-based visualization framework, the existing standalone software or new functions could be easily and flexibly integrated into remote visualization system, which makes the system convenient, flexible and scalable. Two application examples are also provided to demonstrate the usefulness of the developed system.

Keywords: grid computing, remote visualization, GOS2

1. INTRODUCTION

With the rapid development of computer and computing technology, scientific computing becomes the third main method [1] for people to conduct research, which greatly extends scientist’s research ability. Because of the greatly increasing size of the data being generated, scientific visualization becomes an integral part of scientific computing, which provides the most effective means to analyze the data and grasp the insight of it. Normally, scientists, data, and visualization software are distributed in different places in the world. However, visualization software and the cost-expensive hardware equipments needed for visualization are not available in every desktop, so remote visualization service plays an indispensable role in supercomputing centers.

The Supercomputing Center of Computer Network Information Center, Chinese Academy of Sciences (SCCAS) is one of the biggest centers that provide supercomputing service in China. There are two high performance supercomputers (Deepcomp6800, and Dawning2000), a visualization supercomputer (SGI Onyx350), a PC cluster with 20 nodes where each node with an nvidia geforce6200 (Vizcluster), as well as a storage system of 60TB. Deepcomp6800 was installed in SCCAS 2003, with 4.2Tflops. SGI Onyx350 and Vizcluster are mainly for visualizing the scientific simulation result data.

For network security, typically users of SCCAS do not allowed accessing an x server to remotely display the visualization. To provide web-based interface to remote visualization resources in interactive mode, We designed a novel grid-enable visualization service architecture based on GOS2[3].With this novel architecture we provide a transparent, easy web access to visualization resources in SCCAS.

2. RELATED WORK

Works on web-based visualization could be traced to ten years ago. C.S.Ang presented one of the first web-based visualization system in [4], where an applet was employed to display visualization result. Since then, many works such as [5] and [6] make const-ant progress on web-based visualization. As the emerging grid technology, people pay attention to combine web-based visualization and Grid [7-10]. Most notable one is the visportal [7] [8] developed by LBNL and UCD. In this system, web-based visualizati- on for some visualization software is enabled by employing a visualization web application and a portal application server, where the integrated software is limited to those that they are able to work as server. Similar work also could be found in [9] where vtk-based visualization job in batch mode can be submitted through portal. Another interesting work for visualization in grid is done in [10], where an XML application for des-cribing visualization task is developed and the visualization pipes could be distributed on the different Grid resources.

Compared with those grid enabled visualization system, our system focuses on a general framework for easily integrating existing and under developing software and makes these visualization tools more easily accessible by providing web-based interfaces

3. SYSTEM ARCHITECTURE

3.1. Design goal

The goal of our system could be summarized in three points:

Minimize the setup of user’s desktop
No special software in user’s local machine needed and no high performance hardware or special operation system required. Remote visualization system does visualization on remote machine, and returns the results to user by internet.

Offer an easy-use visualization environment
Shield users from complex grid environment and simplify the process of using remotely located tool. Automate some steps which are needed before starting visualization pr-ocess, such as extracting sub-data from huge scientific data, converting data format and so on.

Make good use of existing visualization software
A lot of excellent visualization tool have been developed, most of them have becoming popular. Integrating these software as components into the system will promote usability and quicken development of the system

3.2 Architecture

* this work is supported by “Construction of the network computing application system of astronomy, bioinformatics and computational chemistry”, which is sponsored by Ministry of Science and Technology of China (2005DKA64002), the National Natural Science Foundation of china “Research on parallel algorithms and implementation aimed at problems of characteristic finding”(60673064), and “Research on high performance scientific computation”,973program

The detailed architecture of ScVisGrid is shown in Fig.1. The system is composed of four parts: client, Grid Supporting module, Viz Framework module, Component module.

The left is the client, web-based interface. User could launch Visualization task through the web-based interface and the results are displayed interactively in the browser.

The middle box-Grid Supporting module is in charge of grid related functions of resources, user and task management. It's the foundation of ScVisGrid.

The bottom right is component module, which is a group of application software conformed to certain communication protocol. They could execute visualization tasks under Viz framework module scheduling.

The top right box Viz Framework service module is the core part of the system, which locates in visualization server (in our case, SGI onyx350 and Vizcluster). It manages process of visualization and related data. The visual results produced during the visualization process are put into database automatically or transfer to applicant.

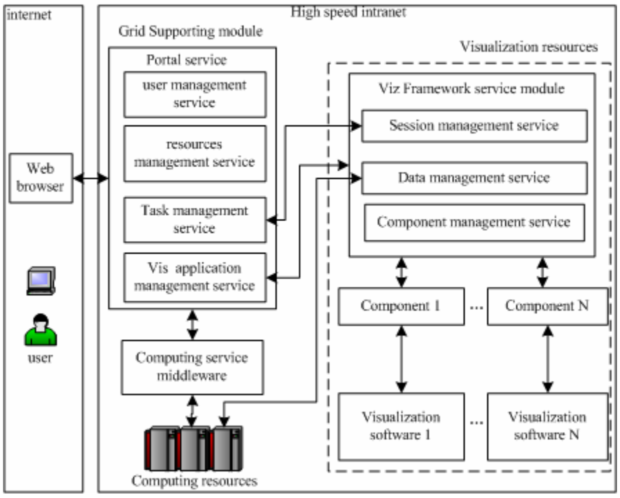


Fig.1. ScVisGrid Architecture

3.2.1 Grid supporting module

Grid supporting module is the foundation of ScVisGrid. It adopts GT2.0 as Grid middleware. The portal is built on GPTK and java CoG. Tomcat/JSP is employed as web engine in web server. Grid supporting module uses the Globus to realize user, task and resources management service. The resources management service is responsible for resource registering, monitoring, and allocation. The user management service is responsible for user authentication. The task management service is responsible for launching, scheduling and monitoring tasks. The Vis application management service is responsible for communication between Viz framework module and client once visualization task started, which is a necessary part for supporting interactive mode.

3.2.2 Visualization Component module

Visualization Component module is the exact module responsible for execute visualization task under Vis framework service module's scheduling. Component is derived from one or more existing software by wrapping the software with a communication layer, which makes component able to communicate with component management service layer. Of course a component also could be standalone software, but it must conform to the protocol of component management service. The wrapped components could call one or more software to execute visualization task and get visual result.

We have developed VM library (visualization module library) and VMU library (visualization module utility library) to simplify the wrapping work. VM provide a simple API for the communication between component and component management service layer while VMU is API based on VM for more convenient component initialization and message parse.

3.2.2 Viz Framework service module

Viz Framework service module is responsible for manage visualization process and related data. It does not directly do any visualization function by itself. The goal of this module is to provide a general purpose visualization framework. It is not limited to some specific software or visualization algorithm. In fact it offers a framework to effectively arrange visualization components to execute visualization tasks. Viz Framework service is composed of three sub-modules: session management service, data management service and component management service. They work together to finish a visualization task. The detailed responsibilities of each sub-module go as following:

Session management service: responsible for creating and deleting session, send user's commands to component management service and require visualization results from data management service.

Data management service: responsible for offer input data to component management service, manage the output from it and visualization results.

Component management service: responsible for schedule all the visualization components.

Fig.2. shows the intra-structure of Viz Framework service module and relationship among the three sub-modules.

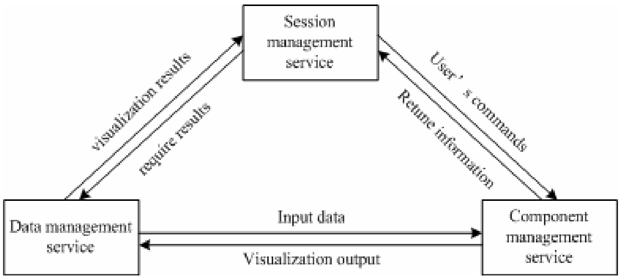


Fig.2. the internal structure of Viz framework service module

To use the ScVisGrid, user login ScVisportal and download the web-based interface after authorization. A user proxy service is created for that user. When user launches a visualization application through the interface, the task information is sent to user proxy service to start a new specific task management service, which consults resource management service to require sufficient resource and locate the application to run the task. After successful allocation, the task management service sends the task to corresponding Viz framework service module on the selected visualization server with sufficient resource, at the meantime creates a specific Vis application management service to manage the visualization session and keep contact with Viz framework service module. All these complex processes are hidden to users.

4. TWO EXAMPLES OF APPLICATION

There are several different services accessible in ScVisGrid, which include the existingstandalone software and the software developed by ourselves. One of the most import-ant commodity usages is to use the remote located standalone software。Here we des-cribe two examples:

(1) GMT in ScVisGrid

GMT is widely used visualization software with command line interface which is hard for user to remember the complex commands. Becker[11] designed Tcl/Tk based interface for GMT, called iGMT(interactive GMT)(Fig.3.)，which make the interface more friendly for users. But iGMT is still standalone software which can't be used in grid. We make it remote accessible by integrating it into remote visualization system with VMU library (Fig.4.).

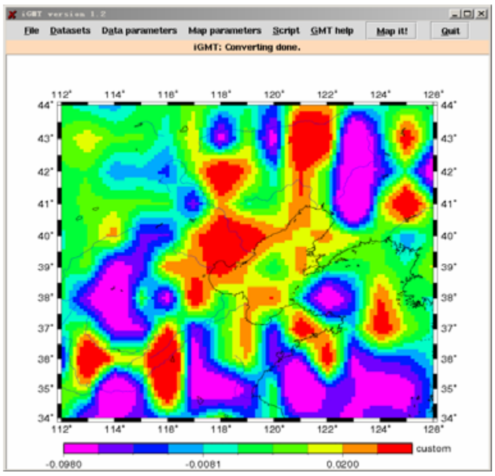


Fig.3. interface of iGMT

GMT integrated in ScVisGrid include all the function supported by iGMT, moreover the functions have been modulated. Multi-instances could created at the same time and multi results produced by functions supported iGMT could be arranged in one visualization result, which is helpful for user to analyze data.

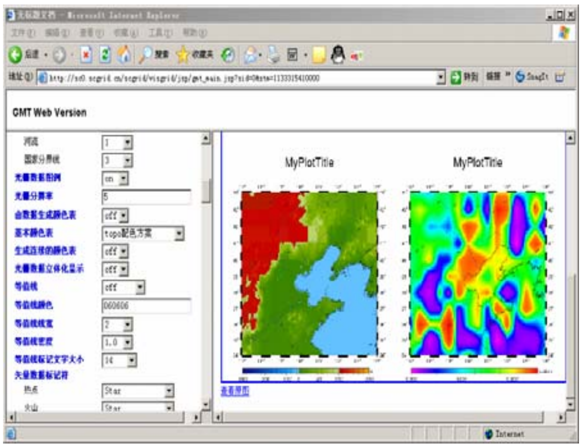


Fig.4. GMT integrated in ScVisGrid

(2) Interactive Isosurface Maker with AVS/Express

AVS/Express is a powerful commercial visualization tool. In the visportal devel- oped by T.J.Jankun-Kelly[7], AVS/Express is only able to work in two modes: one is that it works as a thick client on desktop in their Visportal, where data is download from data server.

But AVS/Express is not available in every desktop since it is an expensive commercial tool; moreover the power of PC is limited; the other mode is that AVS/Express application works as a batch job on server to m-ake a movie where interactive control is not supported. In our case, AVS/Express works in remote server as a back engine and interactive control is available. In fa-ct, AVS/Express itself supports a special "C/S" mode, in that the client and server must run on the same machine. We encapsulate the application client as a comp-onent in our remote visualization system with the VMU. For a more flexible cont-rol, the interface is a java applet in web browser. This greatly extends the capabil-ity of AVS/Express. At present we test it by the isosurface function; other functi-on can be obtained with the similar method.

5. CONCLUSION AND FUTURE WORKS

Visualization service is indispensable in grid computing. We designed a novel architecture which provides a web-based interface to remote visualization resources in SCCAS and simplifies the usage of general visualization software. In this architecture, we make good use of excellent exiting visualization software without changing their source codes. The system is flexible and scalable to encapsulate more general visualization software to provide grid-enable services.
In future, we plan to employ SRB middleware for data management, and integrate some application works more tightly with scientific simulation such as computational steering visualization to dynamically visualize the output from simulation program.

REFERENCES

[1] Cheng Guoliang, parallel *computing.* (High Education Press, Beijing, 2001)

[2] Haili Xiao, Hong Wu, Xuebin Chi, Sungen Deng, Honghai Zhang: An Implementation of Interactive Jobs Submission for Grid Computing Portals. ACSW Frontiers 2005: 67-70

[3] H.Wu, X.B.Chi, F. Xu, "Creation of Web-Based User interface for Supercomputing Environment", 2002 5th international conference on Algorithms and Architectures for Parallel Processing, Beijing.

[4] C.S.Ang, D.C.Martin, M.D.Doyle. Integrated control of distributed volume visualization through the world-wide-web. Proceedings of the IEEE Conference on Visualization 1994, IEEE Computer Society Press, Los Alamitos, CA, USA. 1994

[5] V.P.Holmes, J.M.Linebarger, D.J.Miller, R.L.Vandewart, C.P.Crowley. Evolving the web based distributed SI/PDO Architecture for high performance visualization. 34th Annual Simulation Symposium 04, 2001, Seattle,WA,USA

[6] J.Wood, K.Brodile, H.Wright, "Visualization over the world wide web and it's application to environment data". Proceedings of the 9th Eurographics Workshop on Visualization, 81-86, Los Alamitos,CA 1996.

[7] T.J.Jankun-Kelly, O.kreylos, J..Shalf, K.L.Ma, B.Hamann, K.I..Joy, and E.W.Bethel, "Deploying web-based visual exploration tools on the grid", IEEE

Computer Graphics and Application,3,40-50(2003)

[8] Wes Bethel, Cristina Siegrist, John M. Shalf, S. Shetty, T.J. Jankun-Kelly, Oliver Kreylos, Kwan-Liu Ma , Proceedings of the 3rd Annual Workshop on Advanced Collaborative Environments, 2003

[9] Charlie Moad, Beth Plale, “portal access to parallel visualization of scientific data on the grid”, Indiana University Bloomington, IN. Technical Report TR593. Feb.2004

[10] Ken Brodlie, David Duce, Julian Gallop, Musbah Sagar, Jeremy Walton, Jason Wood Proceedings of IEEE Visualization 2004, edited by Holly Rushmeier, Greg Turk and Jarke J. van Wijk, pp155-162

[11] Becker, T. W. and Braun, A.: New program maps geoscientific data sets interactively. EOS Transactions AGU, 79, 505, 1998.
http://www.seismology.harvard.edu/~becker/igmt/

[12] K. W. Brodlie and J. Wood. Recent Advances in Volume Visualization. Computer Graphics Forum,20(2), The Eurographics Association and Blackwell Publishers Ltd, pp. 125-148, 2001

Simulation Technology of Three-dimension Environmental Field Based on Large-Scale Distributed Computing in VR of Ship Navigation

Jian Deng [2], Liwen Huang [1,2], Yuanqiao Wen [2], Jinfeng Zhang [2]
[1] State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Beijing, China
[2] Navigation School, Wuhan University of Technology, Wuhan, Hubei Province, China
Email: dengjian_whut@yahoo.com.cn

ABSTRACT

Ship navigation simulation is a hot topic in recent years, of which the simulation of Three-Dimension environment field is a key problem. Different from traditional method, this paper focuses its emphasis on how to get a real marine environment. Depending on the single air, sea, wave numerical model, this paper develops a Union Environment Dynamical Forecast System (UEDFS) based on large-scale distributed scientific computing, by which the marine environment can be detail simulated. Through the experiment of a real case, the performance of UEDPS is proved that it can provide a high precise and harmonious environment field. At the same time, the simulated datum of wind, current and wave can be introduced into calculating the ship motion in this marine environment. Virtual reality technology is used to create the three dimension sea scene by VC and OpenGL and it has taken a good result.

Keywords: Environment Forecast, Numerical Model, Virtual Reality

1. INTRODUCTION

Virtual Reality (VR) is a new technology developed mainly on the base of computer graphics technology, simulation technology and sensor technology et al. This technology is now widely used in navigation training, evaluation of hydraulic construction, battlefield environment simulation and so on. In the past few years, this technology has been introduced into ship navigation simulation. By this way, marine environment scene on special place can be reconstructed and the ship movement can be detail described. So, how to construct a “real” three-dimension environment scene is the base of this work. By now, the environment datum, such as wind, rain, are determined just by man-made setting in ship navigation simulator[1]. This subjunctive environment are great different from the reality. Many works has been paid much attention on how to make the scene vision more ‘real’ and how to get a more exact ship movement equation. Through these researches, many progresses have been gotten[2]. In comparison with that, little work has been done on how to provide the real environment datum, which can definitively affect the vision effect and ship movement.

Marine environment is a very complex system including sea surface wind, current, sea height, sea surface temperature, wave height and so on. Some of them, such as sea surface wind, ocean current and wave directly influence the ship’s movement. So how to get a reality marine environment is very important. In general, there are three ways to deal with this problem: geometry, hydrodynamic and statistic method. Numerical simulation is one of hydrodynamic methods and has been widely applied in weather and ocean simulation[3]. But because of the large computing cost problem, large scale numerical simulation is rarely used in ship scene virtual reality. By this way, it can get a near-real environment field and also make the ship navigation simulation results more believable.

2. IMPLEMENT OF THE UNION ENVIRONMENT DYNAMICAL FOCAST SYSTEM

2.1 The Structure of the UEDFS

Marine environment is such a place, where atmosphere, ocean and wave affect each other. Wind can drive ocean current and generate wave. On contrast, it can be influenced by them through roughness length and sea surface temperature. In order to improve the forecast accuracy of them, these kinds of feedback mechanism should be well described. Many models have been developed in the world, but these ones just study one kind of subject. Air model can just simulate the atmosphere motion and ocean model just used research the sea. How to unite these single models and consider the interaction between them is a difficult problem. At the same time, numerical model always bring the problem of large computation cost, which makes the calculation efficiency can not be insured. This paper has done much work on these two questions and developed a Union Environment Dynamical Forecast System (UEDFS), which consists of four parts: atmosphere forecast module, ocean forecast module, wave forecast module and coupler module. The mechanism of their effect to each other is described by fig.1.

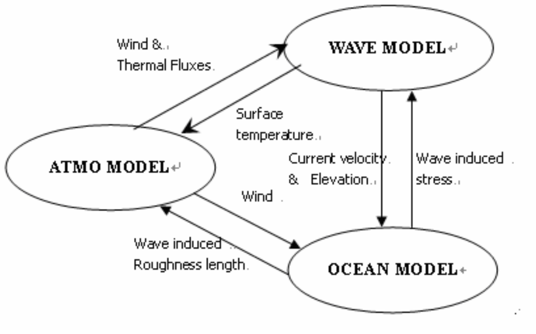


Fig.1. exchange processes between three models

Atmosphere can influence the ocean through momentum flux (wind stress) and heat flux (sensible heat flux, latent heat flux, long wave radiation and short wave radiation). On contrast, ocean can affect the atmosphere by sea surface temperature. At the same time, atmosphere can generate wave by wind and wave can change the sea surface roughness to influence the air. Also, there is current-wave interaction between the ocean and wave. So, multi environment elements can be forecasted together in these models by exchanging their information during their forecasting process.

2.2 Atmosphere Forecast Module

*National Natural Science Foundation of China No. 40275015

Atmosphere forecast module can forecast the atmospheric elements, such as sea surface wind, rain and visibility. Mesoscale Model (MM5) [4] is selected in this paper, which is developed by Pennsylvania State University / National Center of Atmospheric Research. This model is a limited-area, nonhydrostatic, terrain-following sigma-coordinate model designed to simulate or predict mesoscale atmospheric circulation. In this paper, the atmospheric domain covers entire East Asia, which centered at 33°N，125°E. Its grid system is constructed with 187×141 horizontal grid points in medium resolution of 37.5 km. There are 15 sigma levels in the vertical. As to get a high-resolution data, there is also a nest domain with resolution 12.5km covering East China Seas. Of course, the nest domain also covers entire ocean model domain. Time step of atmospheric model used in this study is 120s.

2.3 Ocean Forecast Module

Ocean forecast module is the most important part of UEDFS and used to forecast the current, elevation and sea temperature[5]. This paper chooses the Princeton Ocean Model (POM). It is a three-dimensional, fully nonlinear, primitive equation ocean model. The oceanic model domain covers the entire Yellow and East China Seas (YECS) from 23°N to 41°N by latitude and from 116°E to 131°E by longitude (showed as Fig.2). A rectangle grid is constructed with 164×202 horizontal grid points with horizontal resolution of 10 km. In the vertical, there are 11 sigma levels. Time step of oceanic model is 480s. The ocean domain has five open boundaries including the Changjiang River Estuary, the Taiwan Strait, the Kuroshio inflow boundary located at northeast of Taiwan, the Tsushima (Korea) strait and the Kuroshio outflow boundary located at south of Japan (Tokara strait). As to enable ECOM-si have forecast capacity, tidal forcing is also considered. Four inflow/outflow boundaries are also regarded as open tidal boundaries. The sea surface elevation at four tidal boundaries can be calculated based on equation (1):

$$\eta = \overline{\eta} + \sum_{i=1}^{n} f_i H_i \cos\left[\sigma_i t + (V_{0i} + u_i) - g_i\right] \qquad (1)$$

where η is the predicted sea surface elevation, $\overline{\eta}$ is monthly averaged sea surface elevation (MSL), f_i and u_i are nodal factors in amplitude and phase of the constituents，V_i the phase of corresponding equilibrium constituents, H_i is amplitude，σ_i is angular frequency and g_i is delay phase. In this study, six tidal constituents ($S_2, M_2, N_2, K_1, P_1, O_1$) have been considered.

2.4 Wave Forecast Module

Wave forecast module is used to simulate the significant wave height, wave direction and wave length. It is a key part to form the ocean scene. A fully spectral third-generation ocean wind-wavemodel——Wavewatch-III (henceforth denoted as WWATCH), has been recently developed at the Ocean Modeling Branch of the Environmental Modeling Center of the National Centers for Environmental Forecast (NCEP) for the regional sea wave forecast. It was built on the base of Wavewatch-I and Wavewatch-II as developed at the Delft University of Technology and the National Aeronautics and Space Administration (NASA) Goddard Space Flight Center, respectively (Tolman 1999). In this paper, wave model has the same horizontal grid system as the ocean model.

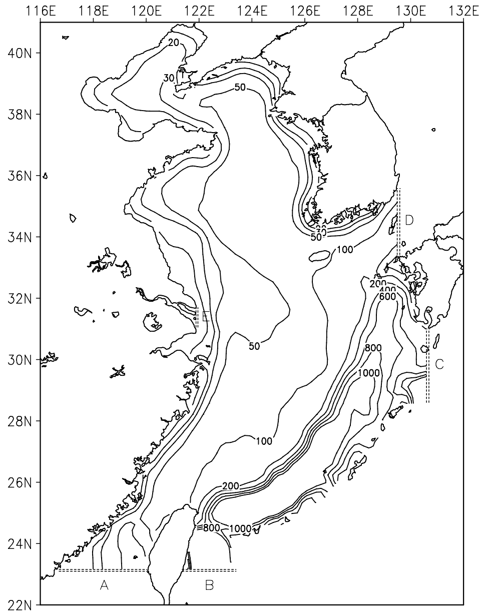


Fig.2. Geography isobaths(m) and open boundary of Yellow and East China Seas.

2.5 Coupler Module

The main function of coupler module is to exchange the information among different models and provide the parallel computing. Former three forecast modules will bring large computing cost, which influence the efficiency of UEDFS. Distribution computing is a useful way to solve this problem. Different from foreign traditional coupler[6], this paper takes atmosphere forecast module, ocean forecast module and wave forecast module as the prototype, designing a distributed coupling computation environment. The realization of this technology is based on SOCKET and PIPE communication. It enables different models run on different computing platform and exchange information through network and keep synchronization. The computing environment can shield the complexity of the system and it is flexible, extensible and collaborative. It is convenient to construct coupled model by using the legacy codes and to realize the coupling of multi models on multi computing environment. It can solve the large computation cost problem and keep effectiveness.

2.6 Test of UEDFS Performance

As to test the performance of UEDFS, a real tropical cyclone case has been selected. Typhoon Winnie No.9711, which occurred between 8 August and 20 August 1997, is a typical landfall one and its intensity is strong. In this paper, UEDFS forecasted the marine conditions from 12 UTC 16 August 1997 to 12 UTC 19 August. The NCEP data with resulotion 1º ×1 º has been selected as the initial field of UEDFS. Figure 3 shows the YECS circulation affected by typhoon at the 48 hour. There is an obvious cyclonic circulation induced by and moved with typhoon. It is accord with the reality. Whether these results are reliable, the simulated significant wave height (SWH) has been compared with the observed datum of TOPEX/ Poseidon altimeter (Fig.4). There is a good agreement between the observed and the model data. All of the results can prove that the UEDFS has a good performance. It can get the near reality environment data and provide the base for ship movement simulation in the real ocean environment conditions.

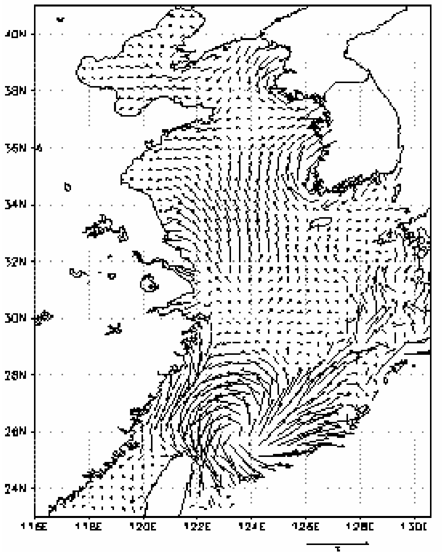


Fig.3. Surface currents field (m/s) at 48 hour

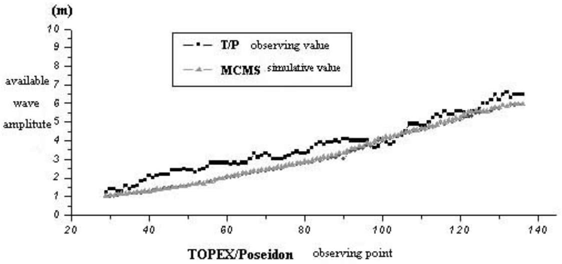


Fig.4. Comparison of SWH simulated by the UEDFS and the observed TOPEX/Poseidon data

By UEDFS, the environment datum on regular grid point can be gotten. Therefore, the information on any other place also can be calculated by interpolation method

3. SHIP NAVIGATION SIMULATION

From above it can be known that it can get the atmosphere and ocean elements at a certain situation by UEDFS, then it can use simulated datum to forecast ship movement and motion. Ship navigation simulation includes two aspects. The first is ship movement (ship track), the other is ship motion (movement pattern). The ship's movement pattern under the various sailing situations has been deeply studied[7]. In this paper, a traditional method is applied.

As to simulate the movement track of ship, sea surface wind and sea surface current are used in the equation (2) and (3).

$$\begin{cases} S_X = (V_S \times Sin\ \alpha + V_C \times Sin\ \beta) \times T + \Delta B_X & (2) \\ S_Y = (V_S \times Cos\ \alpha + V_C \times Cos\ \beta) \times T + \Delta B_Y & (3) \end{cases}$$

where S_X , S_Y are the distances of ship movement in x and y direction respectively, V_S is the ship speed, V_C is the current speed, α , β are the angles of ship velocity and current velocity respectively. ΔB_X , ΔB_Y are the excursion distances induced by wind in x and y direction respectively, which can be calculated by equation (4):

$$\Delta B = K \times (\frac{B_a}{B_w})^{\frac{1}{2}} \times e^{-0.14 V_S} \times V_a \times T \qquad (4)$$

where K is coefficient (0.041), B_a is windward area of hull over waterline, B_w is windward area of hull under waterline, V_s is ship speed, V_a is relative wind speed, T is the length of forecast time.

As to ship pose, its forecast is more complex. Ship pose in the ocean connects with not only hydrodynamic performance, but also the environment conditions. The force on the hull can be defined as follow equation:

$$\begin{cases} X = X_W + X_C + X_F & (5) \\ Y = Y_W + Y_C + Y_F & (6) \end{cases}$$

X , Y are the force on hull in x and y direction respectively, X_W , X_C and X_F are the force induced by wind, current and wave in x direction respectively, Y_W , Y_C and Y_F are the force induced by wind, current and wave in y direction respectively. These variables call can be calculated by corresponding function, such as force induced by wind can be describe by follow equation:

$$\begin{cases} X_W = 0.5 \rho_a A_f V_a^2 C_{Xa}(\alpha) & (7) \\ Y_W = 0.5 \rho_a A_S V_a^2 C_{Ya}(\alpha) & (8) \end{cases}$$

where ρ_a is the air density, A_f is the hull projection area over waterline on x direction, A_S is the hull projection area over waterline on y direction, C_{Xa} and C_{Ya} are coefficients.

By combining the ship hydrodynamic parameters and forecasted environment elements, the 3D virtual sea scene with sky and wave can be constructed. Therefore, the 3D real time dynamic simulation of ship's time-related motions (horizontal swing, horizontal sway and head-sway) in this scene can be realized.

4. VIRTUAL REALITY OF MARINE ENVIRONMENT

After forecasting the marine environment and ship motion, the next step is to use the forecast results to create a 3D virtual sea scene. In this paper, OpenGL is selected to do it. As to make the sea scene more real, virtual reality technology is introduced. In spite that so many environment elements can be forecasted by UEDFS, such as wind, sea surface temperature, current, wave, rain, visibility, cloud and so on, only few of them, such as wave ,current and visibility, is useful for virtual scene making. There are so many simulation software, such as Vega, 3DMax, OpenGL, developed to do these work. Sea scene includes many subjects. This paper just describes how to draw the wave there. Firstly, it is important to get the data on entire eyeshot sea surface. Because UEDFS can only get the wave data on net grid point, it should interpolate these data of grid points to entire eyeshot sea surface. This value is related to significant wave height, wave direction and wavelength. Secondly, the sea surface color should be optimized. Because the value of wave height are gradual changed, sea surface color can also be considered gradual changed. The sea surface colors of maximum and minimum wave height are defined beforehand and the sea surface color of any other wave height can be gotten by interpolation technology. Finally, it is important to set light effect on the wave surface. A typical illumination model in compute graphics is used in this paper.

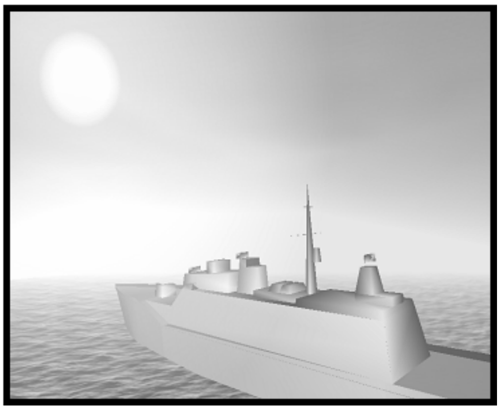


Fig.5. The effect of ship navigation simulation on certain ocean scene

5. SHIP NAVIGATION SIMULATION

As to accurately simulate the ship navigation in marine environment, this paper using numerical simulation technology developed a Union Environment Dynamics Forecast System. Through this way, all of the marine environment parameters have been detail simulated. After that, the simulated datum of wind, current and wave are introduced into simulating the ship movement and motion in this marine environment. At the same time virtual reality technology are used to create the 3D sea scene including marine environment and ship motion. It makes the vision effect better. This research proves UEDFS is a useful tool to simulate the ocean environment. It has good performance and can be widely used in research of virtual reality.

REFERENCES

[1] Zhou M, Xu X G, Zhu T. *Realtime Simulation of Large Scale Sea Battlefield[J]. Act*a ArmamentarII, 27(3)pp6-11, 2006

[2] TRANSAS.INC. *Transas Navigational Simulator NTPro 4000 User Manual*.,2003

[3] LIU X H, YUAN Y L. *Scalar Analysis and Similarity Conditions of Marine Environment Physical Simulation, Advances in Marine Science*, Vol.2, 2006.

[4] Dudhia，J., D. Gill, Y-R.Guo et al.，*PSU/NCAR Mesoscale Modeling System Tutorial Class Notes: MM5 Modeling System Version 3*, 2004.

[5] Zhu J R, Ding P X, *Zhu S X. Numerical simulation of the circulation in the Huanghai Sea and the East China Sea in summertime*. Acta Oceanologica Sinica,Vol.11,2002,

[6] Zhou T J, Yu Y Q, Yu R C, Liu H L, L W and Zhang X H. “Coupled Climate System Model Coupler Review，” *Chinese Journal of Atmospheric Sciences*,Vol.6,2004

[7] Gan L X. “Prediction of ship motion by using the linear response equation of maneuverability,” , *Ship & Ocean Engineering*, Vol.4,2003

Forward-looking Scene Matching Based on Hough Transformation and Phase-only Correlation*

Fangfang He[1], Jiyin Sun[1], Wenpu Guo[1], Libo Sun[2]
[1]Xi'an Research Inst. of High-tech, Xi'an 710025 China
[2]School of Science, Xi'an Jiao Tong University, Xi'an 710049 China
Email: rouqiu1027@163.com

ABSTRACT

Forward-looking scene matching is a fire-new technique for terminal guidance of intending precision-guided weapons. As conventional template matching methods can not overcome perspective transmutation of forward-looking imaging, a new method based on Hough transformation and Phase-only Correlation is proposed in this paper. Firstly use Hough idea to extract the longest line segment both from reference image and real-time image. Then, compute rotation descriptor and reconstruct reference image so as to map the image pair into a same reference frame. Latter matching algorithm with phase-only correlation will be completed between the reconstructed image and real-time image. Experimental results over a range of rotational attacks show a satisfied matching probability and matching speed, which give a demonstration for deep research of forward-looking scene matching.

Keywords: Forward-Looking Scene Matching; Matching Method, Perspective Transmutation, Hough Transformation, Phase-Only Correlation.

1. INTRODUCTION

The early long-distance launching weapons can accurately strike martial object because it is guided by automatic inertia navigation system. In order to keep a shoot straight, terrain aided navigation technique used in the midway of long-distance flight was advanced in the fifties of the twentieth century. Later, digital down-looking scene matching technique was put forward successfully in the terminal guidance of precision-guided weapons. Except that, to enhance pinpoint attack ability of new type weapons launched outside defence area, forward-looking scene matching technique with excellence of speediness and anti-jamming is to be used in the terminal control and guide.

Forward-looking scene matching system has its particular advantage in three aspects. First, it can focus attention upon the interesting area times without number. Second, as imaging forward, it can perform martial scoutcraft and strike. Third, it can discover object in advance and pinpoint them. Without question, forward-looking scene matching technique is sure to show its superiority in aided navigation system along with the progress of relevant research.

However, the research of forward-looking scene matching is only underway in china. There is even no information or papers published in public firsthand to this field, so it's essential to search after new ways and means to bring to success. Down-looking scene matching can use normative correlation method to get matching position, but when the image pair is incoordinate because of perspective transmutation brought by visual angle change, there will exist rotation or size change between reference image and real-time image and the conventional template matching will be disabled because template can only move parallel. On the other hand, the imaging type of forward-looking is perspective which may lead to the distortion of scene. And the distortion can change with many imaging factors, such as the focus of perspective and the stance of object and so on. Under this condition, if we continue to use conventional methods, matching capability certainly will be affected for its ambiquity and instability. So we must plan a new matching idea to overcome the projection transmutation.This paper proposed a new method based on Hough transformation and Phase-only Correlation for forward-looking scene matching.

The paper is organized in the following way. In the next section, we describe the basal idea of forward-looking scene matching. Then the proposed scheme is designed in Section 3. The key techniques are described in Section4. The performance of the method is evaluated in Section 5. Finally, our conclusion is given in Section 6.

2. BASAL IDEA OF FORWARD-LOOKING SCENE MATCHING

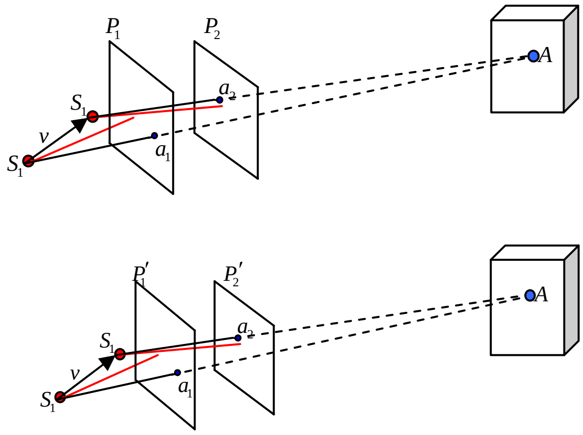


Fig.1. Sketch map of forward-looking scene matching

According to agressire route of precision-guided weapons, the 3D scene of interesting area can be prepared aforehand and forward-looking reference image of 2D is also securable. Thus the 3D scene is described by a series of 2D forward-looking images: $p_1, p_2, \ldots, p_n$. $s_1, s_2, \ldots, s_n$ is focus of each image and v is a speed vector. When precision-guided weapons fly the agressire route in record time, it will also screen a series of images of temporal $p_1^{'}, p_2^{'}, \ldots, p_n^{'}$, then matching p_i and $p_i^{'}$ ($i = 1 \sim n$) can educe the orientation error in flight.

* Supported by the National Defence Preliminary Research Foundation of China.

Having known the flight stance of weapons, we can estimate that $p_i^{'}$ ($i = 1 \sim n$) is a best match of certain sequence ($p_i ,\ldots, p_{i+k}$) which selected from p_1 , p_2 ,…, p_n . As the choosing agressire route which determine the visual angle of 3D scene and forward-looking reference image of 2D is impossible entirely accorded with the real one and the distance of screen time is fixed value (not infinitesimal), there'll surely exist some perspective transmutation between $p_i^{'}$ ($i = 1 \sim n$) and $p_i ,\ldots, p_{i+k}$. (showed in Fig.1)

3. FORWARD-LOOKING MATCHING WITH ROTATION-DESCRIPTOR AND AFFINE RECONSTRUCTION

3.1 Perspective Transform and Affine Transform

Image transform includes geometrical transform and dimensional transform. It's a function to build mapping connection between an image and its transformative one. Perspective transform, affine transform and polynomial transform are the three main form of image transform.

Perspective transform is similar to affine transform in following aspects:

(1) They are both planar mapping.

(2) They can keep shape of line segment in any direction

Considering that weak perspective transformation can be approximated with planar affine transformation, we will use planar affine transformation to express the coordinate mapping among two-dimensional images.

3.2 The Proposed Scheme

As we known, weak perspective transformation can be approximated with affine transformation when runway is represented in perspective, the matching scheme is proposed as showed in Fig.2.

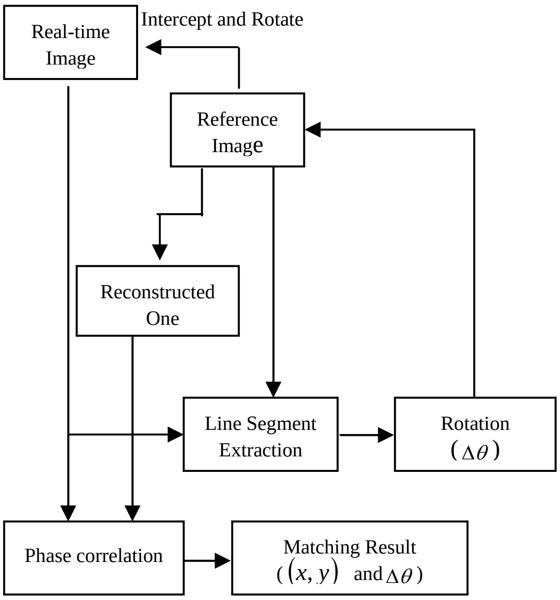


Fig.2. Scheme of proposed

3.3 Matching Algorithm

After the foregoing design, the matching algorithm is as follows:

Input: reference image and real-time image

Output: rotation of the image pair and matching position (x, y)

(1) Step1: detect longest line segment L from reference image and real-time image.

(2) Step2: compute rotation (direction difference $\Delta\theta$).

(3) Step3: reconstruct reference image.

(4) Step4: using phase-only correlation to compute translation of real-time image and reconstructed reference image.

(5) Step5: give outcome (x, y)

4. KEY TECHNIQUES

4.1 Feature Selection

(1) Select line segment of runway as basal feature

Feature is dividing line of objects, such as vertex and edge and so on. To match the image pair, the feature of them must correspond one by one. That is to say the corresponding feature should point to a same position of objects and the selected feature is of course representing the same physical marking point while screened on the two different images.

Airdrome is a place to safeguard all kinds of aero activities where airplanes will flying-off, landing and parking. Airdrome is composed of runway, taxiway, parking apron, blindage, communication equipment, observation window, observatory, magazine and barracks and so on. Among them, runway has the most valuable feature for its distinct line. So the proposed method select longest line segment as the basal feature before affine reconstruction.

(2) Feature extraction

Using the conventional idea of Hough transformation, we can detect line from runway. As the length of runway is a bit longer (generally about 1000m to 5000m), the detected edge of runway will not be continuous and it is easy to miss some part if using other means but not Hough transformation. Anyway, Hough transformation can detect line segment perfectly. The extraction steps are as follows:

1) Suppose the size of airdrome (Q) is $w \times h$, and ρ and θ is variable of polar coordinates, where $-\sqrt{w^2+h^2} \le \rho \le \sqrt{w^2+h^2}$, $0 \le \theta \prec 180$. Apply a memory space, num is accumulation with its upper limit $\left(2 * \sqrt{w^2+h^2} + 1\right) * 180$. Then Apply memory space to memory beginning point and end point of line segment, named $beginX$, $beginY$, $endX$, $endY$ with the same upper limit $\left(2 * \sqrt{w^2+h^2} + 1\right) * 180$. And the adding pace of θ is 1.

2) Scan Q to find point $A(x, y)$ whose grey is 255, then compute ρ from (1) where θ take turns from 1 to 179. In order to compute num , we need change the range of ρ from 0 to $2 * \sqrt{w^2+h^2}$, and this

make us sure that will not be a negative quantity. Add 1 to $num[\rho][\theta]$ according to the corresponding ρ and θ.

$$\rho = x\cos\theta + y\sin\theta \qquad (1)$$

(3) Once get a new $num[\rho][\theta]$, judge $num[\rho][\theta]$ whether it is 1. If it is 1, note down the current coordinate as $(beginX[\rho][\theta], beginY[\rho][\theta])$, or else note down the coordinate as $(endX[\rho][\theta], endY\rho][\theta])$.

(4) When scan over, find the max. of and name the corresponding ρ and θ as $\rho_{\max}$ and $\theta_{\max}$. That means the direction of the longest line segment is $\theta_{\max}$ where coordinate is $(beginX[\rho_{\max}][\theta_{\max}], beginY[\rho_{\max}][\theta_{\max}])$, $(endX[\rho_{\max}][\theta_{\max}], endY[\rho_{\max}][\theta_{\max}])$.

Through the steps mentioned above, the longest line segment L can be detected from Q. Compare the $\Delta\theta(=\theta_{\max}^{reference} - \theta_{\max}^{real-time})$ between reference image and real-time image, affine transformation parameter can be figured out.

4.2 Affine Reconstruction

Having known the affine transformation parameter calculated from 4.1, we can transform reference image under this parameter and this will map reference image into a same reference frame with real-time image. Thus the following matching process can be done between the transformed image and the original real-time image.

4.3 Phase-only Correlation

The aim of this section is to determine the translation of the image pair. We consider a simple two-dimensional translation problem with translation offset (x_0, y_0), thus, $x^{'} = x - x_0$, $y^{'} = y - y_0$. The Fourier transforms of the two images are defined by: $R(u,v) = \Gamma\{r(x,y)\}$ and $S(u,v) = \Gamma\{s(x,y)\}$ with $\Gamma\{\}$ denoting the Fourier transformation. The classical matched filter, which maximizes the detection signal-to-noise ratio, has a transfer function

$$H(u,v) = \frac{R^*(u,v)}{\left|N(u,v)\right|^2}$$

where $R^*(u,v)$ is the complex conjugate of the Fourier spectrum $R(u,v)$ and $\left|N(u,v)\right|^2$ is the noise power-spectral density. If the noise has a flat spectrum with intensity n_ω, the transfer function of the matched filter reduces to

$$H(u,v) = \frac{1}{\left|n_\omega\right|^2} R^*(u,v)$$

and the output of the filter is the convolution of $r^*(-x,-y)$ and $s(x,y)$:

$$q_0(x,y) = \frac{1}{\left|n_w\right|^2} \iint_{-\infty}^{+\infty} s(a,b) r^*(a-x, b-y) dadb$$

This function has a maximum at (x_0, y_0) that determines the parameters of the translation.

5. RESULTS AND DISCUSSION

In order to test the proposed method, program with MARLAB7.0 on a computer whose CPU is 2.0GHz, EMS memory is 512M and OS is Windows2000. Reference image (as showed in Fig.2 (a)) is a satellitic image of airdrome and its size is 512×512. Although this image is not a real perspective projection, the proposed method only uses its line feature under weak perspective projection, so it doesn't matter. Rotate reference image to simulate perspective transformation, and after that blocks of Real-time image with the size 256×256 can be obtained by interception. Note that, memory the coordinate of interception location. Enslaved to length of this paper, we only show real-time image when rotation is 30° (in Fig.4) and Fig.5 shows a curve of calculated rotation compared with those actual values. From the curve we can see that the proposed method can calculate the rotation accurately in most cases. The corresponding reconstructed reference image is showed in Fig.6. Fig.7 is an outcome of matching error of 10 pairs under the measure of summation of absolute errors. Table 1 shows the statistical results of matching when rotation is 10°, 30° and 45°.

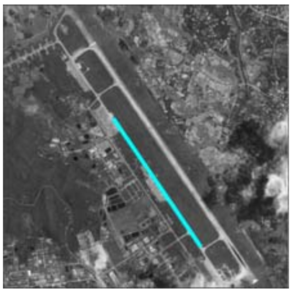


Fig.3. Reference image and its longest line segment

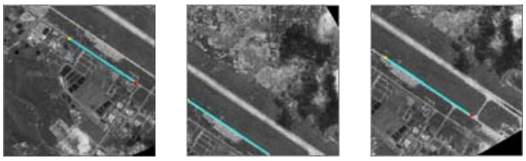


Fig.4. Blocks of real-time image and their longest line segments (Rotation 30°)

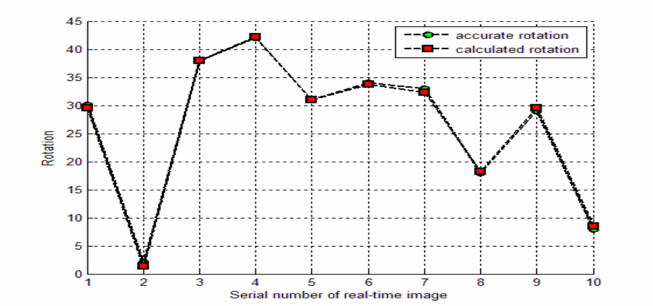


Fig.5. Calculated rotation (in red) and actual rotation (in green)

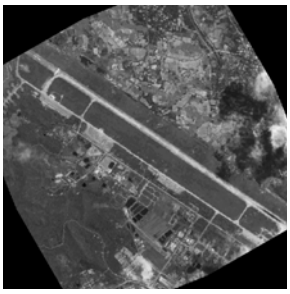


Fig.6. Reconstructed reference image

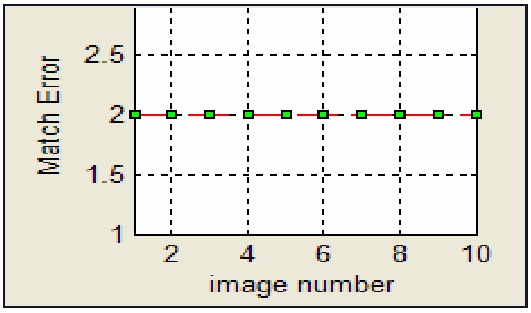


Fig.7. Match error of the proposed method (10 times)

Table 1. Example of matching result

Rotation / *Matching result*	10°	30°	45°
Probability (%)	100	100	96
Precision (pixel)	1	2	2
Time (s)	1.2	1.3	1.3

From Table 1, we can see that even if rotation is biggish, the method has good matching capability. To sum up, we analyze the method as follows:

(1) General method request feature of interesting area is unique and invariable. But feature used in this paper is to progress affine reconstruction and this determines it only needs to have invariability to offer the location mapping.

(2) The method is different from object registration. Object registration is usually use shape of objects, so it request objects has some shape complexity. But the proposed method think a lot of line feature which is not sensitive to perspective transformation and will be more simple and efficient.

(3) As reference image and real-time image having been mapped into a same reference frame, any matching algorithm can be used to calculate similarity measure including down-looking matching algorithm.

(4) General method is invalid to object rotation while this method can counteract rotation and make full use of image information. Thus its matching result can offer more information, such as rotation degree and matching location and so on.

6. CONCLUSIONS

A forward-looking matching method aiming at weak perspective was advanced in this paper. Considering the distinct line feature of runway and invariable direction of its longest line segment, Hough transformation was used to compute affine transformation parameter before reference image can be reconstructed. Plentiful experiments show that this method can resist biggish affine transformation and has a good stability.

7. ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and associate editors for their insightful comments that helped improve the quality of this work. Their comments and suggestions have contributed greatly to the depth and clarity of the presentation.

REFERENCES

[1] Iain Matthews, Takahiro Ishikawa, Simon Baker. “The template update problem.” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, June 2004, 26(6):810-815.

[2] Toshimitsu Kaneko, Osamu Hori. “Template update criterion for template matching of image sequences.” *Proceedings of 16th International Conference on Pattern Recognition*, August 2002, 2:1-5.

[3] L.Lucchese. “A frequency domain technique based on energy radial projection for robust estimation of global 2D affine transformations.” *Computer Vision and Image Understanding*. Jan. 2001, 81:72-116.

[4] Tomas Suk, Jan Flusser. “Projective Moment Invariants.” IEEE Trans. *on Pattern Analysis and Machine Intelligence*. 2004, 26(10):1364-1367.

[5] Q. Ji, R.M. Haralick. “Error propagation for the Hough transform.” *Pattern Recognition*. 2001, 22(6):813-823.

[6] M.K.Ibrahim, E.C.L.Ngau, M.F.Daemi. “Weighted Hough Transform.” *Intelligent Robots and Coputer Vision Techniques*. 1991, 16(07):239-241.

[7] P.S. Modenow, A.S. Parkhomenko. “Geometric Transformations.” *New York: Academic*, 1965, Vol.1

[8] Faugeras O. “Stratification of 3D-vision: Projective, Affine, and Metric Representation.” *Journal of the Optical Society of America*, 1995, 12:465-484.

Research on Window Switching Technology in AVS Decoder*

Cong Zhang[1,2,3], Ruimin Hu[1,2], Chunming Yuan[3]
[1]School of Computer, [2]National Engineering Research Center for Mutimedia Software,
Wuhan University, Wuhan, Hubei, 430079, China
[3]Department of Computer Engineering, Wuhan Polytechnic University, Wuhan, Hubei, 430023, China
Email: hb_wh_zc@163.com

ABSTRACT

AVS is a fundamental audio and video coding/decoding standard developed by the Audio and Video Coding Standard Workgroup of China, which is stepping forward steadily so that it becomes a Chinese official standard early. In this paper, the key technology of AVS audio decoding standard is simply reviewed firstly. Then the research of window switching in AVS decoder is described based on theoretical analysis. Finally, some experimental results and future research directions of AVS decoder are reported.

Keywords: Audio, AVS Decoder, Window Switching, Perceptual Entropy

1. INTRODUCTION

For the sake of obtaining compact digital representations of high-fidelity audio signals and without compromising reproduction quality, tremendous research efforts have been put in the development of efficient digital audio coding technologies. Currently AAC is the latest audio compression standard released by Moving Picture Experts Group (MPEG) which has been working on the standardization of high quality low bit rate audio coding. But China has not own digital audio coding international standards yet. In order to promote China own researches in the area of multimedia , keep up with the development of international advanced technology and lower the burden of equipment manufacturers, the AVS Audio subgroup organized on June 2002, and issued the first call for proposal on stereo audio coding systems in August 2002. After constantly technical proposal evaluation and verification, China AVS Audio subgroup built working draft on stereo audio coding. Now the AVS Audio subgroup is busy at working out conformance test rules.

We know that the central objective of audio standard is high-quality digital audio delivery at low bit rates while achieving transparent signal reproduction and allowing tradeoffs in audio encoding/decoding complexity for different applications. In this paper, according to the standard specification and related experiments, AVS decoding process is discussed and explained how it obtain this resolution.

2. AVS DECODING PROCESS

Compared to AVS audio encoder, its decoding process is relatively simple. The decoding process is approximately reversed one of encoding except without psychoacoustic model. The generalized block diagram of AVS audio decoding process is presented in Fig.2. Now AVS decoder only supports mono/dual PCM audio signal.

In AVS audio decoding schemes, the first parsed step is CBC decoding. CBC decoding is a process that is reversed of CBC encoding. It employs all sub-decoders or partial sub-decoders according to output speed ratio or the received complexion of bit-stream. In each sub-decoder, each bit-plane vector is respectively decoded from higher bit to lower bit and from lower frequency to higher frequency. As a result, by CBC decoding, higher decoding efficiency is achieved.

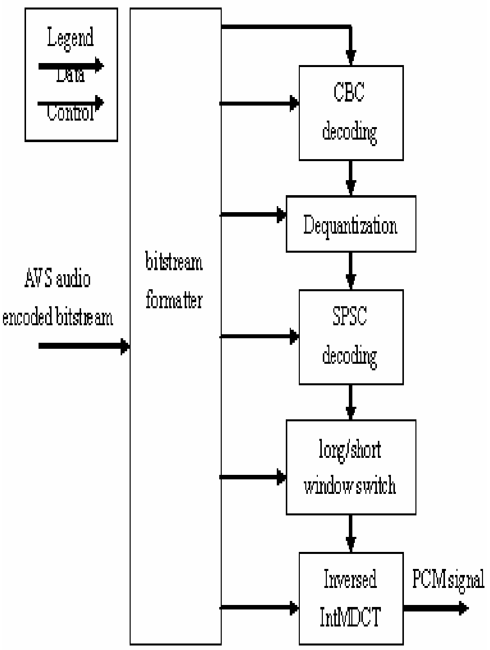


Fig.1. AVS Decoder Block Diagram

After CBC, Square Polar Stereo Coding is applied in frequency domain. The process of SPSC coding is similar with Mid/Side Stereo Coding but adopt L/R mapping and has higher coding efficiency, it is a lossless transform. In AVS audio schemes, SPSC is applied to each paired channels. While paired channels have strong correlation, SPSC will achieve great coding gain. In SPSC, L/R channels have different quantization noises, only one channel has the noise overlap. Therefore, if quantization modules at encoder produce the same quantization noise, quantization noise at decoder for SPSC will be lower than Mid/Side.

In last parsed phase, CHINA AVS employs an inverse modified discrete cosine transform (IntMDCT). We know that most of the current audio coding/decoding schemes use transforms like the Modified Discrete Cosine Transform (MDCT) to calculate a block-wise frequency representation of audio signals in order to increase the time resolution and decrease the pre-echoing effect. But for lossless audio coding, MDCT can not guarantee that the original audio signal will be reconstructed perfectly even there is no any further noisy processing. However, Integer MDCT can commendably address this problem and meets our

* Project (B200618002) supported by the science foundation of Hubei Provincial Department of Education.

expectations. It contains two processes, first time domain windowing and then DCT-IV transform. It can produce integer output values instead of floating point values. It can provide a good spectral representation of the audio signal, overlapping of blocks, critical sampling, good frequency selectivity and a fast algorithm. At the same time, the complexity is greatly reduced which results in improved performance of a lossless audio coding/decoding system employing the IntMDCT.

3. THE RESEARCH OF WINDOW SWITCHING

We know that sometimes audible fake signal starts before true one. This phenomenon is so-called pre-echoing which is one of most notorious audio encoding artifacts. Listeners may discern noises between audio intervals. This will make it sound muddled. The fundamental cause is a lack of time domain resolution.

In modern transform-based audio coding/decoding standards, we tend to employ longer block transform for finer frequency domain resolution in order to obtain more coding/decoding efficiency and more audio quality for stationary signals. But this will reduce time domain resolution—quantization noises spreading over whole transform blocks. For transient signal like castanet, audible pre-echoing appears. So to address this problem, AVS audio standard employs long/short window switching mechanism: long window (2048 samples) for stationary signals and short window (256 samples) for transient ones. A method called ENUPM-WSD (Energy and Unpredictability Measure based Window Switching Decision) is recommended in the AVS standard for long/short window decision.

3.1 Long/short Window Switching Mechanism

ENUPM-WSD employs a two-stage decision-making scheme (shown in figure 2). In the first stage, a frame of 1024 samples signal is evenly partitioned into 16 sub-blocks which has 64 samples. According to the maximal changing of the block energies, or

$$\max_{k=-2,-1,\dots,15}\left\{\left(\sum_{i=64k}^{64k+127} X_i^2 \Big/ \sum_{i=64(k-1)}^{64(k-1)+127} X_i^2\right)-1\right\}, \qquad (1)$$

if it is less than *E_SWITCH* (constant, 3.0) in case the last frame is stationary or 2**E_SWITCH*/3 in case the last frame is transient, long window will be selected for the current frame, where X_i s are input samples.

Only if the last frame is stationary and (1) is greater than *E_SWITCH* will ENUPM-WSD proceeds to the second stage. In this stage, first every 128-point block with 50% overlap is hanning windowed and FFT transformed. Put them in formulas, denote

$$\vec{\mathrm{X}}_k = (X_{64k}, \cdots, X_{64k+i}, \cdots, X_{64k+127}), \qquad (2)$$

And

$$\vec{\mathrm{Y}}_k = (Y_{k,0}, \cdots, Y_{k,i}, \cdots, Y_{k,127}) = \mathrm{FFT}(\vec{\mathrm{X}}_k \bullet \vec{\mathrm{H}}), \qquad (3)$$

where X_i s are as before; $k=-2,-1,\dots,15$, index for each transform block; FFT is Fast Fourier Transform, and $\vec{\mathrm{H}}$ hanning window vector defined by

$$\vec{\mathrm{H}} = (H_0,\cdots,H_i,\cdots,H_{127}), \text{ and } H_i = 0.5(1\text{-}\cos(2\pi(i+0.5)/128)\,. \qquad (4)$$

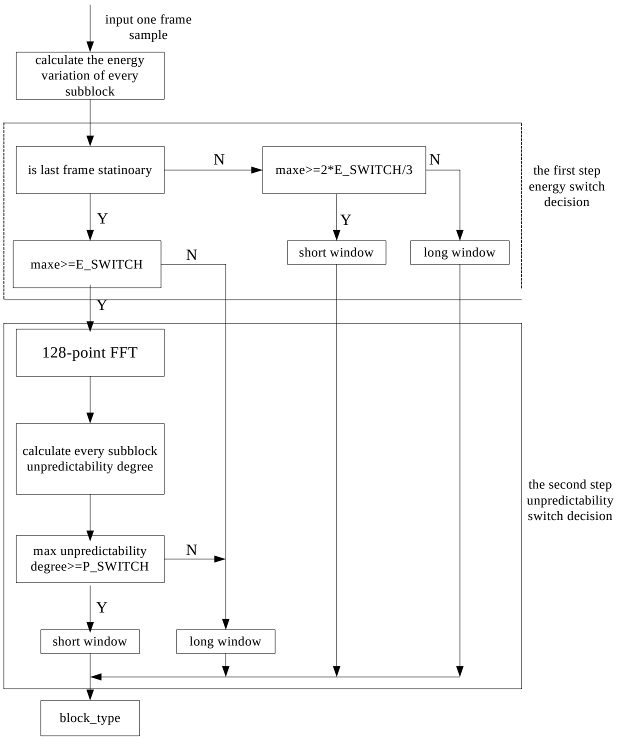

input one frame sample
calculate the energy variation of every subblock
is last frame statinoary
N
maxe>=2*E_SWITCH/3
N
the first step energy switch decision
Y
Y
maxe>=E_SWITCH
N
short window
long window
Y
128-point FFT
calculate every subblock unpredictability degree
the second step unpredictability switch decision
max unpredictability degree>=P_SWITCH
N
Y
short window
long window
block_type

Fig.2. Two Stage Long/Short Window Switching Flow Chart

Then we come to the calculation of unpredictability $\phi_{k,i}$ for each frequency line, or

$$\phi_{k,i} = \frac{\left|Y_{k,i}^{p} - Y_{k,i}\right|}{\left|Y_{k,i}^{p}\right| + \left|Y_{k,i}\right|}, \qquad (5)$$

where $Y_{k,i}^{p}$ is predicated the value for block k line i and $k=0,1,\dots,15; i=0,1,\dots,63$. It is evaluated by linear differential prediction of aptitude and phase respectively, or

$$Y_{k,i}^{p} = \left(2\left|Y_{k-1,i}\right| - \left|Y_{k-2,i}\right|\right) e^{j\left(2\arg(Y_{k-1,i}) - \arg(Y_{k-2,i})\right)}, \qquad (6)$$

Finally we obtain unpredictability Φ for the current frame by finding the maximal weighted sum of $\phi_{k,i}$ s for block 0 to 15, or

$$\Phi = \max_{k=0,1,\dots,15}\{\sum_{i=0}^{63} \left|Y_{k,i}\right| \phi_{k,i}\}\,. \qquad (7)$$

If Φ is less than *P_SWITCH* (constant, 0.1), long window will be applied or short window.

3.2 Comparison with Traditional Perceptual Entropy

By exploiting time domain and frequency domain signal characteristics, ENUPM-WSD has the merits of low complexity: for a major part of most signals, time domain (first stage) decision is sufficient to choose long or short window, which is greatly reduced in computational complexity; and gain higher accuracy: for a small part of signals whose transient is vague if judging only by time domain information, ENUPM-WSD delves into the second stage resorting to unpredictability for high decision accuracy.

As far as traditional audio standards be concerned, Perceptual Entropy (PE) is used to make long/short window decision. If the value of PE is higher than the predefined threshold PE-SWITCH, the window length will change into short window. But a fundamental problem is that what PE discriminate is signal’s noiselessness or tonality, not transient signal. For instance, multi-harmonic signals suitable to long window tend to have higher PE value and short windows may be wrongly applied. Another drawback is its computational complexity (masking threshold for every critical band need to be calculated out). For a piece of castanet ‘si02’ which is famous for its transient, ENUPM-WSD has better discrimination over signal’s transitory property while less short windows (1/6 of overall frames) are requested by ENUPM-WSD compared to that of PE (2/3 of overall frames), which is more efficient from the decoding prospective.

4. CONCLUSIONS AND FUTURE WORK

4.1 Subjective Listening Test

An informal subjective listening test was implemented between AVS and the most popular audio compression format AAC. Four sequences: es02, sc02, si02, and sm02 (Shown in Table 1) that had different character of audio signals were employed in this test. Bit-streams are 128kbps coded. The testing system is a PC with a high quality headphone. The subjective test followed “ITU-T P.800/P.830 seven-point comparison category rating (CCR) method”. The testing evaluation results are shown in Figure 3. We can draw the conclusion that AVS audio at 128kbps stereo is the same as AAC and achieves “indistinguishable” CD audio quality. Thus the goal of AVS has been achieved and AVS audio standard is stepping forward steadily.

Table 1. Four Test Sequences

	Length (s)	Sampling rate(kHz)	Type	mono /stereo
es02	8	48	male speech	stereo
sc02	12	48	symphony	stereo
si02	7	48	castanet	stereo
sm02	10	48	bell	stereo

4.2 Future Work

For China AVS audio subgroup, future work in audio standard will be concerned with the development of multi-channel support, lossless coding and establish a mobile audio standard (AVS-M Audio Standard) to meet the third generation mobile communication**.**

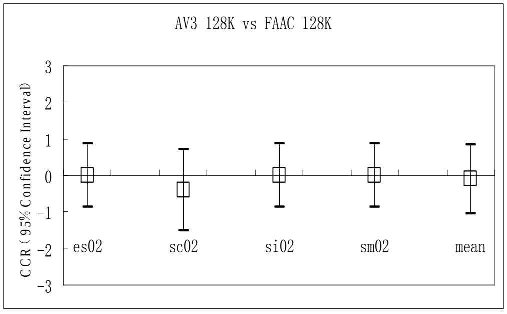


Fig.3. AVS 128kpbs vs. AAC 128kpbs

REFERENCES

[1] China AVS audio subgroup, Information Technology, “Advanced Audio and Video Coding Part 3:Audio” , 2005.

[2] MPEG, “Coding of Moving Pictures and Audio,” *International Standard IS 138187-7,* ISO/IEC JTC1/SC29 WG11, 2004.

[3] Hu Ruimin, Chen Shuixian et.al, “AVS Generic Audio Coding,” *Sixth International Conference on Parallel and Distributed Computing Applications and Technologies (PDCAT'05)* pp. 679-683, 2005.

[4] MiaoLei and Sang-Wook Kim, “Context-dependent Bit-plane Coding (CBC),” *AVS-M1368,* 2004.8.

[5] Yan JiangXin and Zhang ShuHua, “Square Polar Stereo Coding (SPSC),” *AVS-M1436*, 2004.10.

[6] Hung HaiBin, Yu Rongshan et.al, “A High Precision Integer-MDCT Transform And Fast Algorithm,” *AVS-M1366*, 2004.8.

[7] Ai HaoJun, Deng GuiPing et.al, “An Integrated Energy and Unpredictable Measure Based Window Switching Decision ,” *AVS-M1434,* 2004.10.

[8] H. HaiBin, R. Susanto, Y. Rongshan and L. xiao, “A Fast Algorithm of Integer MDCT for Lossless Audio Coding,” *Proc. IEEE ICASSP* 2004.

Cong Zhang is a Ph.D. candidate in school of computer, Wuhan University. He received the Master Degree in Computer Application from Wuhan University of Technology in 1999, and Bachelor Degree in Automation Engineering from Huazhong University of Science and Technology in 1993. His research interests are in multimedia network communication and audio processing.

Ruimin Hu is a Professor, and PhD Director at Wuhan University. He received the Ph.D. degree in Communication and Information System from Huazhong University of Science and Technology in 1994, and the Master Degree and Bachelor Degree in Communication and Information System from Nanjing University of Posts & Telecommunications in 1984 and 1990. He is Younger Director of China Society of Image and Graphics, a senior member of China Audio and Video CODEC Technical Specialist Group. His research interests include multimedia signal processing, multimedia communication system theory and application, pattern recognition, QoS over heterogeneous network.

An Overview of Several Principal Variants of the Ambient Calculi

Minglong Qi, Qingping Guo, Luo Zhong
School of Computer Science & Technology, Wuhan University of Technology,
Ma Fang Shan Campus, 430070 Wuhan China
Email: mlki01@sohu.com

ABSTRACT

In this paper, we present some principal variants of the ambient calculi, of which the expressiveness and the limitation are compared and discussed. For illustrating the communicative mechanism and the expressiveness of the each calculus, some new examples have been coded in each languages, and from which, we could conclude that some variants of the pure Mobile Ambient, such as the Extension of Boxed Ambient (NBA), the Channel Ambient Calculus are more suitable to specify the transport or cryptographic protocols.

Keywords: Ambient Calculi, Mobile Ambient, Boxed Ambient, NBA, Channel Ambient Calculus, Seal Ambient

1. INTRODUCTION

The essential features of a distributed system are distributed locations, mobility of agents that could be mobile codes or messages of communication between two locations or two agents, and concurrency or parallelism of the processes running at the different locations. For modeling these features of distributed systems, Cardelli and Gordon have, in [1], invented a new process calculus called the Mobile Ambients (MA) that has a direct relationship with a variant of the pi-calculus [2]. In the MA, the key notion is the ambient. An ambient is a *bounded* place where computation happens, that can contain both processes and sub ambients, move inside a sibling ambient, and go out of the parent ambient. An ambient can be moved as an entity. In addition, an ambient should have a name.

Even though the MA is very expressive in modeling the distributed systems, according to certain authors [3], it has posed some new hard problems such as the interference of the movement of ambients due to non-determinism. In addition, the pure MA has a simple and rigid communicative mechanism between ambients. For remedying the leak of the MA, some variants of the MA have been born, where we cite for example, the Seal Calculus [3], the Boxed Ambients [4], the Extension of Boxed Ambients (NBA) [5], the Channel Ambient Calculus [6], etc. In this paper, we will be limited to discuss the pure MA, the Boxed Ambients (BA), and the NBA.

The article is structured as fellows: in section 2, we will study three ambient calculi: the pure Mobile Ambients (MA), the Boxed Ambients(BA), and the Extension of Boxed Ambients (NBA). We will try to code some new examples in each language, and discuss the relationship between one and another. In section 3, we will conclude our presentation.

2. SOME PRINCIPAL VARIANTS OF THE MOBILE AMBIENTS

2.1 The Mobile Ambients

In the MA [1], there are two capabilities controlling ambient movements: ***in*** and ***out***. A process headed by the ***in*** capability can make its enclosing ambient to move inside a sibling ambient, whereas a process with the ***out*** capability at the head can make its surrounding ambient to go out of the parent ambient. As an essential capability, the third capability in the MA is ***open*** that allows a process to dissolute a sibling ambient. The reaction of the three capabilities is characterized by the reduction relations presented in the next equations:

$n[\textbf{in}\ m.P|Q]|m[R] \rightarrow m[n[P|Q]|R]$ (2.1.1)
$m[n[\textbf{out}\ m.P|Q]|R] \rightarrow n[P|Q]|m[R]$ (2.1.2)
$n[m[P|Q]|\textbf{open}\ m.R] \rightarrow n[P|Q|R]$ (2.1.3)

In the equations (2.1.1)-(2.1.3), P, Q, and R denotes the processes, m and n denotes the ambient names, the symbol | stands for the parallel composition of the processes, and the square brackets delimit ambients' contents. A process as "**in** $m.P$" makes, if possible, its surrounding ambient to move inside an ambient of which the name is m, and continues to behave as the process P.

The communicative mechanism in MA is very simple, that we could resume in the reduction relation presented in the equation （2.1.4）.

$n[<M>.P|(x).Q] \rightarrow n[P|Q\{x:=M\}]$ （2.1.4）

In the equation （2.1.4）, the process "$<M>.P$" at first drops the message M in the ether of the enclosing ambient and continues to behave as the process P, while the process "$(x).Q$" at first tries to read a message from the ether of the enclosing ambient and save the message in variable x, and continues as the process Q where all occurrences of the variable x have be substituted by the message M. Please notice that the message exchange is anonymous. The symbol "$\{x:=M\}$" stands for substitution of M for the variable x. The message exchange is feasible only inside of a same ambient. The situation presented in the equation （2.1.5） will result to null part unless the ambient m will be dissolved as expressed in the equation （2.1.6） .

$n[m[<M>.P]|(x).Q]$ (2.1.5)
$n[m[<M>.P]|\textbf{open}\ m.(x).Q] \rightarrow n[<M>.P|(x).Q] \rightarrow n[P|Q\{x:=M\}]$ (2.1.6)

Example 2.1.1 *Programming a distributed, mobile and parallel system in the Mobile Ambients for calculating the arithmetic problem (a + b) * (c + d), where a, b, c and d are integers, and the symbol * is multiplication.*

Assume that the network has three nodes in which are running two processes for addition and a process for multiplication. We utilize *add1*[…], *add2*[…] and *mul*[…] these three ambients for standing for the two additional processes and the multiplicative process, that all are immobile. The ambients *na*[…], *nb*[…], *nc*[…] and *nd*[…] represents the ones that input the four integers, and all are mobile. The ambients *res1*[…] and *res2*[…] are the ones that contain the results of $a + b$ and $c + d$, respectively, whereas the ambient *res*[…] contains the final result of the calculus *(a + b) * (c + d)*. We have the next program in the MA(Fig.1):

$$Cal \triangleq (\nu\, na)(\nu\, nb)(\nu\, nc)(\nu\, nd)(\nu\, res1)(\nu\, res2)$$

$$(\nu\, res)(!\, na[\textbf{in}\ add1.\textbf{in}\ res1.<a>] \,|\, !\, nb[$$

$$\textbf{in}\ add1.\textbf{in}\ res1.<b>] \,|\, !\, nc[\textbf{in}\ add2.\textbf{in}\ res2.$$

$$<c>] \,|\, !\, nd[\textbf{in}\ add2.\textbf{in}\ res2.<d>] \,|\, add1[$$

$$!\, res1[\textbf{open}\ na.(n1).\textbf{open}\ nb.(n2).\textbf{out}\ add1.$$

$$\textbf{in}\ mul.\textbf{in}\ res.<n1+n2>]] \,|\, add2[!\, res2[$$

$$\textbf{open}\ nc.(n3).\textbf{open}\ nd.(n4).\textbf{out}\ add2.\textbf{in}\ mul.$$

$$\textbf{in}\ res.<n3+n4>]] \,|\, mul[!\, res[\textbf{open}\ res1.$$

$$(s1).\textbf{open}\ res2.(s2).\textbf{out}\ mul.<s1*s2>]]$$

Fig.1. Programming in the Mobile Ambient

In the above program in MA, the process “**!**.*P*” is a recursive process equivalent to “*P*|**!**.*P*”. The symbol “(ν *x*).*P*” stands for that a fresh name *x* is created for the process *P*. By applying the reductive equations （2.1.1）-（2.1.5）, the process *Cal* is finally reduced to the equation 2.1.7.

$$Cal \,|\, (\nu res) res[<(a+b)*(c+d)>] \quad (2.1.7)$$

2.2 The Boxed Ambients

The **open** capability is crucial in the formalism of the MA, but this capability could cause some serious problems in the real implementation of the calculus. Open an ambient signifies to deliver all its contents into the super ambient, and in the reality this is not always feasible from the view of security. Another limitation of the MA is its rigid and simplistic mechanism of message exchange that is only feasible inside a same ambient. For example, in the ambient *m*[*a*[<*M*>.*P*] | (*x*).*Q*] the process “<*M*>.*P*” is prisoner of its enclosing ambient *a* and can not exchange the message *M* with the process “(*x*).*Q*” unless the ambient *a* will be dissolved. The Boxed Ambients [5] has ameliorated the above consideration dropping the **open** capability and introducing a new mechanism of the message exchange, which allows a crossing boundary communication between parent and children ambients.

In the BA, the process “$(x)^{n}P$” stands for a read access from an ambient named *n* and continues as the process *P* after the reading operation, while the process “$<M>^{n}P$” represents a write access to an ambient named *n* and continues as the process *P* after the writing operation. The process $(x)^{\uparrow} P$ in the ambient $n[(x)^{\uparrow} P]$ stands for a read access from the parent ambient, while the process $<M>^{\uparrow} P$ in the ambient $n[<M>^{\uparrow} P]$ represents a write access to the parent ambient. The BA continues to adopt the **in** and **out** capabilities. The reductive relations are expressed in the next equations:

$$(local) \quad (x)P \,|\, <M> Q \rightarrow P\{x := M\} \,|\, Q$$

$$(input\ n)\ (x)^{n} P \,|\, n[<M> Q \,|\, R] \rightarrow$$

$$P\{x := M\} \,|\, n[Q \,|\, R]$$

$$(input \uparrow) <M> P \,|\, n[(x)^{\uparrow} Q \,|\, R] \rightarrow$$

$$P \,|\, n[Q\{x := M\} \,|\, R]$$

$$(output\ n) <M>^{n} P \,|\, n[(x)Q \,|\, R] \rightarrow$$

$$P \,|\, n[Q\{x := M\} \,|\, R]$$

$$(output \uparrow) (x)P \,|\, n[<M>^{\uparrow} Q \,|\, R] \rightarrow$$

$$P\{x := M\} \,|\, n[Q \,|\, R]$$

Example 2.2.2 *Recoding the example 2.1.1 using the formalism of the BA.*

$$Cal \triangleq (\nu\, na)(\nu\, nb)(\nu\, nc)(\nu\, nd)(\nu\, res)(\mathbf{!}na[<a>] \,|$$

$$\mathbf{!}nb[<b>] \,|\, \mathbf{!}nc[<c>] \,|\, \mathbf{!}nd[<d>] \,|$$

$$\mathbf{!}(n1)^{na}.<n1>^{add1}.(n2)^{nb}.<n2>^{add1} \,|$$

$$\mathbf{!}(n3)^{nc}.<n3>^{add2}.(n4)^{nd}.<n4>^{add2} \,|$$

$$add1[\mathbf{!}(n1)^{\uparrow}.(n2)^{\uparrow}.<n1+n2>^{\uparrow}] \,|$$

$$add2[\mathbf{!}(n3)^{\uparrow}.(n4)^{\uparrow}.<n3+n4>^{\uparrow}] \,|$$

$$(s1)^{add1}.<s1>^{mul} \,|\, (s2)^{add1}.<s2>^{mul} \,|$$

$$mul[\mathbf{!}(s1)^{\uparrow}.(s2)^{\uparrow}.<s1*s2>^{res}.res[(s)^{\uparrow}.$$

$$<s>^{\uparrow} \,|\, \textbf{out}\ mul]]$$

Fig.2. Specifying the parallel calcul (a+b)*(c+d)

in the Boxed Ambient

Because in the BA the communication can be performed crossing the boundaries of ambients, the input ambients *na*[…], *nb*[…], *nc*[…] and *nd*[…] will not be needed to move inside the additional processes *add1*[…] and *add2*[…] to perform their inputs for the four integers *a*, *b*, *c* and *d*. The reading process $(n1)^{na}$ allows to cross the boundary of the ambient *a*, whereas the writing process $< n1 >^{add1}$ permits to cross the boundary of the ambient *add1*, thus go on.

After applying the reductive relations of the BA, the process *Cal* for calculating (*a*+*b*)*(*c*+*d*), can be reduced to the next process:

$$Cal \mid res[< (a+b)*(c+d) >^{\uparrow}]$$

2.3 The Extension of Boxed Ambient (NBA)

The NBA is a variant of the Boxed Ambients (BA). The NBA has introduced the notion of co-capabilities in order to eliminate the non-determinism of the movement of ambients that could happen both in the original pure MA (Mobile Ambients), and in the BA. Citing as an example the process *k*[*n*[**in** *m.P*|**out** *k.R*]|*m*[*Q*]], the two processes "**in** *m.P*" and "**out** *k.R*" in the ambient *n* are parallel and make the behavior of the ambient *n* to be non-deterministic. The capability of movement of a process in an ambient can be performed only and just only in the presence of a correspondent co-capability. The process "**in**<*m*, *k*>" (or **enter**<*m*, *k*>) in which *m* stands for the target ambient of the move and *k* a password, can make its enclosing ambient to enter inside the target ambient *m* only and just only in the presence of a co-capability $"\overline{\mathbf{in}}(x,k)"$ (or $\overline{\mathbf{enter}}(x,k)$) running inside of the target ambient *m*. Please notice that the variable *x* stands for the name of the incoming ambient. The reductive relations for the mobility of ambients in the NBA are shown in the equations （2.3.1）-（2.3.2）. Related to its parent calculus that is the BA, the NBA adopts also the communicative mechanism of crossing ambients' boundaries but simplified. The reductive relations for communication in the NBA are shown in the equations （2.3.3）-（2.3.5）.

Example 2.3.1 *For demonstrating the expressiveness of the NBA, we will try to code the first stage of a famous cryptographic protocol that is the Needham-Schroeder symmetric-key protocol: Alice randomly generates a number called a nonce and sends the nonce, her name and the name of Bob to Trent (a server). The stage of the protocol can be expressed as* **Alice→Trent: R$_{\mathbf{A}}$, Alice, Bob**.

We utilize the ambients *Alice*[*P1*|*P2*|…] and *Trent*[*P1*|*P2*|…] for representing the processes running at the locations of Alice and Trent, respectively. Either is immobile. The processes *mess1* and *messa2* are running at the locations of Alice and Trent, respective. The ambients *NGen*[…] and *NSav*[…] are used by Alice to randomly generate a nonce and to save the nonce. The ambient $M_{A2T}[...]$ is used by Trent to save the message sent by Alice to itself. The above three ambients are also immobile. The ambient *m*[…] is used by Alice to load the message that is moving from Alice to Trent. The code in the NBA is shown in the table 2.3.1. Please notice that the ambient *Lock*[…] is used for forcing the process to load all the messages into the ambient *m*[…] before it exits from the location of Alice represented by the Ambient *Alice*[…]. The process *mess2* running at the location of Trent contains a co-capability $\overline{\mathbf{enter}}(x,k)$ for welcoming the ambient *m*[…] with the variable x representing the name of the ambient *m*[…] and *k* representing the password. In the process *mess2*, the variables *nc*, *an*, and *bn* represent respectively the nonce of Alice, the name of Alice and the name of Bob. Finally, the first stage of the Needham-Schroeder symmetric-key protocol can be specified by the global process ***Alice*[*mess1*]|*Trent*[*mess2*]**, which can be reduced using the equations (2.3.1)-(2.3.5).

3. CONCLUSIONS

In this article, we discussed three principal variants of the ambient calculi: the Mobile Ambients, the Boxed Ambients and the Extension of Boxed Ambients. Through two concrete examples coded respectively in the three formalisms: the one concerning a distributed computing problem, and the other about a cryptographic protocol, we have explained how the problems such as the non-determinism of the movement of ambients, the simplistic communicative mechanism in the Mobile Ambients, etc. have been ameliorated by the descendants of the Mobile Ambients. We demonstrated that the NBA is very suitable for specifying transport protocols.

$$n[\mathbf{enter} < m,k > .P1 \mid P2] \mid m[\overline{\mathbf{enter}}(x,k).Q1 \mid Q2] \rightarrow m[n[P1 \mid P2] \mid Q1\{x := n\} \mid Q2] \qquad (2.3.1)$$

$$n[m[\mathbf{exit} < n,k > .P1 \mid P2] \mid Q] \mid \overline{\mathbf{exit}}(x,k).R \rightarrow m[P1 \mid P2] \mid n[Q] \mid R\{x := m\} \qquad (2.3.2)$$

$$(x).P \mid < M > .Q \rightarrow P\{x := M\} \mid Q \qquad (2.3.3)$$

$$(x)^{n}.P \mid n[< M >^{\uparrow}.Q \mid R] \rightarrow P\{x := M\} \mid n[Q \mid R] \qquad (2.3.4)$$

$$< M >^{n}.P \mid n[(x)^{\uparrow}.Q \mid R] \rightarrow P \mid n[Q\{x := M\} \mid R] \qquad (2.3.5)$$

$$mess1 \triangleq (\nu NGen)(\nu NSav)(\nu Lock)(NGen[< R_A >^{\uparrow}] \mid (n)^{NGen}.< n >^{NSav}.< n >^{m}.Lock[<>^{\uparrow}] \mid NSav[(x)^{\uparrow}.< x >^{\uparrow}] \mid m[(y)^{\uparrow}.< y >^{\uparrow}.< Alice >^{\uparrow}.< Bob >^{\uparrow} \mid ()^{Lock}.\mathbf{exit} < Alice,k > .\mathbf{enter} < Trent,k >] \mid \overline{\mathbf{exit}}(x,k)$$

Fig. 3. Coding the the Needham-Schroeder protocol in the NBA

$$mess2 \triangleq (\nu M_{A2T})(\overline{\mathbf{enter}}(x, k).(nc)^{x}.(an)^{x}.$$

$$(bn)^{x}.<nc>^{M_{A2T}}.<an>^{M_{A2T}}.$$

$$<bn>^{M_{A2T}} \mid M_{A2T}[(nc)^{\uparrow}.(an)^{\uparrow}.(bn)^{\uparrow}.$$

$$<nc>^{\uparrow}.<an>^{\uparrow}.<nc>^{\uparrow}])$$

$$FirstStage \triangleq Alice[mess1] \mid Trent[mess2]$$

$$\Rightarrow Alice[(\nu NSav)(NSav[<R_{A}>^{\uparrow}])] \mid$$

$$Trent[(\nu M_{A2T})(M_{A2T}[<R_{A}>^{\uparrow}.$$

$$<Alice>^{\uparrow}.<Bob>^{\uparrow}])]$$

Fig.4. (cont.) First stage of the Needham-Schroeder protocol

REFERENCES

[1] Luca Cardelli and Andrew D. Gordon,“Mobile ambients,”*in Foundations of Software Science and Computation Strutures: First International Conference, FOSSACS ’98*, Berlin Germany,Springer-Verlag,1998.

[2] G. Berry,“Asynchrony and the pi-calculus,”*Technical Report 1702, IRINA, Sophia Antipolis*, 1992.

[3] J. Vitek and G. Castagna,“Seal: A Framework for Secure Mobile Computations,”*in International Programming Languages*, 1999.

[4] F. Levi and D. Sangiorgi,“Controlling Interference in Ambients,”*in Proceedings of POPL ’00*, pp. 352-364, 2000.

[5] M. Bugliesi, G. Castagna and S. Crafa,“Boxed Ambients,”*in Proceedings of TACS’01*,No.2215,pp. 38-63, 2001.

[6] M. Bugliesi, S. Crafa, M. Merro and V. Sassone. “Communication interference in mobile boxed ambients,”*in FSTTCS’02*, vol. 2556 of *LNCS*, pp. 71-84, Springer, 2002.

Minglong Qi, Associate Professor in School of Computer Science &Technology, WHUT, Ph. D. in visualization of scientific data achieved in the Claude-Bernard University Lyon I (France) from 1984 to 1989, had worked from 1989 to 1999 for a French software development company as analyst-programmer and project manager. He had realized several post-doctoral researches in the same period. His principal interests are in specification and verification of distributed and parallel algorithms, formal methods, etc.

Qingping Guo, Full Professor and head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science & Technology, WHUT. He is a holder of K. C. Wong Award of UK Royal Society (1994); was a visiting scholar of City University and University of West Minster(1986~1988), Visiting Professor of Queen Mary and Westfield College, London University(1997~2000), Visiting Professor of National University of Singapore(2000), Visiting Professor of University Greenwich (2003). He is one of the DCABES international conference founder, was the chairman of DCABES 2001 and 2004, and co-chair of DECABES 2002. He has published two books, over 80 Journal papers, edited two DCABES Proceedings. His research interests are in distributed parallel processing, grid computing, network security and e-commerce.

Novel Intra Prediction Algorithm in H.264 *

Xuqing Xiao[1,2], Ruimin Hu[1], Ruolin Ruan[1], Wei Huang[2], Li Zhu[1]
[1]National Engineering Research Center for Multimedia Software, Wuhan University
Wuhan, Hubei 430072, China
[2] Department of Command, The Second Artillery Command College
Wuhan, Hubei 430012, China
Email: xxq169998@sina.com

ABSTRACT

In the upcoming H.264, the intra-prediction for 4x4 and 16x16 blocks are used for the prediction of the luminance component to compress I frame. However, these intra-prediction modes so far have adopted the reference samples of same field with the current block to predict the current block on interlace video sequence. In fact, the bottom field block and the corresponding top field block have strong spatial correlation. To further achieve high coding efficiency, the paper propose novel intra-prediction mode which adopt the samples of top field block to predict the corresponding bottom block. Experimental results show that the proposed mode can remarkably reduce bitrate which maximum bitrate saving up is to about 14% while maintaining similar PSNR in field coding of full I frame.

Keywords: H.264/JVT, Intra Prediction, Interlace Video Sequence.

1. INTRODUCTION

Due to the strong correlation between current coded macroblock and adjacent macroblock, intra predication is often adopted to lower the spatial redundancy of current image when the process of current image coding can’t offer sufficient temporal correlation information. H.264/AVC takes up more precise and complex intra-prediction techniques than other video compression standards, which is one of key factors to the success of H.264. In the upcoming H.264, for intra mode a prediction block P is formed based on previously encoded and reconstructed blocks and is subtracted from the current block prior to encoding. For the luma samples, P is formed for each 4x4 block or for a 16x16 macroblock. There are a total of nine optional prediction modes for each 4x4 luma block, e.g. Vertical, Horizontal, DC, Diagonal Down-Left, Diagonal Down-Right, Vertical-Right, Horizontal-Down, Vertical-Left, and Horizontal-Up mode. Fig.1 shows eight represent direction of predictions for Intra4x4PredMode; Mode 2 (DC perdition) is modified depending on which reference samples have previously been coded. There are four modes for a 16x16 luma block and four modes for the chroma components, namely vertical, horizontal, DC, and the plane mode.

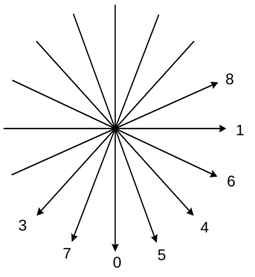


Fig.1. Intra_4x4 prediction mode directions

The enhanced intra-prediction techniques of H.264 remarkably improve its coding efficiency. However, for interlace video sequence (as shown in Fig 2.) we discover that the distance between the top field block and bottom field block is near, so their texture is very similar. In order to further reduce the spatial redundancy, this paper proposed a novel intra-prediction mode called Field-Among mode which adopt the samples of top field block to predict the corresponding bottom block.

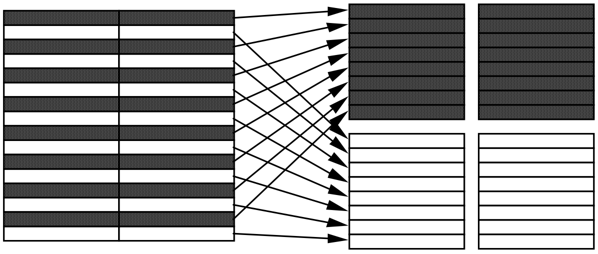


Fig.2. Interlace sequence sampling

2. PROPOSED NOVEL INTRA PREDICTION METHOD[1]

In this part, we proposed a novel intra prediction method, named Field-Among mode. The mode mainly includes two ways: one is samples selection, and other is prediction formula.

2.1 Samples Selection

The proposed mode mends the derivation process for neighboring location. In field macroblock pair (as shown in Fig 3.), when predict VS1 block, the proposed mode gets samples from VS0 block. Namely, the first row of VS1 block are predicted by the first and second row of VS0 block, the second row of VS1 block are predicted by the second and third row of VS0 block, ···, the rest may be deduced by analogy, when the last row of VS1 block are predicted by the last row of VS0 block.

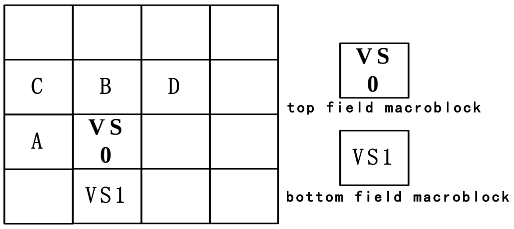


Fig.3. Field macroblock pair and neighboring macroblock

2.2 Prediction Formula

Based on samples section above, we adopted formula (1) to predict the VS1 block.

$$predMatrix[x,y]=(Pt\_v[x,y]+Pt\_v[x,y+1]+1)/2 \qquad (1)$$

Where x denotes horizontal direction, y denotes vertical direction, predMatrix[x,y] denotes predicted value of VS1 block and Pt_v[x,y] denotes reconstructed value of VS0 block.

* This project is supported by the National Natural Science Foundation of China (60472040).

[1] The method has applied patent (200610124585.3).

In addition, the last row value of VS1 block is equal to the last row value of VS0 block.

3. ALGORITHM DESCRIPTION

In according to research result above, for 4x4 luma block the cardinal procedure of Field-Among mode proposed is described as follow:

```
if (img->top_bot==1)
  for(j=0;j<BLOCK_SIZE;j++)
    for(i=0;i<BLOCK_SIZE;i++)
    {
      if(j==BLOCK_SIZE-1)
        img->mprr[field_among_pred][j][i]
              =enc_top_picture->imgY[img_y+j][img_x+i];
      else
        img->mprr[field_among_pred][j][i]
             =(enc_top_picture->imgY[img_y+j][img_x+i]+
         enc_top_picture->imgY[img_y+j+1][img_x+i]+1)>>1;
     }
```

Where img->top_bot==1 denotes that current block is bottom field block, enc_top_picture->imgY[img_y][img_x] denotes the upper-left reconstructed sample of top field block, img->mprr[field_among_pred][j][i] denotes predicted pixel of Field-Among mode to current block.

For the 16x16 luma block and 8x8 chroma block, the cardinal procedure is similar.

4. SIMULATION RESULTS AND ANALYSIS

In this section, the novel intra prediction algorithm was embedded in the H.264 JM11 encoder software and was simulated using the 100 frames testing sequences listed in Table 1. The key experimental parameters are as follow: IntraPeriod=1 (namely the coded frames are I frames), PicInterlace=1 and MbInterlace=1 (namely field coding is adopted).

Table 1. Video sequences used for analysis

Frame Format	Sequences	QP
cif(352x288)	News	21, 25, 28, 35
	Paris	29, 34, 36, 42
	Mobile	38, 40, 43, 48
	Tempete	35, 38, 40, 45

Simulation results are listed in Table 2. Figs 4 describe the rate-distortion for the four tested sequences used original JM11 and modified JM11 which added proposed Field-Among mode. It can be clearly seen that proposed algorithm can dramatically reduce bitrate whereas similar Peak Signal to Noise Ratio (PNSR) is still maintained. The percentage of bitrate decreases about 14.5% to 1.153% over original JM11.

The intra-predicted images using original JM11 and modified JM11 which added proposed mode for News sequence are shown in Fig 5. It can be seen that the predicted images by original and modified JM11 have very similar visual quality.

5. CONCLUSIONS

Based on the strong correlation between bottom block and top block for interlace video sequence, this paper proposed a novel intra prediction algorithm called Field-Among mode. Proposed algorithm adopts top block to predict bottom block, and further compress the spatial redundancy for I frame, so evidently improves the coding efficiency. Experimental results show that the decease percentage of bitrate is up to about by 14.5%, average decease percentage of bitrate is to about 9.94%, while maintaining the similar PNSR in field coding of full I frame.

Table 2. Bitrate and PNSR for different sequence and QP

sequences	Qp	Original JM11		Modified JM11		Compared results	
		Bitrate (kbits/s)	PSNR （dB）	Bitrate (kbits/s)	PSNR （dB）	△PSNR (db)	△Bitrate (%)
News	21	4860.94	43.16	4293.69	43.09	-0.07	-11.6696
	25	3577.89	40.42	3112.18	40.34	-0.08	-13.0163
	28	2759.78	38.1	2370.89	38.04	-0.06	-14.0913
	35	1451.16	32.7	1240.03	32.69	-0.01	-14.5491
Paris	29	4740.63	35.22	4263.79	35.14	0	-10.0586
	34	3120.51	31.17	2771.46	31.14	-0.01	-11.1857
	36	2513.5	29.49	2227.87	29.49	0	-11.3638
	42	1281.2	24.97	1168.1	25.05	-0.01	-8.82766
Mobile	38	3441.12	26.19	3089.78	26.16	-0.03	-10.21
	40	2807.56	24.67	2526.62	24.67	0	-10.0066
	43	2016.45	22.45	1846.35	22.5	0.05	-8.43562
	48	915.41	19.07	906.22	19.12	0.05	-1.00392
Tempete	35	2833.45	29.49	2474.87	29.44	-0.05	-12.655
	38	1941.25	27.12	1723.88	27.13	0.01	-11.197
	40	1497.96	25.77	1353.95	25.71	-0.06	-9.6137
	45	692.1	22.81	684.12	22.82	0.01	-1.153

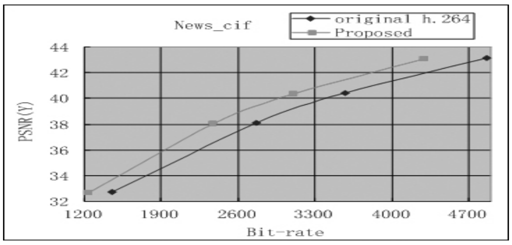


Fig.4. (a) Rate-Distortion for News sequence

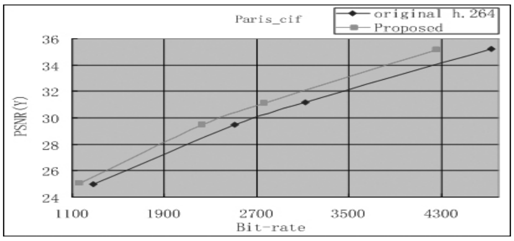


Fig.4. (b) Rate-Distortion for Paris sequence

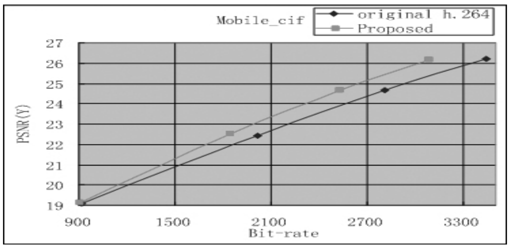


Fig.4. (c) Rate-Distortion for Mobile sequence

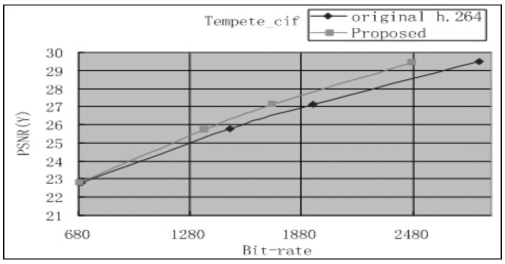
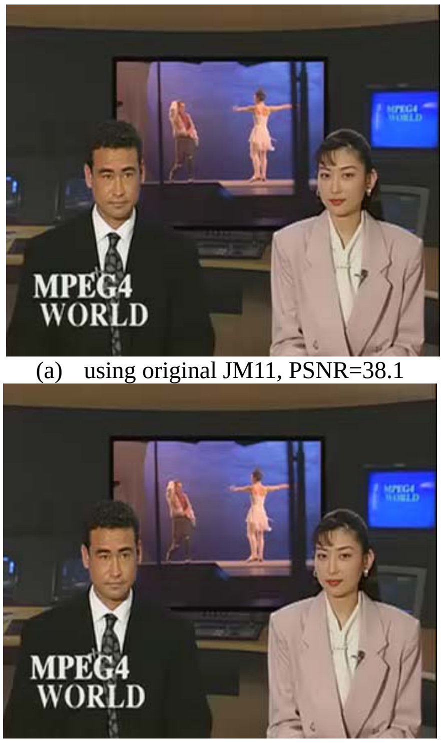


Fig.4. (d) Rate-Distortion for Tempete sequence



(a) using original JM11, PSNR=38.1

(b) using JM11 added proposed mode, PNSR=38.04

Fig.5. Intra-predicted images at Qp=28

6. REFERENCES

[1] Emerging H.26L Standard: Overview and TMS320C64x Digital Media Platform Implementation, White paper, UB video Inc., Feb 2002.

[2] Iain E. G. Richardson, “H.264/MPEG-4 Part 10 White Paper: Intra Prediction”, www.vcodex.com, May 2003.

[3] ITU-T Rec. H.264/ISO/IEC 11496-10, “Advanced video coding for generic audiovisual services”, *Telecommunication standardization sector of ITU* , March 2005.

[4] ITU-T Rec. H.264/ISO/IEC 11496-10, "Advanced Video Coding”, H.264/JVT Working Draft, Document WD2R8, April 2002.

[5] Zhen Han, Qiong Liu, Xuqing Xiao and Ruolin Ruan “Improvement of intra field-among prediction algorithm for AVS P2 X-profile”, *AVS Doc. AVS_M1922,* Dec 2006.

[6] Bojun Meng, Oscar C.Au, Chi-Wah Wong and Hong-Kwai Lam, “Efficient intra-prediction algorithm in H.264”, Image Processing, 2003. Proceedings. 2003 International Conference on, Sept. 2003.

XuQing Xiao was born in Hunan, China, in 1975. He is presently pursuing the doctorate in National Engineering Research for Multimedia Software, Wuhan University. His research interests include image coding, digital video compression and communication.

Research on the Model of Integrating Chinese Word Segmentation with Part-of-speech Tagging*

Xiaojun Tong[1] , Minggen Cui[1] , Guolong Song[2]
[1] School of Computer Science & Technology, Harbin Institute of Technology at Weihai, Weihai 264209, CHINA
[2] School of Information Science & Engineering, Northeastern University, Shenyang 110004, CHINA
E-mail: tong_xiaojun@163.com

ABSTRACT

In this paper, we presented a model of integrating Chinese word segmentation with part-of-speech tagging. In the early stage, we reserve the top N segmentation results as candidates. After Unknown words recognition and POS tagging, we choose the best result form the top N segmentation candidates by evaluating every one. We also implemented a Chinese lexical analyzer based on this model. The experiment results show that the overall accuracy of the proposed analyzer is 98.1% for segmentation and 95.1% for POS tagging respectively. The research is meaning for Chinese information processing and Chinese search engine based on web.

Keywords: Chinese Word Segmentation, Part-of-speech Tagging, N-shortest Paths Method

1. INTRODUCTION

Word is the independent and meaningful atom in natural language. Unlike English, there is no delimiter to mark word boundaries and no explicit definition of words in Chinese, therefore, Chinese lexical analysis is foundation and key of Chinese information processing[1,2]. In the lexical analysis of natural language based on Chinese characters, people used to segment Chinese word and tag POS respectively all the time. In fact, Chinese word segmentation and part-of-speech tagging are in close connection. More than 90% of segmentation ambiguity can be solved with grammatical level information, and those involved syntactic knowledge is few[3] . It is obvious that integrating Chinese word segmentation with part-of-speech tagging organically helps to dispel ambiguously and to improve whole efficiency.

This paper introduces N-shortest paths method to construct a model of integrating Chinese word segmentation with part-of-speech tagging, and we also implemented a Chinese lexical analyzer using this model. Through Unknown words recognition and POS tagging for the top N segmentation results, we get N candidates which are chosen from the best one. This method can integrate morphological level information with grammatical level information organically.

2. MODEL OF INTEGRATING CHINESE WORD SEGMENTATION WITH POS TAGGING

2.1 Description of The Question on Integrating Chinese Word Segmentation with POS Tagging

According to noisy channel model, we describe question on integrating Chinese word segmentation with POS tagging to be: a string of words tagged with POS $<W,T>=<w_1,t_1><w_2,t_2>\ldots<w_n,t_n>$,(Among them, $<w_i,t_i>$ represents the word w_i with the POS t_i), has passed a noisy channel and lost word boundaries and information of POS. In the end, it becomes a string of Chinese characters in output port. We should find $<W,T>$ corresponding of C, and choose the best result $<\mathrm{W,T}>^*$ with the greatest probability.

$$<\mathrm{W},\mathrm{T}>^* = \arg\max_{W,T} P(<\mathrm{W},\mathrm{T}>|C) \qquad (1)$$

2.2 Stochastic Model of Integrating Chinese Word Segmentation with POS Tagging

Eq.(1) describe model of integrating Chinese word segmentation with POS tagging. In order to integrate Chinese word segmentation with POS tagging and merge t morphological level information and grammatical level information as the evaluation basis, we introduces N-shortest paths method to get the top N results.

According to a string of Chinese characters C, we can get the top N results (W_1, W_2, $\ldots W_n$) with the greatest probability after segmentation. Then, we get word bunches ($<W_1,T_1>'$, $<W_2,T_2>'\ldots<W_n,T_n>'$) with a little information of POS after unknown words recognition, reserving those POS information. After POS tagging, word bunches ($<W_1,T_1>$, $<W,T_2>\ldots<W_n,T_n>$) with complete information of POS develops. In the end , we choose the most probable result. The processing shows in Fig 1.

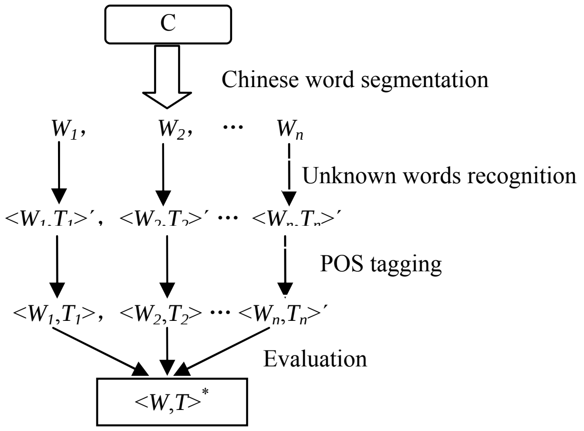

C
Chinese word segmentation
W1, W2, ··· Wn
Unknown words recognition
<W1,T1>′, <W2,T2>′ ··· <Wn,Tn>′
POS tagging
<W1,T1>, <W2,T2> ··· <Wn,Tn>′
Evaluation
<W,T>*

Fig.1. The model of integrating Chinese word segmentation with POS tagging

2.3 Model of Chinese Word Segmentation

According to noisy channel model, we imagine that a string of words become a string of Chinese characters, which have lost word boundaries because of the noise when it passed a noisy channel. So, the question on Chinese word segmentation can be described as looking for the string of words with the greatest probability.

This work was supported by finance in Shandong Province science and technology plan project of china through the grant number 2006110

$$W^{'} = \arg\max_{W} P(W \mid C) \quad (2)$$

On the basis of Baye’s Theorem, it can be induced that:

$$W^{'} = \arg\max_{W} P(W \mid C) = \arg\max_{W} \frac{P(W)P(C \mid W)}{P(C)} \quad (3)$$

In this formula, P(C) is the probability of a string of Chinese characters which is a constant, so it needn’t be considered. Because the word bunch corresponding with a Chinese characters bunch is one and only, P(C|W) is also a constant and is equal to one.

Therefore, above formula can be simplified as:

$$W^{'} = \arg\max_{W} P(W) \quad (4)$$

We apply Uni-Gram to acquire the word bunch probability, so

$$P(W) = \prod_{i=1}^{n} p(w_i) \quad (5)$$

In other words, the string of words with the greatest probability is the best [4,5,6].

2.4 Model of POS Tagging

In POS tagging, we try to find POS bunch $T^{'}$ with the greatest probability, according to a given word bunch W.

Therefore: $T^{'} = \arg\max P(T \mid W)$ (6)

On the basis of Baye’s Theorem, it can be induced that:

$$P(T \mid W) = \frac{P(T)P(W \mid T)}{P(W)} \quad (7)$$

In above formula, the denominator P(W) is the probability of the word bunch W, which is a constant, so the above formula can be simplified as:

$$P(T \mid W) = P(T)P(W \mid T) \quad (8)$$

We assume words are independent with each other, and the appearance of a word only depends on the POS of itself. Under the condition of the string T of POS, the probability of the word bunch W can be expressed with the product of the conditional probability of each word with the known POS.

$$P(W \mid T) \approx P(w_1 \mid t_1)P(w_2 \mid t_2)P(w_3 \mid t_3)\cdots P(w_n \mid t_n) \quad (9)$$

We assume that the probability of a POS depends on the one in front of it, then, the probability of T can be expressed with the product of probability of each item of it.

$$P(T) \approx P(t_1 \mid t_0)P(t_2 \mid t_1)P(t_3 \mid t_2)\cdots P(t_n \mid t_{n-1}) \quad (10)$$

Because t_0 is supposed, $P(t_1 \mid t_0)$ is equal to $P(t_1)$.

All the words, the statistic model of POS tagging can be described as:

$$T^{'} = \arg\max \prod_{i=1}^{n} P(t_i \mid t_{i-1})P(w_i \mid t_i) \quad (11)$$

We describe the question on POS tagging with Hidden Markov model. We consider the word bunch $W=w_1w_2\ldots w_n$ as output sequence that can be observed, consider the corresponding POS bunch as hidden state transfer sequence, consider the probability of the appearance of one word with one POS as the probability of lunching a sign when states transfer. During the processing of solving the problem, we use Viterbi algorithm. Viterbi algorithm has three steps: (1) Initializing; (2) Deducing; (3) Stopping and Reading the path [8].

(1) Initializing

The probability of fist word with state j (its POS is j) is

$$\delta_1(t^j) = P(t^j \mid w_1) \quad (12)$$

(2) Deducing

The probability of word i+1 with state j (its POS is j) is

$$\delta_{i+1}(t^j) = \max_{1\le k\le T}[\delta_i(t^k)\times P(w_{i+1} \mid t^j)\times P(t^j \mid t^k)], 1\le j\le T \quad (13)$$

When word i+1 is in the state j (It is tagged as POS j), the most probable state (POS) of the word i is:

$$\Psi_{i+1}(t^j) = \arg\max_{1\le k\le T}[\delta_i(t^k)\times P(w_{i+1} \mid t^j)\times P(t^j \mid t^k)], 1\le j\le T \quad (14)$$

(3) Stopping and Reading the path

Among them, $t_1,t_2\ldots,t_n$ is the POS bunch of the word bunch $w_1,w_2\ldots,w_n$.

$$t_n = \arg\max_{1\le j\le T} \delta_n(t^j) \quad (15)$$

$$t_i = \Psi_{i+1}(t_{i+1}), 1\le i\le n-1 \quad (16)$$

$$P(t_1,\cdots,t_n) = \max_{1\le j\le T} \delta_{n+1}(t^j) \quad (17)$$

2.5 Evaluation Function

The word bunch corresponding with a string of Chinese characters is one and only, so

$$P(C|W)=1\rightarrow P(CW)=P(W) \quad (18)$$

$$\begin{aligned} P(W,T|C) &= P(T|CW)P(W|C)=P(T|W)P(W|C) \\ &= P(T)P(W|T)/P(W)\times P(W)/P(C) \\ &= P(T)P(W|T)/P(C)=P(T)P(W|T) \quad (19) \end{aligned}$$

Apply HMM to expand $P(T)P(W|T)$, and introduce the co-occurrence probability now.

$$P(<W,T>|C) = \prod P(t_i \mid t_{i-1})P(w_i \mid t_i) \quad (20)$$

$$P^{*}(W,T) = \ln P(W,T) = \sum \ln P(t_i \mid t_{i-1}) + \sum \ln P(w_i \mid t_i) \quad (21)$$

Evaluation function is as following:

$$R^{\#} = \arg\max_{W,T}[\sum P(t_i \mid t_{i-1}) + \sum P(w_i \mid t_i)] \quad (22)$$

3. DESIGN AND IMPLEMENTATION OF SYSTEM OF INTEGRATING CHINESE WORD SEGMENTATION WITH POS TAGGING

This system of integrating Chinese word segmentation with POS tagging is made up of five parts: pretreatment module, Chinese word segmentation module, unknown words recognition module, POS tagging module, evaluation module. The systematic structure shows as following Fig 2.

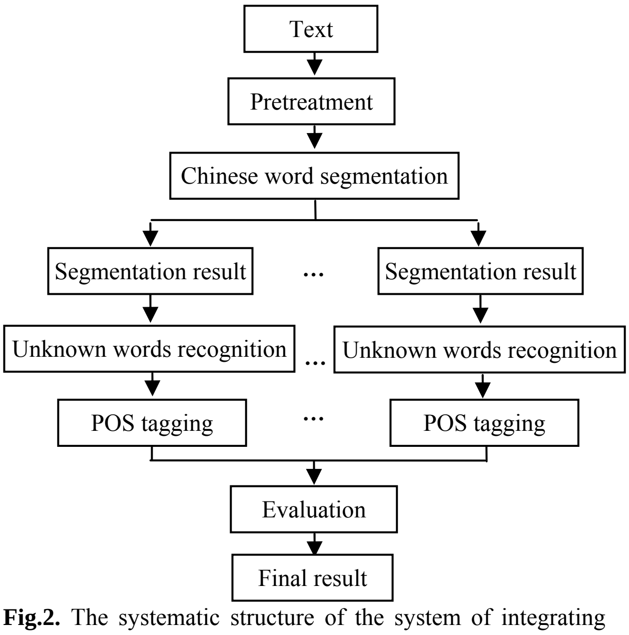


Fig.2. The systematic structure of the system of integrating Chinese word segmentation with POS tagging.

3.1 Pretreatment

In the mainland region, people generally use National Standard Code (GB2312-80).In this encoding system, Chinese character say with two bytes and the ASCII yard of each byte is greater than 127, which can distinguish Chinese characters with others.

In pretreatment, we need to scan the text twice. The first time, we scan the text to be dealt with so that we can withdraw sentences from the text according to the punctuation. The second time, we scan every sentence to distinguish Chinese characters with others, and recognize figures and English word bunch with the automatic machine. The string of Chinese characters will become the input of Chinese word segmentation module.

3.2 Chinese Word Segmentation and Unknown Words Recognition

In Chinese word segmentation, we need to find the top N word bunches (The Value of N can be set through user's interface according to the need). While calculating probability, because the probability of each word is a very little positive number (smaller than 1), the probability of a word bunch approaches zero, which is unable to be expressed on the computer. In order to solve this problem, we replace probability with the cost; the cost of word bunch is calculated according to the following formula:

$$Fee(W) = \sum_{i=1}^{n} -\log P(w_i) \tag{23}$$

When dealing, we scan Chinese character bunch from left to right, list all candidate words in order, and keep it in the array. Then we scan the candidates, use dynamic layout way to get the top N forerunners with minimum accumulated cost, and keep the information of the forerunners. If the present word is a terminal word, through rollback, we can get the top N word bunches with minimum cost (with the greatest probability).

Unknown words recognition is made up of Chinese personal name recognition and place name recognition, whose course is as Fig 3. Chinese personal name recognition is based on statistics method, using the sum of surname cost and name cost as its one. If the personal name cost of a bunch of Chinese characters is smaller than a given value, we consider it as a person name. Place name recognition is based on regular method. Through checking its suffix, we can recognize the place name. We keep the POS information of the recognized unknown words.

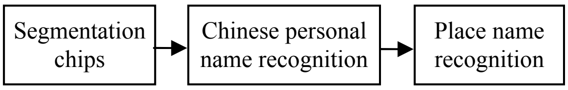


Fig.3. The processing of unknown words recognition

3.3 POS Tagging and Evaluation

POS tagging is used to deal with a word bunch making up a sentence, and the goal is to find POS bunch with the greatest probability. We add all POS and its cost of every word to its node (The POS information of unknown words is from unknown words recognition and others from the dictionary). Then, we scan the word bunch from left to right, calculate the cost of every word with its one POS and look for the most probably POS of the word in front. If the present word is a terminal word, through rollback, we can get the POS bunch of the word bunch.

After unknown words recognition and POS tagging, we get N words bunches with POS. We give a mark to each word bunch according to Eq. (22), the one with mnimum cost is the final result. In this module, we still recognize reduplicative words and merger part unknown words.

4. EXPERIMENT AND ANALYSIS

The system of integrating Chinese word segmentation with POS tagging uses the Beijing University collection of POS. There are a word form of 114,758 records, a POS form of 114,758 records and a name frequency form of 2,328 records in our dictionary. This paper tested this system openly with the articles including 5,221 words which are drawn from People's Daily in January of 1998. The results is as Table 1 shows.

Table 1. The test of system performance

The value of N	The total number of words	Chinese word segment		POS tagging	
		TWS	TA (%)	TWT	TA (%)
1	5221	5090	97.49	4920	94.23
2	5221	5109	97.85	4956	94.92
3	5221	5121	98.08	4964	95.07
4	5221	5123	98.12	4964	95.07

Note:

(1) TWS: the total number of words segmented right;
(2) TWT: the total number of words tagged right;
(3) TA: the accuracy

We can find from the data in the table: (1) The system performances well, whose accuracy of segmentation and POS tagging in opening tests both are higher than the accuracy of 96% and 94% in documents [9]. (2) After adopting N-shortest paths method, the accuracy rate of segmentation and POS tagging both are improved, which proves that it is suitable to keep a few word segmentation results. (3) The introduction of the POS information can improve the accuracy of Chinese word segmentation. (4) With the increase of N value, system performance is promoted . In the face of this question, considering of the operation efficiency, we choose 3 as the value of N.

5. CONCLUSIONS

This paper has applied N-shortest paths method to structure a model of integrating Chinese word segmentation with POS tagging, and implemented a Chinese analyzer based on this model. The tests show this system has high accuracy of Chinese word segmentation and POS tagging, which proves the method is effective.

Unknown words recognition and probability dictionary structure are important factors of influencing the system performance. It needs further study to handle these problems well.

ACKNOWLEDGEMENTS

This work was supported by finance in Shandong Province science and technology plan project of china through the grant number 2006110.

REFERENCES

[1] Zhang Hua-Ping, Liu Qun.，“Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Metho,”*Journal of Chinese information processing*,16(5).pp1-7, 2002

[2] ZHANG Hua-Ping, LIU Qun, Zhang Hao and Cheng XueQi “Automatic Recognition of Chinese Unknown Words Recognition,” *First SIGHAN Workshop attached with the 19th COLING*, pp.71-77, Aug.2002.

[3] He Ke-kang, Xu hui, Sun Bo, “Design principles of the expert system of written Chinese word automatic segmentation,”*Journal of Chinese information processing*,5(2),pp1-14 , 1991.

[4] Chen Xiao-He *Automatic Analysis of Contemporary Chinese*. Beijing Language & Culture University Press, pp127-130, 2000

[5] Yuan S. C, Henry T., Probability Theory, Springer-Verlag New York Inc., 1978 PP.324-338.

[6] Christopher D. Manning, Hinrich S., *Foundations of statistical natural language processing*, MIT press ,pp197-202, 1999

[7] Ma Qing,Isshara H, “Maoson SA Mnhi-neuro Tagger Applied in Chinese Texts”.,*Proceedings 1998 International Conference on Chinese Information Processing*，Beijing,1998-11-18/20:200.

[8] Shai Fine, Yoram Singer, and Naftali Tishby, “The hierarchical Hidden Markov Model: Analysis and applications,”*Machine Learning*, 32(1):41, 1998.

[9] Bai Shuan-Hu,*Method of Integrating Chinese word Applications（JSCL－95）*，1995，PP.56－61

Xiaojun Tong: A vice professor, birth 1963, reading Ph.D. at Harbin Institute Of Technology
Research direction: chaos cryptography, information security.
Communicate address: Harbin Institute Of Technology (Weihai) in school 20#506 room,
Weihai city, 264209, P.R. of China
E-mail: tong_xiaojun@163.com
Minggen Cui: Professor, the director of Ph.D candidate.

Efficient Image Retrieval in P2P Using Distributed TS-SOM and Relevance Feedback

Xianfu Meng, Changpeng Feng, Yingchun Wang
Department of Computer Science and Engineering, Dalian University of Technology
Dalian 116024, China
Email: chpfeng@gmail.com

ABSTRACT

Nowadays, features of an image in most image retrieval systems in centralized networks are represented by vector whose dimension will reach up to tens or even hundreds easily. Therefore, a wide range of researches has being triggered to solve the above problems. We propose a strategy to combine DHT and TS-SOM to attack the Curse of Dimension as well as application in P2P environment. In addition, strategy of multiple sample images is supported in this system to change the situation that single sample image cannot sufficiently reflect the users' aim. In the end, a method of changing the weight of every feature dynamically is employed by this system, which is supposed to improve the latency of feedback efficiently. Final evaluation shows that this system outperforms the referential retrieval system: PicSOM in retrieval accuracy and load distribution. Theoretically proves that it also is a promising system especially in computing resources saving.

Keywords: CBIR, P2P, DHT, TS-SOM, Relevance Feedback

1. INTRODUCTION

In the past decade, content-based image retrieval (CBIR) has attracted extensive research interest. The color histogram method introduced by Swain and Ballard has shown to be very effective and simple to implement. However, the main disadvantage of the color histogram method is that it is not robust to significant appearance changes because it does not include any spatial information. Similarly, querying images based on single feature, say shape only or texture only, cannot get satisfactory results. Recently, lots of researchers are doing related research about that and a lot of methods are being proposed. For instance, [1] was implemented for two different similarity measure, Euclidean distance based and human perception based. [12] proposed a soft rule mining algorithm to infer image relevance from the collective feedback. Besides, relevance feedback also attracts a large number of researchers concern. As by adopting this method, it does improve the performance of the system significantly.

Recently, many famous CBIR systems emerges, like QBIC [IBM], MARS [UIUC], photoBook [MIT], DISCOVER [Hong Kong], all of which operate only in high dimension space that will require large amount of computing time and memory storage. By adopting Distributed TS-SOM, which makes the best of Tree structure and Self-Organized Maps, saves the computation resources remarkably.

The reminder of this paper is organized as follows. In the following two sections, we describe related work and feature integration respectively. Section 4 and section 5 present in details about Distributed Tree Structured Self-Organized Maps and Relevance Feedback. We report and analyze our system in section 6 and give final remarks and conclusions in section 7.

2. RELATED WORK

During the query of Content Based Image Retrieval (CBIR), the features of an image can be represented by vector, whose dimensions will reach up to tens or even hundreds easily. Experimental results show that, when the dimensions of a vector exceeds 20, then the existing index methods become unstable, which is so-called Curse of Dimension [2]. So this problem raises wide researcher's concern. Till now, a lot of methods about image indexing have popped up accordingly.

How to partition the data space and manage data according to the partitioning method is the key point of indexing strategy [2]. The category of R tree includes R tree as well as some other AVL trees evolve from R tree, say, R+ tree, R* tree, SR tree, just to name a few. Beside R tree category, there are varieties of other index structures, for instance, K-D tree category, Quad tree category, VA- File, etc.

R Tree
R tree is a kind of index structure with good performance. Overlaps are allowed among MBRs, which guarantees the usage factor of space over 50%. But in the case of high dimensional, the indexing times and storage space will increase sharply due to freely overlap, which degrades the query performance significantly [3,5].

Self-Organizing Maps (SOM) and Tree Structured –SOM
A SOM consists of a (usually two-dimensional) regular lattice or grid of map units. The most common grid type is probably the hexagonal grid but a more natural choice with images is to use a rectangular grid. To speed up the best-matching unit (BMU) search, Koikkalainen and Oja introduced a variant of SOM called the Tree-Structured Self-Organizing Map (TS-SOM)[13]. TS-SOM is a tree-structured vector quantitative algorithm that uses normal SOMs at each of its hierarchical levels. The search space for the BMU on the underlying SOM level is restricted to a fixed-sized portion just below the BMU on the above SOM.. Denote C(A) the union of the children of a set of map units A and $N_u(m)$ the neighboring units of a given map unit m, and suppose the BMU at the current level is b. Then the candidate set for searching the BMU at the next level is

$$G(b) = C(\{b\}) \bigcup C(N_u(b)) \qquad \text{Eq.(1)}$$

The above strategy guarantees high probability for finding correct BMUs while reduces the searching time from O(M) to O(logM), which takes advantage of the essence of tree structure, where M is the size of the SOM. This saving in computation by using hierarchical structure facilitates the creation and use of large SOMs and is beneficial to indexing huge image repository.

3. FEATURE INTEGRATION

Automatic analysis of image content is a challenging problem. The ability to extract and describe objects in a complex scene is crucial for image understanding. Another problem facing the researchers is automatic segmentation, which barriers the integration of various methodologies. In practice, there is a trend of using information obtained from multiple cues such as color, structure and texture to do image retrieval [4].

We use a linear combination of the distances in the product space of structure, color and texture for retrieval. Distances are properly normalized to take into account the difference in image size, and to make sure that the histogram intersection measure is symmetric. Weights are associated with distances, which assign the degree of importance attached to features extracted from different methodologies. The distances in these spaces were pre-normalized in the range [0,1]. However, it may be possible that a relatively larger value in a feature space biases the calculation of the weighted distance. To overcome this problem, we have used the following Gaussian normalization that puts equal emphasis on the distances in each of the three feature spaces, before taking a linear combination. Let d = (d_i) be a sequence of distances in any of the above-mentioned three feature spaces. Gaussian normalization results in the mapping: $d_i \rightarrow (d-\mu)/3\sigma$. Where μ and σ represent the mean and the standard deviation of d_i.
Let $d_i = (d_i - \mu)/3\sigma$. This normalization ensures that probability of the normalized value, d_i, being in the range [-1,1], is 99%. Values outside this range may be forced to map to either -1 or 1. Finally, the mapping $d_i \rightarrow (d_i + 1)/2$ normalizes distances in [0,1].

4. DISTRIBUTED TS-SOM

4.1 DHT

Structured P2P system in the form of Distributed Hash Tables (DHT) is a promising approach for building data management platforms for such scenarios. But soon the database community has started to adapt techniques from distributed and parallel database systems to DHTs. DHT are able to cope with very high numbers of parallel transactions that process huge sets of (key, value)-pairs. DHT follow the p2p paradigm, i.e. they consist of many autonomous nodes, there is no central coordinator and global knowledge is not available. Examples of DHT are CAN, Chord, Pastry or P-Grid. The proposals mainly differ in the topic of the key space and the contract selection, i.e., how to distribute the (key, value)-pairs among the peers and which are the nodes a peer communicates with [6, 7, 8].

4.2 Distributed TS-SOM

Image retrieval, in a sense, has to do with the user's interest. Every one interested in one or some type of images will look them up much more often than images from other categories. TS-SOM has the ability of clustering and it can classify the images into categories, which facilitate the user's query and save much computation resources. So we combine DHT and TS-SOM.

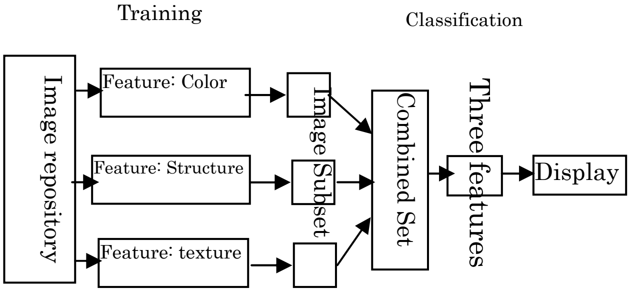


Fig.1. Two-stage structure of TS-SOM with 3 features

Before that, we give an overview on how different data structures are used with TS-SOM as an supplement of Section 2. More details refer to Fig.1. The first operation is training. It builds and organizes a TS-SOM structure, which is stored in the name space with a given name. The second important operation is SOM classification. This operation divides the data records of a given data frame into subgroups according to located BMU (neuron). The result is a classified data which includes as many classes as there are neurons in the TS-SOM. Like a weight matrix, one class corresponds to one neuron and includes the indexes of the data records classified to that neuron. If the neuron does not get any data record in the classification, then an empty class is created corresponding to that neuron.

Combining DHT and TS-SOM

In this paper we point out that the SOM mapping can also be used as a hash coding, a scheme for providing rapid access to data items which are distinguished by some key, and it is able to implement the index of the multi-dimensional feature vectors by a distributed Self-Organizing Map. We named the new indexing structure Distributed and Tree-Structured SOM (DTS-SOM) since it evolves from TS-SOM. Like many exiting DHT designs, all participating nodes in the proposed system work together to maintain the whole index. Each node is responsible for a part of the index and maintains its neighborhood at different levels, which save a great amount of storage.

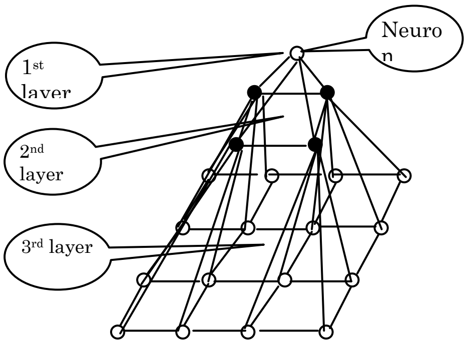


Fig.2. Architecture of TS-SOM with 3 layers

The SOM mapping is by nature a trade-off, which depends on the size of SOM, between clustering and topology ordering. The above facts motivate us to make the most of the SOM approximation on the 2-D lattice. Many operations, such as looking for neighbors or the low-pass filtering, in the query procedure can be approximated on the 2-Dlattice rather than in the high-dimensional feature space if the distributed index is well consistent. Thus we employed large maps during a query because only the score, i.e. a float number, is stored for each position of specific map unit. This saving is especially important when the query involves multiple features.

4.3 Routing

The data peer who owns the best matching unit is called the Best Matching Data Peer (BMDP). Routing or finding BMDP is a very important subroutine in image publishing and querying. There are two types of routing procedures for this purpose in our design: routing by vector and routing by position.

The first procedure takes a vector as the input and returns the node that owns the BMU of the input vector. This procedure is similar to the counterpart of the centralized TS-SOM, except replacing the BMU candidate set with the following one:

$$G'(b,D) = C(\{b\}) \bigcup [C(N_u(b)) \bigcap c(N_u(D))] \qquad \text{Eq(2)}$$

Where D is the data peer who owns the BMU b at the current level. The additional intersection restricts the communication within the direct neighborhood. In analogous to the centralized version, finding the BMU in a DTS-SOM starts from the top level and then downwards the hierarchy. In DTS-SOM training, the BMDPs at all levels are returned.

During queries, only the BMDP at the bottom level are returned. The input of the second procedure is the position of a map unit at the bottom level and the output is the data peer who owns that unit. The routing of this type is a special case of the first type. It also iterates from the top level to the bottom, but the proximity between map units is approximated by 2-D grid distance. This saves a substantial amount of computation compared with the using Euclidean distance in the high-dimensional feature space.

The above procedures run within a working group. If more than one working group is available, the routing request is also forward to other groups through friend links. Take multi-features query for example, the data information about weights and scores will be transferred among TS-SOMs. If the working group is specified, the routing procedures return the result from that group. Otherwise, they end up with the first returned BMDP and discard the others.

The procedure can also be described by the following pseudo-code:

```
Find_win_path(x(t),root)          //x(t) denotes inputting at
//time t; root denotes root neuron.
Begin
    Initialize k;          //which is used as path search factor.
    For i : 1 to L   initialize $\lambda_i$.  //initialize the level of
//emphasis at each layer.
    V=root; $e_v$=||x(t)-$W_v$||;          //compute error at root.
    Vi=(v, $e_v$);
    (v, $e_v$ )=Find_win_child(x(t),l,$V_l$);
   // find winning children recursively until a winning leaf is found.
    Win_leaf=v;
     //assign the win_leaf as neuron_error pair v returned by
//Find_win_child
     Win_path=Trace_parent(v); //trace the winning path from
//the winning leaf to root.
End.

Find_win_child(x(t),l,$V_i$)
//l denotes the layer number; Vi denotes set of neuron-error
//pairs, (v, $e_v$) at l-th layer
Begin
   If(i=L) return first element of $V_i$;   //return winning leaf
//neuron.
   Initialize $V_{i+1} = \phi$;    //begin with an empty set of winning
//neurons
For each (v, $e_v$) $\in$ $V_i$
        $V_{i+1}$= $V_{i+1}$ $\bigcup$ Compute_error(x(t),i,v,$e_v$); //compute
//errors for all children of node v
Delete all except k neuron-error pairs with smallest error values
in $V_{i+1}$
     Return Find_win_child(x(t),i+1, $V_{i+1}$);
End

Compute_error(x(t),l,v,$e_v$)          // $e_v$  denotes error of v.
Compute  $e_{u_j} = \| x(t) - w_{u_j}(t) \| + \lambda_i e_v \quad \forall u_j \in child(v)$
Begin
Return $\bigcup_{j=1}^{n} (u_j, e_{u_j})$
End.
```

4.4 Query Processing

We classify the nodes in the network into 3 categories to facilitate our following description, see Fig.3:

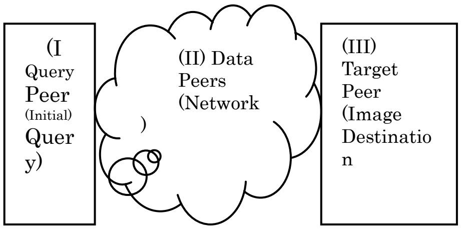

(I Query Peer (Initial) Query)
(II) Data Peers (Network)
(III) Target Peer (Image Destination

Fig.3. The classifications of nodes

Suppose k images are displayed in each round, the interactive CBIR query procedure in distributed context is as follows:

1. The query peer will extract its feature vectors and route to the respective BMDPs. k images are then randomly retrieved from those data peers.
2. The user marks the relevant images. He/she will exit the system in the case of enough relevant images are found or the user is tired.
3. The relevance feedback forms positive and negative impulses on the 2-D lattice and spread to their neighbors by the low-pass filtering. The list of positions with highest scores in each feature map is obtained. Next, route to the respective data peers for each position in the list, and retrieve the new image candidates from individual features.
4. For each candidate image from a feature map, locate its BMUs in the other maps. Scores of each candidate image are then represented by the ones of its BMUs in the individual feature maps, and the total score is obtained by summing up the individual scores. Display the candidate images of top k total scores. Go to step2.

4.5 Varying Sample Images

A lot of CBIR systems support sample image retrieval but most of them allow only one sample image. So we proposed the multiple sample images strategy to improve the situation.

4.5.1 multiple sample images

As the significant gap between high-level semantic concepts and low-level visual features of an image, it is difficult to depict an image accurately, in particular, to list out sample images to do retrieval. It is easily to understand that one image is a little similar to that one in color or in structure, or in texture, which crosses refers to section 5 relevance feedback. Obviously, it has to do with semantic analysis which still is a great challenge to image retrieval. For reasons of brevity, we compute the similarity of the images by the following equation:

$$D(X_j,S) = \min_k d(X_j,S_k) \qquad \text{Eq(3)}$$

Where, X_j denotes the J_{th} image in the image repository X in a certain peer, S denote a set of query images provided by the user. D represents the distance of the image X_j to the set of images S, and d is the distance of X_j from an image S_k, which is contained in S. The equation above essentially defines the distance of a candidate image and its nearest neighbor in the query set.

4.6 Maintenance

The periodically updating, intrinsic ability of TS-SOM, provides self-balancing algorithm for the index, which simplifies the process of joining and leaving.

4.6.1 Joining

When a node D joins the network, first it connects to one of the participating nodes D0 in the network by some external knowledge and sends a random working group ID as well as a random position p to D0. D0 then routes the joining request to the BMDP D' of P in the specified working group. D' forwards the joining request to its neighboring node N_n(D'). For each $D_i \in N_n$ (D'), D_i compares p and its own center point. If the condition of neighbor-ship holds, then D_i and D add each other to their sets of neighboring nodes. So do D and D'. Later the center points and related index elements are updated in the self-organization manner.

4.6.2 Leaving

When a node is about to leave, it sends a notification to each $D_i \in N_n$ (D). D_i gets the information and simply removes D from Nn (D_i). The charging regions and index sub-trees will later be updated by the periodical self-balancing communications. If D_i receives no response from D within a predefined time threshold, D_i will deem D as inelegantly left, namely node failure. In this case the indexing part of D, including the image links and the model vectors has been lost in its working group. D_i has to recover such information by retrieving the counterparts in other working groups. So we found that the whole network can work through node failures smoothly and without any downtime.

5. RELEVANCE FEEDBACK

5.1 Multiple Level Feedbacks

Traditional relevance feedback framework requests the user to announce the retrieved image as either "extremely excellent" or "exactly irrelevant", which does not take images that fall into the medium into consideration. So Rui et al.[14] proposed five relevance levels to better capture the user's perception subjectivity about the image relevance, namely, '"highly relevant", "relevant", "no opinion", "irrelevant" and "highly irrelevant" for relevance feedback annotation. Experimentally [11], we found that although it is more convenient to express the user's subjectivity about relevance by "highly relevant", "relevant" and "no opinion", it is difficult for the user to discriminate "irrelevant" and "highly irrelevant" retrieved images. Since the image repertory may have extremely diverse content in general CBIR systems. It is hard to say which image is more "irrelevant" if the image does not contain the information that the user seeks for. Hence, in our system, we provide the user with four levels of relevance feedback, say: good, fair, not care or bad. Each of them has a corresponding weight: 0.5, 0.1, 0, -0.1 for the above four levels respectively.

5.2 Changing the Weight of Features Dynamically

When the candidate images are returned. We will mark each feature of an image with a certain score. As some of the images may be very similar to the target image in color, while some may be similar in structure, and some in texture probably. Then the system knows why this image is good, or fair, or not care, or bad by completing above tasks. Note that, the image(s) that do not get scores will be discarded implicitly; it won't be involved into the next query round. Any feature of an image has a default weight, for instance, 0 for color, structure and texture respectively. In case of part of features of an image are marked, the default value of other features will be taken by the system to refine the query vector. Next, the feedback results will be collected to refine the query vector and adjust the weight of every feature. So the weight of every feature is adjusted after each query round, and user's opinion can be used to refine the results timely.

6. EVALUATION

Theoretical analysis:

If 4-byte float numbers are used for representing the scores and the query includes, say 3 different features (color, structure and texture) of a sample image, supposed 3 sample images are used in the query, the memory space required for a 64×64 score field is 64×64×4×3×3 =144 Kbytes, which can easily be handled by most contemporary computing devices. The respective amount for the 256×256 setting is about 2.3 Mbytes, which is also acceptable for most personal computers. On the other hand, a DTS-SOM whose bottom level is of size 256×256 can support the image repository that contains up to several million of images. This is sufficient for most applications of special interests.

Multi-sample images can reflect the user's intention much better than a single image, since it can represent the target image from different aspects. It is true that extracting the feature vectors of multi-sample is more resource consuming than dealing with a single image, but empirically data shows it is more cost-effective in the long run. Since the feature extraction takes place only once, and it will be more time-consuming and increase the burden to the whole P2P network when query is re-continued.

Experimental results:

Our experiments were carried out on basis of Peersim [15] in a PC (Pentium 4 2.8G, 1G RAM, Windows XP) The experimental results were compared with that of PicSOM's which was developed by Jorma Laaksonen et al [16]. Because of conditions limit, we manually collected 2078 images in total, which were classified into three categories: natural scene, manmade object and persons. During simulation, the number of nodes was set as a fixed number, 100, and the images were distributed into 100 files according to standard distribution.

To measure the retrieval performance, we have applied a quantitative figure denoted by Γ measure. For obtaining the Γ value, it is assumed that the user is searching from an image repository D for an image labeled *l* belonging to an image class $C \subset D$. Before the target image is found, the user guides the search by marking all shown images, which belongs to class C as relevant. This process is then repeated for each image in C. Now, the Γ measure equals the average number Nc of images the system retrievals before the target one is found.

Γ=Nc/N Eq(4)

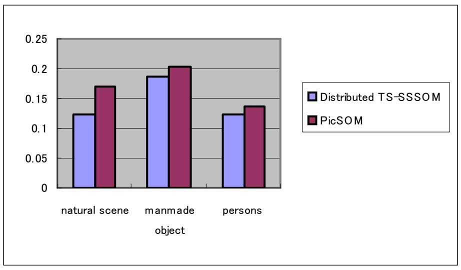


Fig.4. Comparison of Distributed TS-SOM,using 3 sample images, with PicSOM.

Where N denotes the total number of images.

Data in Fig.4 proves that Distributed TS-SOM with interactive relevance feedback is a promising method and outperforms PicSOM as a whole. Retrieval of images from person category is superior to that in PicSOM by 9.6%; regarding retrieval of manmade object images, 9.1%; and extraordinary superiority occurred for retrieval natural scene images, say 37.1%. It is easily to understand the color takes higher proportion in most natural scene images, and it also proves to be the most direct viewing and most sensitive feature to human beings [9]. Furthermore, the weights of feature vectors can be adjusted timely by incorporating the interactive relevance feedback.

The second experiment was carried to measure the recall of the system, which can also be regarded as an variance of Γ, with different times of feedbacks, where TTL was set as 3. Results were shown in Fig.5.

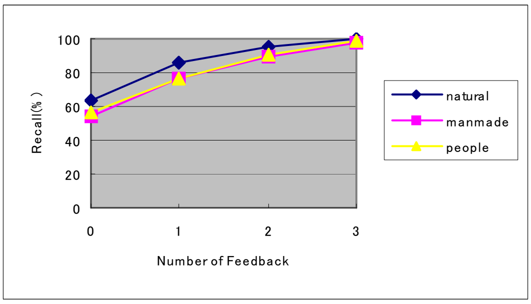


Fig.5. The recall result

The third Experiment aimed at measuring the load distribution, say the number of node searched during a query. The more the node searched, the more overhead it put to the network. DHT support exact search quite well, given a query, it can get the node easily and with fewer overheads. Furthermore, TS-SOM classifies the images into groups, which can also facilitates the query. The data in Fig.6 proves our opinion. The number of searched nodes was kept at a low level. Note that, the number of nodes shows in Fig.6 means the distinct node during a query, that is, one node was counted once no matter it is searched two or more times (due to feedback).

It is worth pointing out that, during experiments, the number of searched nodes became "huge" for several times. Since images were distributed into files according to Standard Distribution. It is true that some images falls into rare files with certain probability. It is not a mistake we found, actually, it proves our experiments performs perfectly well.

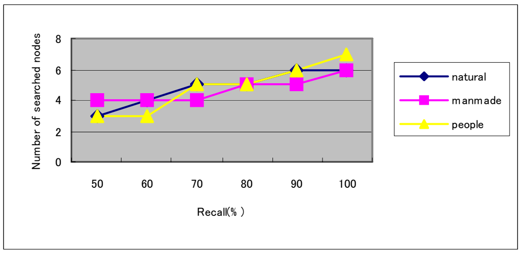


Fig.6. Load distribution

The results above does not suffice to support us to conclude the system performs better than most of traditional ones, because the scale of simulation, say the number of nodes and images used in the experiments was rather smaller than that in the real network. More experiments needs to be done to prove our conclusion.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we took full advantage of Tree Structured-Self Organized Maps (TS-SOM) and Distributed Hash Table (DHT) and combined them together to build index for images, which saved a lot of computation resources. Multiple features were integrated to perform query, say color, structure and texture constitute the query vector with individual weight. In course of relevance feedback, our proposed method can change the weight of each feature dynamically according to user's feedback. Evaluation results demonstrate it is a promising strategy and outperforms most of other methodologies on certain aspects. More future work, for instance, how to incorporate the Top-K method into this system, which is supposed to save much network overhead [10], since the system will make good use of every returned result. Another aspect needs more consideration is the failure tolerance with concurrent node failures.

REFERENCES

[1] Sanjoy K. Saha , Amit K. Das , "Bhabatosh Chanda, Image retrieval based on indexing and relevance feedback," *Pattern Recognition Letters* (2006).

[2] QU Ji - lin , KOU Ji - song , LI Min – qiang, "Research on the Indexing Technique in Content Based Image Retrieval," Vol. 24 , No. April, 2006.

[3] Norbert Beckmann, Hans-Peter Kengel ,Ralf Schneider, Bernhard Seeger, "An Efficient and Robust Access Method for Points and Rectangles," *ACM 089791365* /!90/0@35/0322

[4] Qasim Iqbal, J.K. Aggarwal, "Feature Integration, Multi-image Queries and Relevance Feedback" in *Image Retrieval.6th International Conference on Visual Information Systems (VISUAL2003)*, Miami, Florida, Sep. 24-26, 2003,pp. 467-474th Heldin conjunction with9 International Conference on Distributed Multimedia Systems (DMS'03).

[5] http://www.cnweblog.com/hengfei/archive/2006/01/03/62194.aspx.

[6] Indranil Gupta, Ken Birman, Prakash Linga, Al Demers, Robbert van Renesse, Kelips: "Building an Efficient and Stable P2P DHT through Increased Memory and Background Overhead".

[7] P.A. Felber, E.W. Biersack, L.Garces-Erice, K.W. Ross, G.Urvoy-KellerData, *Indexing and Querying in DHT Peer-to-Peer Networks*.

[8] XIE Yao, LI Jin-sheng, HONG Pei-lin, HiMCAN: *a Novel DHT Based P2P CAN Network*, TP393.04.

[9] SHI Jun, CHANG Yi-lin, "Overview of image retrieval," *JOURNAL OF XIDIAN UNIVERSITY*, Aug.2003 Vol.30 No.4.

[10] HE Ying-Jie, WANG Shan, DU Xiao-Yong, "Efficient Top-k Query Processing in Pure Peer-to-Peer Network, "*Journal of Software*, 2005, 16(4):540 552. DOI: 10.1360/jos 160540.

[11] Sam Y. Sung, Tianming Hu, "Iconic pictorial retrieval using multiple attributes and spatial relationships," *Knowledge-Based Systems* (2006).

[12] Peng-Yeng Yin, Shin-Huei Li, "Content-based image retrieval using association rule mining with soft relevance feedback," J. Vis. *Commun. Image R. 17* (2006)

1108-1125.

[13] P.Koikkalainen. “Progress with the tree-structured self-organizing map”. In A. G. Cohn, editor, *11th European Conference on Artificial Intelligence*, pages 211-215. European Committee for Artificial Intelligence (EC-CAI), John Wiley & Sons, Ltd.

[14] Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval,” *IEEE Trans*. Circ. Syst. Video Technol. 8 (5) (1998) 644-655.

[15] http://peersim.sourceforge.net/.

[16] Jorma Laaksonen, Markus Koskela, Sami Laakso and Erkki Oja, “Self-Organising Maps as a Relevance Feedback Technique in Content-Based Image Retrieval,” *Pattern Analysis & Applications* (2001)4:140–152 . 2001 Springer-Verlag London Limited.

Feature Selection Based on the Circle Window in Image Classification*

Xiang Zhang[1,2] Xiaoling Xiao[1,3]
[1] Yangtze University, Jingzhou, Hubei, 434023, China;
[2] Key Laboratory of Exploration Technologies for Oil and Gas Resources, Ministry of Education, Jingzhou, 434023
[3] School of Computer Science and Technology, Wuhan University of Technology,
Wuhan, Hubei, 430063
Email: zx_jr_xl@163.com

ABSTRACT

A new method of feature selection for support vector machines in pixel classification is presented. Since outliers always exist in the margin of the square window, the circle window is proposed to reduce the effects from the outliers in feature selection. At each location in 2D images, we use gray intensity features based on the circle window as inputs to the SVM classifier. The comparative experiments are carried out with different window shapes and sizes in terms of the classification precision and the computation cost. Comparative experiments show the proposed features obtain the better classification precision and the lower computation costs than other features using multi-class SVM methods.

Keywords: Feature Selection, Window Shape, Outlier, Classification, Support Vector Machine

1. INTRODUCTION

Support Vector Machine (SVM) is the most recent classifier in machine learning which was proposed by Vapnik and is based on The Statistical Learning Theory[1]. In the classification case, the key feature of the SVM is that its target function attempts to minimize the number of errors made on the training set while simultaneously maximizing the ‘margin’ between the individual classes. This is an effective ‘prior’ for avoiding over-fitting, which leads to good generalization.

In addition to performance of classifiers, the quality of image classification is highly dependent on the quality of the features used to describe the image. For a SVM classifier to work properly, feature selection plays a crucial role. Putting together too many irrelevant features only degrades the performance of classifier. The optimal selection of features is important since it maximizes classes contrast differentiation in feature space, while minimizes the computational complexity when they are used. We want to find out the most important features of image data to avoid overfitting. Many classification approaches reported in the literature simply use the gray intensity features of the pixels [2][3]. Lssam, et al. proposes feature extraction based on gray intensity features of region pixels for support vector machine classifier [4]. In addition to the current pixel intensity, the pixel intensities in the neighborhood were used as additional features; for example, each pixel forms with its eight nearest neighbors a 9-element features vector. Feature extraction based on region pixel intensities possesses the ability of robust classification and noise elimination compared with the chosen current pixel intensity, and the less computational complexity compared with the texture extraction. However, it is difficult to determine the size of the window used in extracting gray intensity features. Too small a sub-image region will not have enough information to separate classes, while a large sub-image region are more likely to include different types of feature information, specially for medical images.

In this paper, we present an approach in which the circle window is used to reduce the effect of mixture of different types of feature information. At each location in 2D images, we use gray intensity features in the circle windows as inputs to the SVM classifier. Experiments show that the proposed method could solve the problem of the effect suffered from the neighborhood classes in pixel classification.

2. THE CIRCLE WINDOW FOR FEATURE SELECTION

Choosing the SVM approach as the classifier for pixel classification, we need to specify which information should be provided as input. Many classification approaches simply use the gray intensity features of the pixels in the window. The square window is always used when features of image region are extracted. However, it is a very key problem to determine the size of the input window, because the size of square window is heavily related with the precision of classification. A small window region will not have enough information to separate classes, while a large window region are more likely to include feature information from other different classes. For example, the information is incomplete when the size of square window is 5×5, while the information is the mixture of information from different classes when the size of square window is 7×7. The optimal size of square window is between 5×5 and 7×7, but the size of window is odd. Since outliers always exist in the margin of the square window, feature extraction based on the circle window can reduce the number of the outliers. The radius of the circle window is defined by the block distance between the maximal row and the maximal column based on the current pixel as the origin position, as following:

$$r = |i_{\max}| + |j_{\max}| \qquad （1）$$

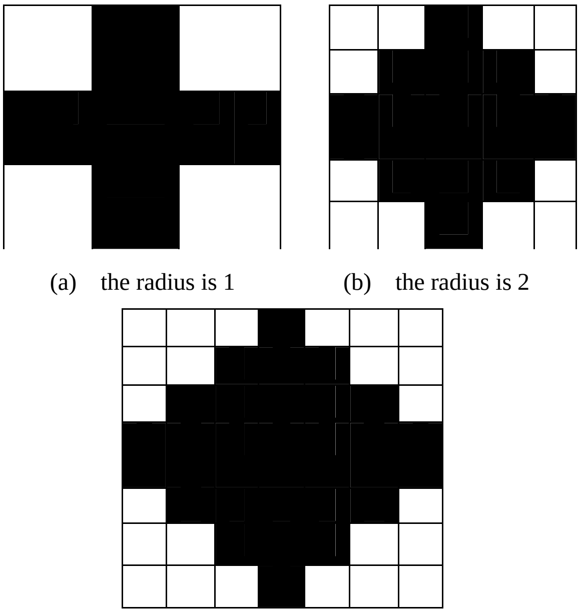
Where, r is the radius of the circle window, $i_{\max}$, $j_{\max}$ is the maximal row and the maximal column coordinate, respectively. According to the definition of the radius of circle window in Eq. 1, the relation between the radius of circle window and the length of square window is defined as following:

$$r = \mathrm{int}(w/2) \qquad （2）$$

Where, w is the length of square window.

Fig.1 show a comparison between the circle window and the square one with window size 3×3 , 5×5 and 7×7 , the range of the circle window is shown in black block that is defined by the corresponding radius 1,2 and 3. From the Fig. 1, we can see that feature extraction based on the circle window is helpful for removing the effect of outliers.

* This work was funded under Project 60673189 supported by the National Natural Science Foundation of China, Project 2004ABA043 supported by the Natural Science Foundation of Hubei Province and Project D200612002 supported by the Key Scientific Research Project of Hubei Education Department.



(a) the radius is 1 (b) the radius is 2

(c) the radius is 3

Fig.1. A comparison between the square window and the circle one

SCALING

Scaling each feature before applying SVM is very important. The main advantage is to avoid features in bigger numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation, for kernel values usually depend on the inner products of feature vectors, e.g. the linear kernel and the polynomial kernel, large feature values might cause numerical problems. So it is necessary to scale features before using them. The scaling processing is made in the following equation.

$$F_a = \frac{2 * F - F_{\max} - F_{\min}}{F_{\max} - F_{\min}} \qquad （3）$$

Where, F and F_a is feature before and after the scaling processing respectively, $F_{\max}$ and $F_{\min}$ is the maximum value and the minimum value of feature respectively.

It is recommended to linearly scale each feature to the range [-1, +1]. Of course we have to use the same method to scale testing data before testing.

3. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed features, the comparative experiments, in which multi-class SVM method is performed to classify the brain tissue, are made using features based on both the circle window and the square window in different size of window. The simulated MR images available from the BrainWeb [5] were experimented. The size of image volume is 181×217 pixel^2. At each location in 2D MR images, we applied SVM classifiers with a Gaussian RBF kernel $\sigma^2 = 0.5$ and c=1000 to determine which one of the three brain tissues (WM, GM and CSF) and the background the pixel belongs to. For the training data sets, each class is given a respective label (0: background, 1:WM, 2:GM, 3:CSF). In the training and testing of the SVM classifier, the gray intensity features of the input image was normalized to [-1,1]. The training data sets were obtained by randomly selecting 1000 pixels from a 2D image.

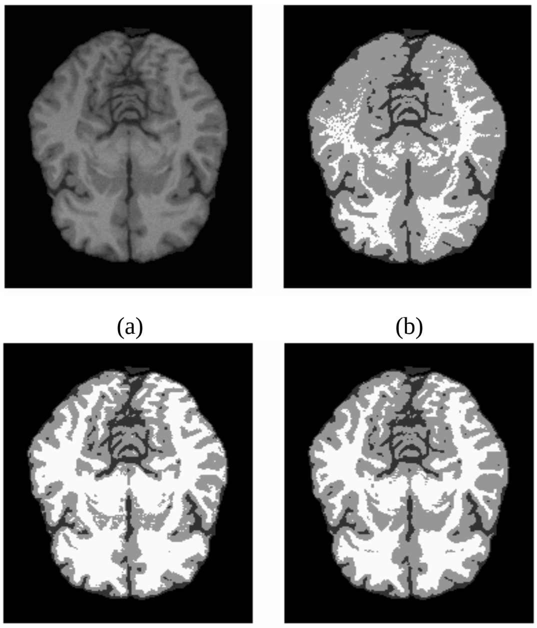
Classification Precision Comparison

The comparative results based on features of both the square window at the different window sizes 3×3, 5×5 and 7×7 and the corresponding circle window at different radius 1,2 and 3 are listed in Table 1. From table 1, we can see that, the classification error rates of $22^{\#}$ slice for the square window sizes 3×3, 5×5 and 7×7 are 9.6 %, 6.9% and 5.8% respectively, while that of the corresponding circle window with the radius 1, 2 and 3 is 11.8%, 3.7% and 4.5% respectively, and the optimal features are features extracted from the circle window with the radius 2. The similar results and conclusions are obtained from $24^{\#}$ slice.

Table 1. The classification error rates for different features

Type of region			$22^{\#}$ slice	$24^{\#}$ slice
Type	Size	Radius	The classification error rates（％）	The classification error rates（％）
Square	3		9.6	6.9
Circle		1	11.8	10.7
Square	5		6.9	6.0
Circle		2	3.7	5.5
Square	7		5.8	9.1
Circle		3	4.5	7.2

The brain tissues segmentation results of $22^{\#}$ slice MRI were performed with different features in Fig.2, while the comparative results with the square window sizes 3×3 and 5×5 were also obtained in the same figure. Fig.2a shows the original MR image. Fig.2b shows the under-segmentation results using the gray intensity features with the square window size 3×3, and Fig.2c shows the over-segmentation results using the gray intensity features with the square window size 5×5. The best results are obtained in Fig.2d with the circle window with the radius 2. It can be seen that by using the circle window in feature extraction, the performances of brain tissues classification are greatly improved.



(a) (b)

(c) (d

Fig.2. The original MR image and the SVM segmentation results with different window shapes and sizes

a The original MR image
b The SVM segmentation results with the square window size 3×3
c The SVM segmentation results with the square window size 5×5
d The SVM segmentation results with the circle window with radius 2

From table 1 and Fig.2, we can see that the classification error rates are very high when the window size is too small or big. When the window size is 3×3, the extracted features are incomplete, which is difficult to construct the optimal classification surface in low dimension space. For obtaining the more feature information, the big window size should be chosen. But the extracted features contain information from other outliers, which affects the classification precision. Since the outliers are always in the margin of the window, feature extraction based on the circle window reduces the number of the outliers, which improves the classification precision of the classifier.

Computation Cost Comparison

The good features have not only the high classification precision but also the low computation cost. This section gives a comparative experiment between textures and the proposed features in term of computation cost using the two multi-class SVM methods: one-against-one and one-against-rest. Table 2 shows the average computational times using both textures and gray intensities based on the circle window.

Table 2. The average computational times for different features

Silce No.	Features	Total times including training and testing(s)			
		one-against-one		one-against- rest	
		Binary output	Probability output	Binary output	Probability output
$22^{\#}$	Texture	7.38	6.56	9.52	12.13
	This proposed feature	6.17	5.84	8.5	10.56
$32^{\#}$	Texture	7.89	10.3	19.03	34.56
	This proposed feature	4.52	4.64	6.86	8.8

From table 2, gray intensities based on the circle window show less computation times than textures for brain tissues classification using SVM classifiers. Especially the classification using the proposed features takes one fourth of computation times using the texture features in the probability output of the one-against-rest method.

4. CONCLUSIONS

Feature selection is one of the key problems in pattern recognition. Too many irrelevant features degrade the performance of classifier. The optimal selection of features is important since it maximizes classes contrast differentiation in feature space, while minimizes the computational complexity when they are used. This paper proposes the gray intensity features based on the circle window for feature selection. Comparative experiments show the proposed features obtain the better performance of classification in terms of the classification precision and the computation costs using multi-class SVM methods.

REFERENCES

[1] VAPNIK, V.: ‘The nature of statistical learning theory’, *Springer-Verlag*, 1995.

[2] Mitchell,J.R.;Karlik,S.J.;Lee,D.H.;Fenster,A. “Computer-assisted identification and quantification of multiple sclerosis lesions in MR imaging volumes in the brain”.*J.Magn.Reson.Imaging4*, 1994, pp197-208.

[3] KIM,K.I..,JUNG,K.,PARK,S.H.’Supervised texture segmentation using support vector machines’, Eelectronicsetters,1999,35(22),pp.1935-1937.

[4] Issam EI-Naqa,Y.Y.Yang,M.N.Wernick.et al. A Support Vector Machine Approach for Detection of Microcalcifiations IEEE Trans. On Medical Imaging. 2002,21 (12):1552 -1563.

[5] C.A. Cocosco, V. Kollokian, R.K.-S. Kwan, A.C. Evans : "BrainWeb: Online Interface to a 3D MRI Simulated Brain Database" NeuroImage, vol.5, no.4, part 2/4, S425, Proceedings of 3-rd International Conference on Functional Mapping of the Human Brain, Copenhagen, 1997.

Xiang Zhang was born in 1969. He is currently an associate professor. He did Postdoctor research in Department of Computer Science and Technology in Tsinghua University in 2005-2006. He received his Ph.D. in Institute of Pattern Recognition and Artificial Intelligence in Huazhong University of Science and Technology in 2004. His research interests include pattern recognition, and image processing. He has published over 50 technical papers in domestic, international conferences and journals, in which 14 papers were indexed by SCI, EI, and ISTP. He has gained some honors including the first-class award in the science and technology progress prize by China Petroleum and Chemical Industry Association in 2005, and the second-class award in the Science and Technology Progress Prize by the government of Hubei province in 2005.

Construction and Research of Digitized Campus Data Center

Cailan Zhou[1], Xin Li[1], Rong Zhu[1]
[1]College of computer science and technology, Wuhan university of Technology, Wuhan, Hubei, 430070 China
Email: francislix@tom.com

ABSTRACT

This paper focuses on the Digitized Campus Data Center's design and research, and notes that the data center as a whole can be divided into foundation technology platform, unified information portal, unified database platform, unified authentication platform and application system integration which is based on the data center. Finally, the paper makes a simple exposition on the key technology of the construction of the whole data center.

Keywords: Digital Campus, Database, Application Framework, Data Center

1. INTRODUCTION

Informational education, as the lead to promote education modernization and education development by leaps and bounds, has become a strategic choice for the development of education of every country. So for, almost all colleges and universities have been constructing a variety of information management systems, which play a key role in their research, teaching and management. However, with the expansion of the Campus scale, the method of storing and sharing this information which has been accumulated by the various systems over the years has become a new problem that the universities have to solve for further development.

The first reason for emergence of this problem is，in the long process of applications constructing, colleges pay little attention and awareness for information .They are not aware that they should unify design and construct the digital campus from the whole and on a high level. Secondly, bound by the material conditions and rapid development of computer technology, the systems that are built in different period, are different in operating system and developing tools. This result in the reconstruction of source and the data can not be shared. It has specific performance in the following aspects:

1) It differs greatly at information level and means of technology. Both the systems are with the J2EE technology development, there is no system which carries on the management database using excel
2) Information multiple maintain, and sharing is not efficient and synchronized. Every system preserves and maintains the organization of university and basic information of teachers and students. Data source is not unified, so authority is not guaranteed.
3) The absence of a unified information criterion and information coding inconsistencies in different department force every department uses their own coding criterion. Data items are different uniform code and the ambiguity problem is difficult to solve.
4) Coexistence of multiple heterogeneous databases, database revenue, small ACCESS databases and some large enterprises apply the database such as SQL Server, Oracle database.
5) Information sensitivity and confidentiality requirements and some other non-technical are barriers for data sharing.

As mentioned above, at this stage, the college digital campus' construction concentrate mainly on the aspect of system integration, the original network with the expansion and upgrading of the infrastructure, data center as a platform to integrate its existing applications through a certain system of integrated and unified data format to achieve smooth sharing of information.

2. DIGITAL CAMPUS APPLICATION PLATFORM'S TECHNOLOGY ARCHITECTURE

Combining with the actual situation of domestic College campus' information construction at present, digital campus' structure should be developed in the four following levels, the overall construction architecture and arrangement is shown in fig.1.

The bottom is the network platform, including network infrastructure and network's security construction. The main network of campus utilizes the existing network infrastructure, rehabilitation and replacement of the old network equipment. Ensure teaching and administrative departments involved in network connectivity, allow investment in the maximization of network coverage, and network management software for effective monitoring and management. In this layer, network security's hardware and software's selection is crucial.

Network platform is based on foundation technology platform. Foundation technology platform is the digital campus basis. Digital Campus data center or central database is built on this platform. From the functional division, Foundation technology platform except includes the application integration、unified identity authentication、authorization system and Unified database platform, but also includes its related operating systems, each kind of system software (application service software, directory services software), database and development tools,etc.

The public components platform (Application Platform) is above Foundation technology platform. It is the important component of the whole system and structure. It consists of the public tool module. It also lets developers to fast Construction of efficient business function modules. It is characterized by:

1) Application framework is based on open or shared criterions, and bring about the product-oriented practical component library system with openness, extensibility;
2) Support heterogeneous environment framework components of the Internet and communications and implement of the new and old system compatibility.
3) It is guided by the important components criterion, and has the features such as transparent localization platform-independent.
4) System configuration and exchange of data is based on XML and Java standardized format.
5) It supports personalized information services and customized menu Reconstruction.

6) It provides a consistent, robust, scalable, flexible framework for Application's development.
7) The framework provides services provider interface (Service Provider Interface) and freedom to choose corresponding services that is provided by the third-party software such as workflow engine, authentication services, and guarantee platform scalability.

Its basic components include: service engine, workflow engine, news engines, engine entities, job scheduling engines, engine statements, log engines, and Web Application Framework (WAF).

The application components are above the application platform. Any campus construction projects are based on the application of components. Application components are consists of the foundation business entities such as teachers, organization, workers and the basic stability of the composition of business services. This design has made the functional reorganization of business very easy. If organizations changes or the functions of the body change in scope, the system does not need any re-development, just right components for assembly and re-authorization will satisfy the new demand. Application components in the design are guided by the following general principles:

1) It supports future changes: suitable for the description of the rules of business processes and the use of rule engine to achieve to avoid re-coding.

2) The simplicity of the components should be insured, in order to functionally reuse and remodel business.

3) Portfolio priority is in the use of inheritance and components for easy maintenance and expansion.

The top two tiers are digital campus application's specific business systems and unified information portal. These business systems and functional modules are established on the application components and public components based platforms to meet different business needs. The construction of the operational system is based on application components, a business process are achieved by a few components. Users are role based to use the functions in addition to the existing administrative organizations.

Unified information portal is a platform for various application components integration and deployment. In this system the separation of the different functions effectively organize themselves to a variety of users to provide a unified information service entrance. The Portal provides the customization tools of website page style、layout、content for developers and other personalized service.

3. DATA CENTER PLATFORM

3.1 The definition of data center

Data center is a center database that unifies planning and design from the overall college. Storage is involved with the college principal business' coding and data. It is different from the shared database. The main purpose of sharing database is to extract the basic information of college and the basic data of business sector. The central database is a center that consists of a core database and a number of business databases. This center is a logical center that can be geographically dispersed so as to meet different sectors of data storage requirements.

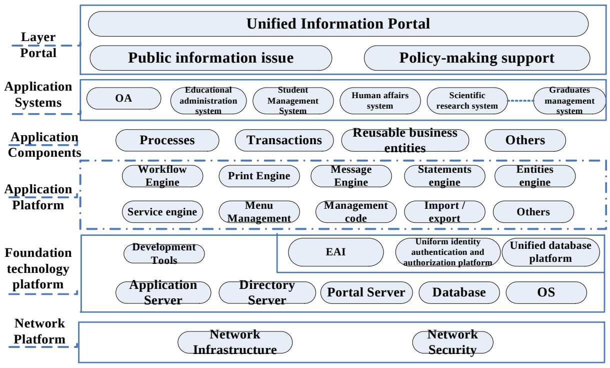


Fig.1. Digital campus application platform's technology architecture

3.2 Data Center's construction content

Digital Campus database consists of a central database and a number of databases which cover campus' business. It combines various business sector databases into an organic whole. Meanwhile with a number of middleware server components it makes up the digital campus application service platforms. The center's database in the whole digital campus should serve as a framework for the control of the central role. the construction of center database includes:

1) **The establishment of the Information Criterion**
The construction of Information criterion mainly includes the constructing information model、Formulation of data criterions and the establishment of the data dictionary etc. At present, in China's higher education field, the state criterions, industry criterions, provincial and municipal criterions and college internal criterions coexisted. Various criterions can not be completely identical, and there are many incompatibilities. This means different departments; different business' operational criterions for the use have duplication, deletion, contradictions and many other problems. In information criterions construction process, it is needed to synthesize national criterions, the department criterions, and the definition of criterion by college, establishes a unified university organization and coding, jobs coding, personnel coding and basic identity information. In the each application systems, uniform application of these basic information can realize the standardization of college organize and personnel information.

2) **Unified database platform**
Unified data platform in addition to construct a basic, broad, professional and authoritative huge database, It includes the following three levels of the constructing:

a) data criterions specifications
- Council of the system to criterion data
- Code criterions conversion

b) data management operations :
- the data add, delete, update, query operation
- Data importing and exporting
- Data correction

c) Data services
- Flexible statements
- General Query
- college grassroots statements

3) Unified Identity Authentication and Authorization System

Unified Identity Authentication and Authorization Systems (platform) are the important components in digital campus' construction. The platform provides uniform identification and delivery of competence for all digital campus 'users. Users through a certification get corresponding authority and could use all application service provided by the system. Besides, we must have a more stable and unified dynamic password along with the authentication system. So that, each user can have static password and an evolving dynamic password. Administrators can set up different areas where user's static password and dynamic password can be used; this from another angle improves the entire information security. In addition, through the designation of the centralized authentication corresponding technical specifications, provides a unified application user interface management. Ultimately all new users' authentication system unified centralized management and the focus on the real significance of certification. The realization of the application system of "centralized certification" can be completely changed fragmented, loosely user management model. Colleges could take full management responsibilities of internal network management department, standardize user behavior and strengthen user consciousness to reasonable using of network resources.

Unified Identity Authentication and Authorization System, includes three key areas: user information, storage and centralized management. User focus on certification, access permissions centralized control and management. Its structure, as shown in Fig.2:

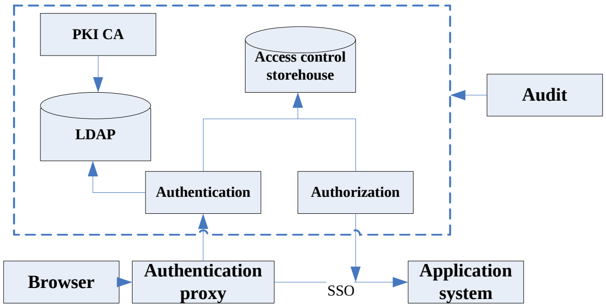


Fig.2. the Structure of Authentication and Authorization System

4) Application System's Integration

In order to make digital campus application system's integration first, we must have a deep understanding of college user's demand. China's current institutions of higher learning are mostly built with a more comprehensive campus network. The existing application systems are office information systems, academic management system, the user's accounting system and library management systems. One of digital campus' constructions focuses on the transform rich, unified and expanded for variety application systems. Therefore integration has to solve many problems; it has the following main categories:

a) Existing network infrastructure needs improvement and transformation, including the upgrading of network bandwidth, network-wide expansion, network equipment upgrading and optimization, net management, security enhancement, etc.

b) Its necessary to establish an advanced Internet-based application development system, fully utilizing J2EE, XML and other leading technologies. Create better infrastructure to support application development, deployment, unified platform, the integration of existing applications and plan the various systems and data resources to accommodate future network variety of flexible application development.

c) The network application platforms and systems need to be enhanced and improved, including the Web, e-mail services, visit schedule of services further optimization, and the implementation of application security. Based on the integration of unified user management-based campus wide e-mail, library systems, distance education system expand the scope of the application.

d) Unified information portal should be established, including one-time login, personalized service, and centralized access to online FAQs and other functions, which achieve the on-demand service requirements in digital environment.

e) The external interface needs to be fully integrated to increase access modes and types (such as Unified Messaging services) further expand colleges and the propaganda impact and external services.

4. THE KEY TECHNOLOGIES OF SYSTEM'S IMPLEMENTATION

4.1 Authentication mechanism

Identity Authentication is to show identity in two mutual main. In usually case, The user that access to the system from a single system is usually required to provide a username and password to verify the user's identity; More generally, authentication is used for two-way process test that between mainframes or processes. Web- authentications mostly use user's name and password for testing, which includes many safety hazards. In the certification system design, we can use a series of measures to strengthen the security of communications: The PKI architecture is for identity authentication and verification of the user's online identity. PKI architecture put public key cryptosystem and symmetric key together with, to realize automatic key management on the network, so ensure the on-line data's confidentiality, integrity; in addition, we could adopt time-synchronized pattern to prevent delay attack; using security encryption and SSL-based channel in transmission guarantee password security; adding random factor in the process of encryption to withstand a dictionary attack; and so on.

4.2 Unified authorization mechanism

As the college information system perfect growing is bounded to the security of information, power management, system interaction between the demands for enhanced and corresponding to the campus organizations, the role of posts complex applications is needed. Thus digital campus's Authorization System is more complicated. Digital Campus System can be divided into functional competence, operational competence and data access permissions. Therefore, the system should determine the system function, operation and data access according the Identity based on unified model. Uniform identity authentication and authorization model based on RBAC model, shown in Fig. 3. RBAC: A mandated strategy is based on mandated role as a bridge. In the business college personnel should know following concepts clearly:

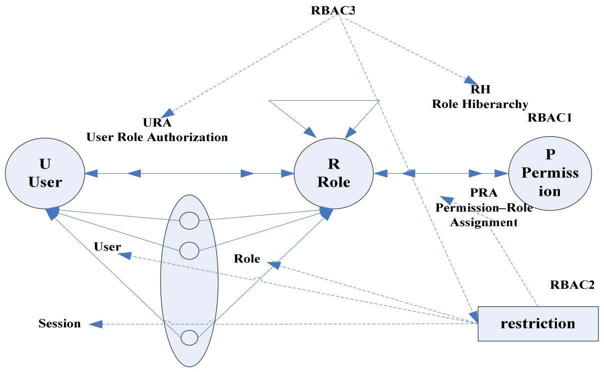


Fig.3. RBAC Model

- Organizations : correspondence college is a form of organization
- Resources competence: that the system provides all functional items, the performance of the functions of the menu items.
- Role : equivalent to the identity of the user, can be interpreted as a post
- Users: is the system for each user's account, we often used to refer to a teacher or student.

All the accesses to central database platform, first through the reunification of identity authentication, according to user's identity inquiries user's role, obtains the corresponding rules of the role. Based on the interpretation of rules to determine the functional competence of the users, and operators to restrict data access permissions.

4.3 Application system Integration

With the increasing demand for applications and resources for the accumulation, the demand for various departments' data exchange and sharing is increasing rapidly. When data center is in the design phase, we must fully consider the application platform and system integration. The digital resources based on information management platform should be integrated and consolidated information, and can provides the basis for the college leadership through data warehouse for storage, can use data mining technology. Key technologies of System integration involved are:

1. Use Web Service technology (XML-based Integration) to cross-platform firewall and information Integration.

2. By the application procedure API it can make up the data's integration between heterogeneous databases.

3. Using Portal, SSO technology and JSR-168, Wrsp criterions, through personalized portal platform integrated college's relatively independent business systems.

4. Data conversion operations (job) between the Isomorphism database and snapshot services.

5. CONCLUSIONS

Campus data center is like a heart in digital campus's construction project. It should be proved and designed in detail. We must establish an organic, intelligent network security system to protect the campus data center critical applications and to provide data security. In this paper, we presented a digital campus application framework and a central database platform structure, which have good extension and can adapt to the college system later construction. It not only gives the college constructing more advanced database schema, but also a mature system in the future. As the digital campus is a huge project, this article has only made a simple introduction to its realization of essential technology; specific details will be subject to further research.

REFERENCES

[1] Xinhua Zhang, Zhenghe Liang, Guobao Zhang, HaihongBian. "Research and Design of Digital Campus Application Framework Based on Central Database". *Microcomputer Applications(research and design edition)*,2007,23(1):19-22.

[2] Na Yin, Ming Fu. "Programming and Establishment of Digital Campus". *Journal of Lanzhou Railway University(nature sciences edition)*, 2002, 21（4):142-144

[3] XinPing Yan, Anfu Zhang, Hong Tian. "The Thought about Digitized Campus Construction". *Higher Educational Research in areas of Communications* 2001, (2) :7-9.

[4] JianBin Chen. "Key Points and Strategies on Constructing Digital Campus". *Journal of North China Institute of Technology(social sciences edition)* ,2002, (1):96-98.

[5] Yan Dao, XiaoHong Dao. "Enrich Internet Information and Accelerate the Building of Digital Campus". *Acta Scientiarum Naturalium Universitatis Sunyatseni* ,2001, (3) :121-124.

[6] Xin Xu, XinNing Su, Naigang Wu. "Design and Implement of SDC in Digital Campus". *New technology of library and information service*, 2005,6:48-53.

[7] Kai Han, Lihua Yue, Yang Yang. "Transformation and Integration of Heterogeneous XML DTDs". *Mini-Micro Systems*, 2005, 26 (1) : 119-123.

[8] Yugang Zhen, Luying Liu, Jianchu Kang. "Architecture and Implementation of an XML-based Heterogeneous Database Integration System". *Computer Engineering*, 2006, 32 (2): 85-87.

[9] Bin He Fang Guo. "Research and Implementation of Digital Campus Software Architecture". *Journal of Donghua University (Natural Science Edition)* 2005, 05.

[10] SandhuR,CoyneE,FeinsteinH,etal."Role–basedAccessControlModels" [J]. In *IEEE Computer*, 1996, 29 (2):38-47.

Quantum-behaved Particle Swarm Optimization for Medical Image Registration

Jingquan Xie[1.2], Daojun Wang[1], Wenbo Xu[1],
[1]School of Information Technology, Southern Yangtze University, Wuxi, 214122
[2]Wuxi Institue of Technology, Wuxi, 214121
Email: lucky_xiejq@hotmail.com

ABSTRACT

Medical image registration is the first step in the image fusion and other imaging processes. In this paper, the image edges are first detected by using canny operators, then the contour feature points are extracted by K-Means algorithm, and translation parameters are calculated by using Quantum-behaved Particle Swarm Optimization (QPSO) algorithm. Experiments indicate that the QPSO algorithm get better fitness comparing with the PSO algorithm and GA, it's suitable for medical image registration.

Keywords: Medical image registration; Feature control points; Quantum-behaved Particle Swarm Optimization (QPSO)

1. INTRODUCTION

Medical image fusion, a subject that involves information science, computer science and medical imaging, is one of the hotest spot in medical image processes. As a basic issue of medical image fusion, medical image registration has great significance. It is the first step of image fusion, so its precision influenced the effect of image fusion directly.

The goal of image registration is to find a transformation that aligns one image to another. Mainly, there are two kinds of methods in medical image registration, one is statistic-based, and the other is feature-based. Much work has focused on statistic-based approaches recently, in which the intensity values (color, or gray level) are used to compute similarity measures between images. Experiences shows that this method has its own shortcomings such as a large amount of complex calculations and its high computational complexity need a long-time registration process that is unacceptable in medical clinic practices. Feature-based registration is versatile in these conditions that it can be applied to any image at least in theory, no matter what the object or subject is. Moreover, in the feature-based registration, the set of identified points are sparse compared with the original image content, which makes for relatively fast optimization procedures[1,2,3]. The later was adopted in our experiences

A wide various optimization techniques have been applied in medical image registration. Local methods, such as Powell's direction set method, conjugate gradient , Levenberg–Marquardt , or the Nelder–Mead simplex algorithm [4,5] , are generally used. These methods still frequently trapped in local optima, as the global optimum may not be present in lower resolutions [6,7]. Therefore, global optimization is required. Simulated Annealing (SA) and Genetic Algorithms (GA) methods are new probabilistic heuristic algorithms which have been successfully used in solving optimization problems, GA method is usually faster than SA method because of its parallel search ability. But recent research has identified that it may degrade in highly epistatic objective functions [8].

Particle Swarm Optimization (PSO) is a modern population based evolutionary search technique, which first introduced by Kennedy and Eberhart, it has successfully optimized a wide range of continuous functions and have generated higher quality solutions in shorter time, at the same time, it has more stable convergence characteristic than other stochastic methods [9]. Quantum-behaved Particle Swarm Optimization (QPSO) is first proposed by Sun, it is more effective for actual social system, because a social system is far more complex than that formulated by anyone of equation and even the thinking mode of an individual of the social system is so intricate that a linear evolvement equation is not sufficient to depict [10, 11]. In Sun's paper, the experiment results indicate that the QPSO algorithm works better than standard PSO on several benchmark functions and it's a promising algorithm. The experiment results of image registration problem are shown in Section IV, and it prove that QPSO is better than PSO in this case.

2. REGISTRATION METHODS

2.1 Image Transform Model

Let $V=(v_x,v_y)^T$ be the points of image $f_1(x,y)$,and $U=(u_x,u_y)^T$ be the points of image $f_2(x,y)$,the relation between them can be described as a rigid transformation model :

$$V = sRU + T$$

$$R=\begin{bmatrix}\cos\theta & \sin\theta \\ -\sin\theta & \cos\theta\end{bmatrix}$$

Where R is a rotation matrix, s , θ and $T=\left(t_x,t_y\right)^T$ is the scale, the angle and the translation vector between the above images respectively, the superscript T denotes transpose.

2.2 Feature Control Points Extraction

Obtaining feature control points is an important step in a feature-based registration. In this paper, feature control points are extracted out in three steps.

First, images are convolved with the canny operator [12] and the edges are detected by the threshold value-0.15.

In the second step, the contours are detected from the above edges based on the 8-connected boundary tracking method.

After getting the contours, K-Means algorithm [13] is introduced to obtain the feature control points (80 points used in this paper).

2.3 Feature Control Points Matching Function

The main idea is to find a transformation, so a finite number of feature points in template image A can be mapped to the feature points in the reference image B. We usually use a distance function measuring the differences between points set X and set Y. The special function is defined as:

$$C(R,T) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\min_{j=1}^{M}\left\|Rx_i + T - y_j\right\|}$$

where x_i is the feature points in template image A, y_j is the feature points in the referencing image B, N and M are the numbers of feature points in the sets X and Y respectively, R is the rotation matrix, T is the translation vector. In order to obtain the optimistic values of R and T, we calculate the minimum value of C. In this paper, we propose a new optimization approach to solve this problem.

3. PARTICLE SWARM OPTIMIZATION AND QUANTUM-BEHAVED PARTICLE SWARM OPTIMIZATION

3.1 Overview

Particle Swarm Optimization (PSO) algorithm, originally introduced by Kennedy and Eberhart in 1995, is a population-based evolutionary computation technique. It is motivated by the behavior of organisms such as fishing schooling and bird flock. In a PSO system, each particle corresponding to individual of the organism is a candidate solution to the problem at hand. Particles of the population fly around in a multi-dimensional search space, to find out an optimal or sub-optimal solution by competition as well as by cooperation among them [9].

Quantum-behaved Particle Swarm Optimization (QPSO) algorithm, proposed by Jun Sun, the search space and solution space of the problem are two spaces of different, and the search space of an individual particle at each iteration is the whole feasible solution space of the problem, the experiment in results Sun's paper indicate that the QPSO works better than standard PSO on several benchmark functions [10, 11]].

3.2 Particle Swarm Optimization (PSO)

The PSO algorithm maintains a population of particles, where each particle represents a potential solution to an optimization problem. Let S be the size of the swarm, each particle i can be represented as an object in the D-dimensional space as:

$$X_i(t) = (X_{i1}(t), X_{i2}(t), \ldots X_{iD}(t))$$

and $V_i(t) = (V_{i1}(t), V_{i2}(t), \ldots V_{iD}(t))$.

The particles move according to the following equation:

$$V_i(t+1) = \omega V_i(t) + c_1 * rand1() * (pbest_i - X_i(t)) + c_2 * rand2() * (gbest - X_i(t)),$$

$$i = 1, 2 \cdots S,$$

$$X_i(t+1) = X_i(t) + V_i(t+1),$$

Where ω is the weight factor, c_1 and c_2 are the acceleration constant, $rand1()$ and $rand2()$ are the uniform random value in the range [0,1], $V_i(t)$ is the velocity of particle i at iteration t, $X_i(t)$ is the current position of particle i at iteration t.

The weight factor ω provides a balance between global and local explorations. The constants c_1 and c_2 represent the weighting of the stochastic acceleration terms that pull each particle toward the $pbest$ and $gbest$ position. Low values allow particles to roam far from the target regions before being tugged back. On the other hand, high values result in abrupt movement toward, or past, target regions [9].

3.3 Quantum-behaved Particle Swarm Optimization (QPSO)

The dynamic behavior of the particle in Quantum-behaved Particle Swarm Optimization (QPSO) is widely divergent from that of the particle in traditional PSO system in that the exact values of V and X can't be determined simultaneously. In QPSO, the particle moves according to the following equation:

$$mbest = \frac{1}{S}\sum_{i=1}^{S} pbest_i = \left(\frac{1}{S}\sum_{i=1}^{S} pbest_{i1}, \frac{1}{S}\sum_{i=1}^{S} pbest_{i2}, \ldots \frac{1}{S}\sum_{i=1}^{S} pbest_{iD}\right)$$

$$p_i = \varphi * pbest_i + (1-\varphi) * gbest,$$

$$\varphi = \text{rand1()}$$

$$X_i = p_i \pm \alpha * \left|mbest - X_i\right| * In(1/u)$$

$$\text{u=rand2()}$$

$$\alpha = (\alpha_1 - \alpha_2) * \frac{(MAXITER - t)}{MAXITER} + \alpha_2$$

where mbest is the mean best position among the particles, p_i is a stochastic point between $pbest_i$ and $gbest$ stand for the local attractor of the particle, φ and u are random values in the range [0,1], α is a parameter of QPSO that is called Contraction-Expansion Coefficient, α_1 and α_2 .are the initial and final values of the parameter α, t is the current iteration number and MAXITER is the maximum number of allowable iterations [10,11].

The only parameter in the algorithm is Contraction-Expansion Coefficient, α, which is called Creativity Desire of the particle and works on individual particles convergence speed and the performance of the algorithm [11].

3.4 Implementation of the QPSO Algorithm

The proposed QPSO approach has been tested on image registration using the following steps:

1) Choose the population size and number of generation
2) Select transform parameters(θ , T_X , T_Y), as state variables { X_i, i= 1, 2, 3}
3) Generate randomly S particles { $X_i(0)$, i= 1, 2, 3}, where $X_{ik}(0)$ is generated by random selecting a values with uniform probability over the k^{th} optimized parameter search space $[X_{\min}, X_{\max}]$

   Evaluate the fitness of each particle according to the objective function.
4) For each particle, set the initial position as its pbest position, and set the particle of best fitness as the gbest position.
5) Set time counter t=1
6) Calculate the mbest by:

$$mbest = (\frac{1}{S}\sum_{i=1}^{S} pbest_{i1}, \frac{1}{S}\sum_{i=1}^{S} pbest_{i2},$$
$$\dots \frac{1}{S}\sum_{i=1}^{S} pbest_{iD})$$

7) Recalculate the position of each particle by:

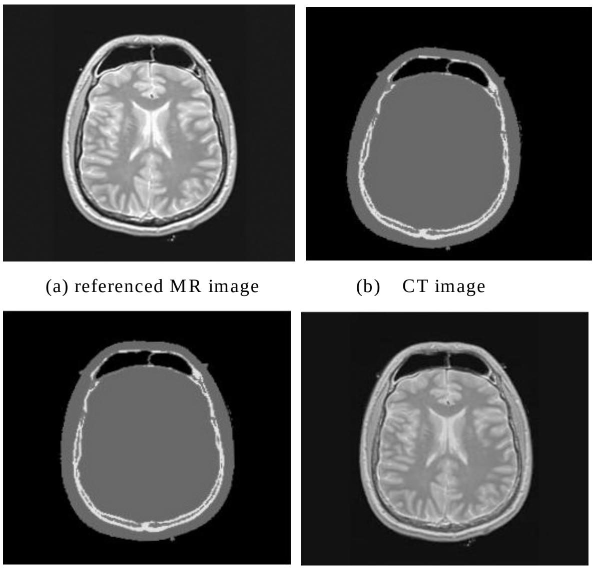
$$X_i = p_i \pm \alpha * |mbest - X_i| * In(1/u)$$
$$p_i = \varphi * pbest_i + (1-\varphi) * gbest$$

8) For each particle X_i, it's the fitness is better than the $pbest_i$, update $pbest_i$; update $gbest$ if the best fitness of this generation is better.

9) Set time counter $t = t+1$, if $t > MAXITER$, then the program stop; else go to step7.

4. RESULTS AND ANALYSIS

The combination of MR and CT is beneficial, where the former is better suited for delineation of tumor tissue (and has in general better soft tissue contrast), the latter is needed for accurate computation of the radiation dose.
Fig.1 shows experimental results of MR-CT medical image registration.



(a) referenced MR image (b) CT image

(c) moved CT image (d) fused MR-CT image

Fig.1.the experimental results of MR-CT medical image registration.

The proposed (QPSO) method has been applied to MRI-CT registration compared with the PSO method and the GA method. As the value of Root Mean Square(RMS) is fairly small, the matching error is calculated by $RMS^2 \times N\,(\sum_{i=1}^{N} \min_{j=1}^{M} \|Rx_i + t - y_j\|)$.

Table 1 shows the fitness of QPSO, PSO and GA method in 50 times:

Table 1. 50 times date

	Times	Iteration	Mean Fitness	Standard Deviation
GA	50	200	213.46	17.7091
PSO	50	200	199.62	9.3564
QPSO	50	200	189.91	5.0119

Fig.2. shows that the evolution process of PSO method, QPSO method and GA method.

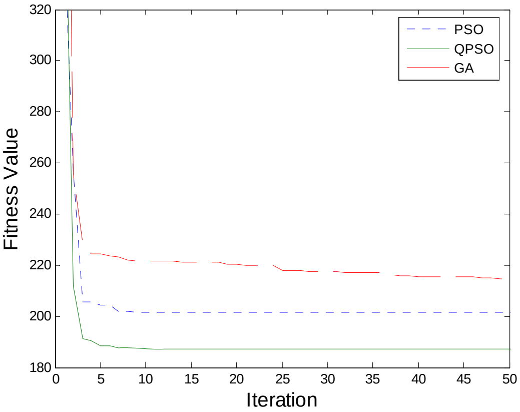


Fig.2. the evolution process of three types algorithm

It is evident that the QPSO method is better than PSO method both in speed and final fitness. The mean fitness of QPSO is 189.91 compared with 199.62 using PSO. Referring to standard deviation, QPSO is 5.0119 and PSO is 9.3564.

Table 2 shows the best fitness of PSO and QPSO compared with GA:

Table 2. THREE TYPES ALGORITHM

	Best Fitness	(θ, T_X, T_Y)
GA	199.39	(-5.8186,0.7016,-0.6966)
PSO	193.03	(-7.0291,3.7393,-4.2898)
QPSO	187.42	(-4.9510,-1.4438,0.6191)

5. CONCLUSIONS

This paper presents an enhanced PSO method--QPSO method for medical image registration. The proposed method utilizes an enhanced global searching method for complex problem, and we successful used it in the MRI-CT registration. The results show that QPSO is better than PSO method**.** We have only considered the two-dimensional registration. Our further work is to apply the method to three-dimensional registration.

REFERENCES

[1] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," Phys. Med. Biol., vol. 46, pp. R1–R45, 2001.

[2] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," Med. Image Anal., vol. 2, pp. 1–37, 1998.

[3] L. Brown, "A survey of image registration techniques," ACM Comput. Surv., vol. 24, pp. 325–376, 1992.

[4] J. L. Bernon, V. Boudousq, J. F. Rohmer, M. Fourcade, M. Zanca, M.Rossi, and D. Mariano-Goulart, “A comparative study of Powell’s and downhill simplex algorithms for a fast multimodal surface matching in brain imaging,” Comput. Med. Imaging Graph., vol. 25, pp. 287–297,2001.

[5] F. Maes, D. Vandermeulen, and P. Suetens, “Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information,” Med. Image Anal., vol.3, pp. 373–386, 1999.

[6] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, “Medical image registration,” Phys. Med. Biol., vol. 46, pp. R1–R45, 2001.

[7] M. Jenkinson and S. Smith, “A global optimization method for robust affine registration of brain images,” Med. Image Anal., vol. 5, pp.143–156, 2001.

[8] K. Price, "Differential Evolution: A Fast and Simple Numerical Optimizer," *1996 Biennial Conference of the North American, IEEE*, 1996, pp. 524-527.

[9] J.Kennedy and R.C. Eberhart, “Particle swarm optimization,” in *Proceedings of the IEEE International Conference on Neural Networks*, IV,1995, pp. 1942-1948.

[10] J. Sun, W.B. Xu, and B. Feng, “A Global Search Strategy of Quantum-Behaved Particle Swarm Optimization,” *Conference on Cybernetics and Intelligent Systems*, Proceedings of the 2004 IEEE, Singapore, 2004, pp.111-116.

[11] W.B. Xu and J. Sun, “Adaptive Parameter Selection of Quantum-Behaved Particle Swarm Optimization on Global Level,” ICIC 2005, Springer-Verlag Berlin Heidelberg 2005.

[12] Canny J. “A computational approach to edge detection [J].” *IEEE Transactions on Pattern Analysis and Machine Intelligence*.1986,8(6):679-698.

[13] Ursem, R.K.: Diversity-Guided Evolutionary Algorithms.Proceedings of the Parallel Problem Solving from Nature Conference. (2001).

Image Data Quality Control Based on Bootstrap Algorithm*

Wei Yi, Haifeng Ni, Yiwei He
School of Automation, Wuhan University of Technology
Wuhan, Hubei, 430070, P. R. China
Email: shirley_yi_wei@yahoo.com; nihaif@sina.com; airari@gmail.com

ABSTRACT

A digital image is the projection of a real scene through imaging systems. Although there are large numbers of advanced imaging equipments with high resolution and sensitivity, various defects still exist and may corrupt images. As many image analysis algorithms and their performances can be affected by the qualities of the original images, the preprocessing techniques are always adopted firstly to eliminate those defects. However in some biomedical applications, no preprocessing can be used in order to keep the original data for the further analyses. Based on the Bootstrap algorithm, this paper conducts the data quality control of protein spots in the electrophoresis gel images. The quantization of polluted spots is achieved without damaging the original data. Experiments show the proposed method is effective in the real data extraction.

Keywords: Bootstrap Algorithm, Particle Filter, Data Quality Control, Gel Image, Elliptical Paraboloid Fitting

1. INTRODUCTION

It is well known that a digital image is the 2D projection of a 3D scene. In the ideal situation, the image only contains the background and foreground. However due to the physical characteristics of the imaging systems and other factors, images are usually corrupted by the defects such as the noise, blur, shadow etc that consequently affect the analyses. Some necessary preprocessing techniques are applied in those cases to reduce the noise, improve the image quality and offer “clean and clear” images. In some biomedical applications, the scientists require original data to analyze the biological properties, and the preprocessed data are no longer considered as reliable although the image quality is improved.

Based on the Bootstrap algorithm[1][2][3], this paper targets the quantization problem in the 2D electrophoresis gel analysis field. Through recursive sampling and estimation procedures, it carries out the quantization of protein spots polluted by various defects without damaging the original data. Experimental results demonstrate its better performance over the traditional quantization approach.

This paper is organized as follows: the second part briefly introduces the principle of the Bootstrap algorithm. The third part gives the detailed quantization method based on the Bootstrap sampling and estimation. The fourth part shows the experimental results on the image data quality improvement, and the conclusions are drawn in the final part.

2. PRINCIPLES OF THE BOOTSTRAP ALGORITHM

The Bootstrap algorithm is a resampling technique that was firstly introduced by Efron in 1979. The Bootstrap idea is to replace the unknown population distribution with the known empirical distribution through the Bootstrap sampling. Because of its generality, it has been applied to a wild class of problems in many areas and is also referred to as Particle Filter.

Efron’s Bootstrap is defined as follows: Given a sample of n independent identically distributed random vectors $X=(X_1,X_2,\dots,X_n)$, and θ is the distribution parameter under estimation. $\hat{\theta}=\theta(X_1,X_2,\dots,X_n)$ is a real-valued known estimator of θ. A procedure of the Bootstrap is to assess the accuracy of $\hat{\theta}$ through the empirical distribution function $\hat{F}_n$, where $\hat{F}_n$ has the probability mass $\frac{1}{n}$ in each observed value $x_i\ (i=1,2,\dots,n)$ of the random vectors X_i.

From the Bootstrap sampling, a Monte-Carlo approximation is among one of the most frequently used approaches to obtain the estimation. It contains four basic steps: (1) Generate a sample of size n from the empirical distribution; (2) compute the $\hat{\theta}$ by using the Bootstrap sample in place of the original population; (3) Analyze the probability characteristics and accuracy of $\hat{\theta}$; (4) Repeat the above steps.

3. BOOTSTRAP BASED IMAGE DATA QUALITY CONTROL

3.1 Gel image Analysis

The electrophoresis is one of the most frequently used techniques to separate proteins. Combined with high-resolution imaging systems, a gel image of the protein spots can be obtained. As the gel image analysis offers quantified evidence, it has been playing an important role in the bioscience field. The traditional inspections of the gel images are visually completed by the human. Intensive labor cost and largely subjective factors unavoidably lead to the limited reliability of the analyses. Therefore it is significant to automatically interpret the gel images through computers. The gel image analysis mainly focuses on the spot detection, spot quantization and spot matching. The spot detection separates and outlines the border of each spot. The quantization accumulates the grey value of each pixel inside the spot as its volume. The matching is the volume comparisons between the unknown spot and reference spot.

A gel image is very likely to be polluted by many artifacts during the electrophoresis and imaging process. What is more, spots themselves are irregular in shapes and may partly overlap on each other. The existing detection

This work is funded by the Ministry of Education Overseas Scholar Returning project 2005[383].

techniques[4][5][6][7][8] fail to correctly segment spots in all circumstances and directly bring fault data into the grey value accumulation where the final matching is based. If the preprocessing is applied before the spot detection, the processed spots are no longer valid for the analyses from the bioscientific research point of view.

3.2 3D Surface Fitting

The protein spots have various shapes in a gel image. The contour of a spot is a closed irregular curve. If the grey level of each pixel represents the third dimension, a 3D coordinate system can be built. For an ideal clean and full spot, the 3D view is a hill like surface as shown in the Fig. 1. When the gel images are contaminated, the 3D view of a protein spot contains the spikes on the top of its hill as shown in the Fig. 2. Based on the Bootstrap sampling and estimation approach, sample pixels from a polluted spot are randomly and repeatedly taken to fit 3D surfaces and the corresponding volumes are calculated. For each spot, a volume selection is run to pick up all valid volumes that meet the statistical threshold. The final quantization of such a spot is the mean of all valid volumes.

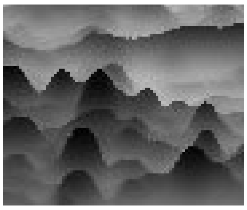


Fig.1. 3D view of the clean spots

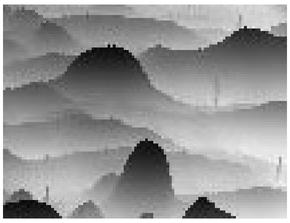


Fig.2. 3D view of the polluted spots

Based on the hill surface shown in the Fig. 1, the elliptical paraboloid quadric equation is used to describe the 3D surface of a protein spot

$$z = Ax^2 + Bxy + Cy^2 + Dx + Ey + F \qquad (1)$$

where (x, y) is the coordinate of a pixel inside the spot, and z is the corresponding grey value of the pixel.

Randomly sample the pixels within a spot. In order to calculate the coefficients in (1), the minimum sample number n must be 6. In practice, the condition $n >> 6$ must be met to achieve the high accuracy of the 3D fitting. Apply the samples to (1), and the grey level equation of the spot is

$$z = A\Delta x^2 + B\Delta x\Delta y + C\Delta y^2 + D\Delta x + E\Delta y + F \qquad (2)$$

where $(\Delta x, \Delta y)$ is the centralized coordinate of the sample.

Based on (2), the grey value correction of the sample i is

$$er_i = A\Delta x_i^2 + B\Delta x_i\Delta y_i + C\Delta y_i^2 + D\Delta x_i + E\Delta y_i + F - z_i^0 \, (3)$$

where z_i^0 is the original grey value of the sample i, er_i is the difference between the fitting and original grey values, and $i = 1,2,...,n$ is the number of the samples. If the matrix format is used here, (3) can be rewritten as

$$Er = KT - Z^0 \qquad (4)$$

where

$$\begin{cases} Er = \begin{bmatrix} er_1 \\ er_2 \\ \vdots \\ er_n \end{bmatrix} \quad T = \begin{bmatrix} A \\ B \\ \vdots \\ F \end{bmatrix} \quad Z^0 = \begin{bmatrix} z_1^0 \\ z_2^0 \\ \vdots \\ z_n^0 \end{bmatrix} \\ K = \begin{bmatrix} \Delta x_1^2 & \Delta x_1\Delta y_1 & \Delta y_1^2 & \Delta x_1 & \Delta y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Delta x_n^2 & \Delta x_n\Delta y_n & \Delta y_n^2 & \Delta x_n & \Delta y_n & 1 \end{bmatrix} \end{cases} \qquad (5)$$

The least square solution of the coefficient matrix T in (4) is

$$T = (K^T K)^{-1} K^T Z^0 \qquad (6)$$

The elliptical bottom must be firstly under rotation and translation normalization to obtain the standard ellipse format,

$$A' x_{tr}{}^2 + C' y_{tr}{}^2 + F' = 0 \qquad (7)$$

and the volume of the elliptic paraboloid is

$$V = \frac{1}{2}\pi abh \qquad (8)$$

where

$$\begin{cases} a^2 = \left| -\frac{F'}{A'} \right| \\ b^2 = \left| -\frac{F'}{C'} \right| \\ h = F \end{cases} \qquad (9)$$

3.3 Data Quality Control Based on the Bootstrap Algorithm

Due to the unavoidable pollution in the gel images, the volume calculated from only one fitting process is not close enough to that of the real protein spot. The data quality control based on the Bootstrap algorithm is necessary to be adopted here. That is, after the multiple sampling and fitting to one spot, a set of volumes V_i $(i = 1, 2, ..., m)$ can be obtained. Get the mean value and standard deviation of these m volumes

$$\begin{cases} \overline{V} = \frac{1}{m}\sum_{i=1}^{m} V_i \\ \sigma_V = \pm\sqrt{\frac{\sum_{i=1}^{m}(V_i - \overline{V})^2}{m-1}} \end{cases} \qquad (10)$$

Let $2\sigma_V$ be the threshold and compare each V_i with the mean $\overline{V}$. Only when the differences between V_i and $\overline{V}$ meet

$$\left| V_i - \overline{V} \right| < 2\sigma_V \qquad (11)$$

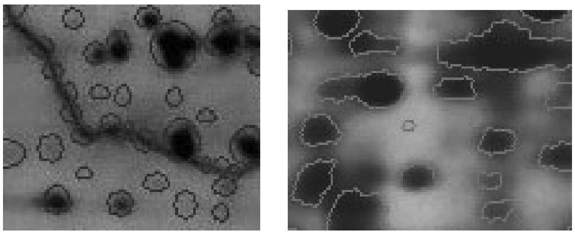
is the V_i considered as valid and still reserved. Otherwise V_i is discarded. Repeat the above process until all V_i $(i = 1,2,...,l)$ satisfies (11). The final volume of the spot $\overline{V}'$ and its accuracy $\sigma_{\overline{V}'}$ are

$$\begin{cases} \bar{V}' = \frac{1}{l}\sum_{i=1}^{l} V_i \\ \sigma_{\bar{V}'} = \frac{\sigma'_V}{\sqrt{l}} \end{cases} \tag{12}$$

where

$$\sigma'_V = \pm\sqrt{\frac{\sum_{i=1}^{l}(V_i - \bar{V}')^2}{l-1}} \tag{13}$$

It is worthwhile to mention here that noise is not the only factor affecting real protein spots. As stated above, the spot detection may fail to correctly outline a spot boundary in some cases as can be seen in the Fig.3.



(a) (b)

Fig.3. False spot detections

Fig.3 (a) and (b) are the real spot detection results. It is not difficult to find out that the image crack and speckle are wrongly recognized as the protein spots. In those cases, the 3D views of these fake spots do not fit the elliptical paraboloid model and therefore generate invalid parameters of (6). The algorithm jumps out and redoes the fitting from the new samples. If no valid parameters can be obtained for a spot, the algorithm finally gives out the warning message and reminds the user the spot may be extracted by a wrong detection.

4. EXPERIMENTAL RESULTS

Large experiments have been carried out on the gel images with different qualities. Some results are shown in Table 1. The results in Table 1 demonstrate that the fitting volume is close to the accumulation volume if the spot is clean. The volume difference is much lower than the $3\sigma'_{\bar{V}}$ statistical threshold. However if a spot is contaminated by the noise, the accumulation volume is much larger than the fitting volume. The volume difference is bigger than the statistical threshold. In addition, as stated in the last section, the algorithm can reject fake protein spots as no valid parameters can be obtained.

Another experiment is designed to show the proposed quantization approach outperforms the traditional one. A clean and full spot (original volume = 23964) is chosen with gradually increased noise adding on it. The volumes are calculated through two methods and the results are shown in Table 2. The first column is the noise intensity. The second and the fourth columns are the fitting volumes and accumulation volumes respectively. The third column is the differences between the fitting volume and the original one. The fifth column is the differences between the accumulation volumes with the original one. It is obvious that when the noise intensity increases, the fitting volume is still statistically reliable and stable. On the other hand, the accumulation volume has no power to resist the noise and therefore the corresponding results are equally polluted as the original spot.

Table 1. Volume comparisons

Fitting volume	Accumulation volume	Volume difference	$3\sigma'_{\bar{V}}$
24154	23964	190	721.65
30957	30071	886	664.53
72196	90839	18643	3905.97
109611	156994	47383	9646.95
41686	50648	8962	2734.05
30384	38379	7995	2125.65

* The data unit in the above Table is pixel.

Table 2. Simulation experiments

Noise intensity	Fitting volume	Fitting volume difference	Accumulati on volume	Accumulati on volume difference
0	24151	190	23964	0
0.004	23892	72	25489	1525
0.006	24133	169	26017	2053
0.008	23985	21	26116	2152
0.01	24170	206	26205	2241
0.02	24395	431	26264	2300
0.03	24423	459	24423	2923

* The data unit in the above Table is pixel.

5. CONCLUSIONS

The Bootstrap algorithm is a statistical tool that analyzes the whole population through the samples. Due to its flexibility and simplicity, the Bootstrap approach has been widely used in many areas such as the quality control, object recognition and tracking etc. This paper presents a novel algorithm to complete the spot quantization based on the Bootstrap algorithm. After the sampling, 3D fitting and quality control, the corrupted spot data can be well reconstructed and yield more accurate and reliable quantization results.

REFERENCES

[1] M. R. Chernick, Bootstrap Methods, *A Practitioner's Guide*, John Wiley & Sons Inc., 1999.

[2] S. J. Li, H. Wang, T. Y. Chai, "A t-Distribution-based Particle Filter for Target Tracking," *Robot*, Vol. 28, No. 6, November, 2006, 598-604.

[3] X. J. Yang, Q. Pan, R. Wang, H. C. Zhang, "Development and Prospect of Particle Filter," *Control Theory and Applications*, Vol. 23, No. 2, April, 2006, 261-267.

[4] P. S. Umesh Adiga and J. Flint, “An efficient tool for genetic experiments: agarose gel image analysis,” *Pattern Recognition, Vol. 36*, 2453-2461, 2003.

[5] I. Bajla, I. Hollander, K. Burg and S. Fluch, “A novel approach to quantitative analysis of electrophoretic gel images of DNA fragments,” *IEEE International Symposium on Biomedical Imaging* , Washington DC, USA, 899-903, 2002.

[6] A. Efrat, F. Hoffmann, K. Kriegel, C. Schultz and C. Wenk, “Geometric algorithms for the analysis of 2d-electrophoresis gels,” *Proceedings of the 5th Annual International Conference on Computational Biology*, Montreal, Canada, 114-123, 2001.

[7] T. Akutsu, K. Kanaya, A. Ohyama and A. Fujiyama, Point matching under non-uniform distortions, *Discrete Applied Mathematics, Vol. 27,pp.* 5-21, 2003.

[8] R. Wilson, “Modelling of 2D gel electrophoresis images for proteomics databases,” *16th International Conference on Pattern Recognition*, *Vol. 1,* Quebec, Canada, 767-770, 2002.

Wei Yi is currently an associate professor at School of Automation, Wuhan University of Technology. She received her BEng and MSc from Xidian University and Wuhan University in 1994 and 1997 respectively, both from P. R. China. In 2001, she obtained Ph.D. from Dept. of Electrical and Electronic Engineering, University of Strathclyde, Glasgow, UK. Dr. Wei’s main research interests are pattern recognition and machine vision. She has published over 20 technical research papers in these areas and authored the monograph Advanced Statistics and its Applications in Shape Analysis. Dr. Wei is in charge of several Municipal and Ministerial Projects. She is now working as a visiting scholar at Dept. of Electrical and Computer Engineering, University of Waterloo, Canada.

A Parallel Modeling Algorithm for Semantic Image Hierarchal

HongXia Shi[1], Yi Ouyang[2]
[1,2]College of Computer and Information Engineering, Zhejiang Gongshang University
Hangzhou, Zhejiang 310035, China
Email: oyy@mail.hzic.edu.cn

ABSTRACT

A parallel modeling algorithm for semantic image hierarchal is proposed. The image semantic network frame(ISNF) structure is based on a 'weighted' similarity measure for comparing pairs of images data, composed by two distances, the so-called color feature and texture feature. The parallel image hierarchal create (PIHC) algorithm links the image semantic information and image Features together, through related semantic of the characteristic of low-level of the pictures and concept entities on the high-level of commercial affair. In order to show the performance of the proposed PIHC algorithm, a simulation study and two illustrative applications are discussed, on which our proposed approach as well as other related techniques are implemented and compared. Extensive experiments are conducted to investigate the effectiveness of our PIHC algorithm, which is found to be consistently better than other approaches.

Keywords: Semantic Frame, Image Feature, Parallel Compute, Image Retrieval

1. INTRODUCTION

With the rapid development of computer technologies, powerful distributed servers systems are becoming ubiquitous. At the same time, more and more pictures information are applied to the commercial website. The digital information has become increasingly popular in recent years. Effective searching of large-scale image database remains as challenges for computer systems.

Although the content results can be obtained in searching by manual tagging and establishing index, due to the multitude of Internet Image Database, it urgently needed an feature automatically labels Semantic Image Retrieval System. The Content-Based Image Retrieval (CBIR) can meet this demand better, and it is one of the key technologies of improving human-computer interaction too, so it is becoming a hot topic in recent years. The references below are to be taken as examples of related work, not as the complete list of work in the cited area. [1-6] .

We proposed a parallel image hierarchal creating algorithm, which can automatic organization of images into hierarchal so that images within a class have high similarity in comparison to one another, but are very dissimilar to images in other classes.

2. THE SEMANTIC NETWORK MODELING

Information retrieval system of commercial image (Image Information Parallel Retrieval System IIPRS) including three stages, the semantic of drawing stage, the semantic segmentation the commercial image and image features matching stage. All commercial semantic information are constructed a network by the commercial concept entity.

2.1 Image Semantic Network Frame

In order to distinguish the semantic relation of image information, this paper, according to the frame structure of multi-level classifies the commercial semantic concept entity at first, and then divides the concept entity of the commercial field into three levels. On the senior level in order that the semantic of the commercial image is perceived layer, the middle layer is semantic relation level of the commercial image. The first floor is feature layer of the commercial image.

Firstly, we define R as the relation among the commercial concepts, and S as the constraint, which can be representing as (C_S, R_S); C_S describes the constraint of the concept, and shows the semantic link relationship between commercial concept nodes; R_S means the constraint of the relation, it can be described as $R_S(E_i, E_j) = \frac{1}{R(i,k) + R(k,j)}$

Definition 1 Image Semantic Network Frame(ISNF) is a triple M=<C,R,S>, C is a hierarchical for the commercial concepts. It shows the level structure of the commercial field concept;

Among the ISNF, each concept entity has its' own link relation with other concept entities, but the different level have the different power weight. So, they construct a Semantic Concept Network structure. At the same time, the inheritance tree with link nodes provides the primary nodes. It can be located using a frame through the sub-class or upper class. Unit provides the structure of main slot, which will be used for storing and processing messages of a given theme. Slot provides the data structure which stores the special attribute messages of a frame class, including relation slot, attribute slot, method slot, and rule slot. With inheritance, the inherent value and the default value may or may not be inherited from classes (upper class), and it passes the inherent value and the default value to a sub-class. The relations of the four elements are shown in Fig.1

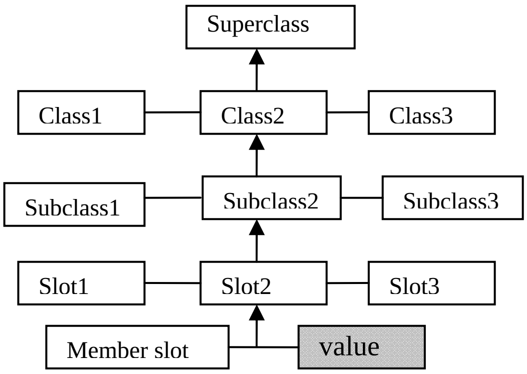


Fig.1. ISNF semantic frame

2.2 Automatic Feature Extraction

In the image searching fields, the color features is used the most extensive vision characteristic. The main reason lies in the color and object or scenes in the picture are closely related. In addition, compared with other vision characteristics, the dependence on size, direction, visual angle of the image itself of the color features is relatively small, thus the color features has better robust character. Researches on color features mainly adopt colored histogram, color moment [7], color collect [8], histogram refinement [9] and color correlation diagram [10], etc. method to color at present.

The purpose of extracting image features lies in describing the images in the database with the features of low-level. These image features are useful for calculating similarity degree among different images, so it is obviously smaller than the primitive image on the size.

- **The color features:**The grayscale histogram extracted from an image is a vector that has 256 dimensions. All vectors are contained in histogram space S(H). The histogram formula of the distance as follows: $d_i(H')=\sqrt{(H'-H_i)^t(H'-H_i)}$
- **Gabor texture feature:**Researches on the texture at present, commonly used methods have Tamura[11], wavelet transform[12][13], etc. Moreover, Gabor filters method can reduce the space and frequency to the greatest extent. Set the g(x, y) as a Gabor function. $g_{m,n}(x,y)=a^{-m}g(x',y'),a>1$ $x'=a^{-m}(x\cos\theta+y\sin\theta)$, $y'=a^{-m}(-x\sin\theta+y\cos\theta)$, m=0,1,…S-1. After the Gabor filtering, the mean of the image that is $\mu_{m,n}$,and the square of the standard deviation is $\sigma_{m,n}$. $\mu_{m,n}=\iint|M_{m,n}(x,y)|\,dxdy$ $\sigma_{m,n}=\sqrt{\iint(|M_{m,n}(x,y)|-\bar{\mu}_{m,n})^2dxdy}$ Where $M_{m,n}(x,y)=\int I(x,y)g'_{m,n}(x-x_1)(y-y_1)dx_1dy_1$ I(x,y) is a picture, and use the $\mu_{m.n}$ and $\sigma_{m,n}$ as components, these components constitute an eigenvector that can be used in the process of retrieval. Each Image can be regarded as the texture feature of the sub image. $X=[\mu_{00},\sigma_{00},\mu_{01},\sigma_{01},....\ \mu_{57},\sigma_{57}]$ The distance $d(X_i,X_j)$ formula as follows: $d(X_i,X_j)=\sum_{k=1}^{N}(X_{ik}-X_{jk})^2$

3. FRAMEWORK OF THE PIHC ALGORITHM

The global image database is DB, the amount of images is N. Let $P_1,P_2,....,\ P_n$ to be the distributed computer nodes, which only have the message transfer function. DB_i is the division database of DB, and it has Ni of items. $DB=\bigcup_{i=1}^{n}DB_i$, $N=\sum_{i=1}^{n}N_i$

Definition 2 For one of itemset X, the local frequent images item in local database DB_i is defined as X_c^i , which is the count of X including images.

Definition 3 A global frequent item GF_c^i refers to an item that belongs to some global frequent itemset.

Definition 4 The global support GS_c^i of an image is the percentage of images containing the itemset, which similarity is greater than threshold value.

In the leaf node of ISNF structure, there is a link between relevant semantic concept entities and the image in the image database. Each link is connected with a weight $w_{i,j}$ (where i represents the index number of the image, j is the symbol of the concept entity). It shows the related degree among the concept entities in ISNF and the image.

Based on [11], we put forward a kind of image-concept relation power weight feedback approach that can acquire correlation from the query images and the return images set which are coming from relevance feedback cycle of the user. The results of mark will be strong relevant, weak relevant or irrelevant results.

The searching procedure of relevant feedback is divided into two stages. The first stage is the sample study stage, where T is the minimum confidence degree about the number of searching times, domain expert finishes this stage mainly. Chorse,Ccloth, Ccandy, Capple. Their sytem links relation follow as:

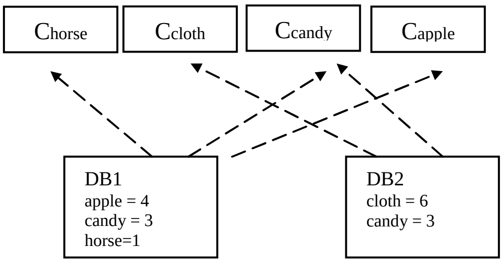


Fig.2. ISNF concept link relation diagram

Intuitively, a class C_i is good for a document M_j if there are many global frequent items in M_j that appear in many documents in C_i.

A global frequent item is class frequent in a class Ci if the item is contained in some minimum fraction of documents in C_i.

$Score(C_i-M_j)=[\sum_x f(x)*c\sup(x)]-[\sum_{x'} f(x')*g\sup(x'')]*$

x represents a global frequent item in M_j and the item is also class frequent in C_i.

x' represents a global frequent item in M_j but the item is not class frequent in C_i.

f(x) is the frequency of x in the feature vector of M_j.

f(x') is the frequency of x' in the feature vector of M_j.

Firstly, all images are marked by sample training stage, and the correlation among each concept leaf node of ISNF and the subclass in classification image database are created.

The second stage is the system self-learning stage, through searching based on the knowledge database to modify the semantic relevant degree between image and each of the concept entity in subclass.

For each concept entity $E_j\in DOM$, the system will find out subclass, then system choose Parallel Image Hierarchal Create(PIHC) algorithm to search the similar images in DOM fields. It will assign the weight among the concept node. The framework of the algorithm PIFS is as follows:

Algorithm PIHC(M, D) //parallel image hierarchal create

```
Input: M: image feature array; D:Domain image-concept feature: HIC: Hierarchal image class
Output: HIC
Begin
Initial M to getting color feature and the texture feature;
Select a superclass $C'_k$ concept node from D;
    For(Concept Entity $E_j \in C'_k$)
        If Score(M, $E_j$)> $\mu_{\text{threshold}}$   then
          ICB(M,CH,HIC)
        Else
          Partition D into subpartitions L and R
          Solve L using PIHC (M,L)
          Solve R using PIHC (M,R)
        End if
        Next
End
```

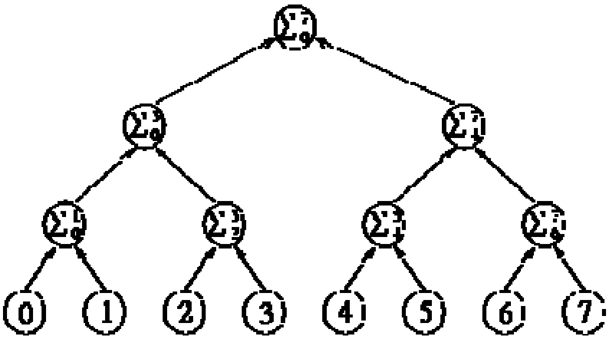


Fig.3. Tree structure for divide-and-conquer summation algorithm with N=8. The N numbers located in the tasks at the bottom of the diagram are communicated to the tasks in the row immediately above; this each perform an addition and then forward the result to the next level. The complete sum is available at the root of the tree after logN steps.

```
Algorithm ICB(c, CH,HIC) //insert class balance
Input:CH: concept hierarchical HIC: Hierarchal image class
C: the concept ready to enter HIC
Output: HIC
Parent(X,c):concept c's parent node in concepts set X
Children(X,c): concept c's children node in concepts set X
Minparent($c_1$, $c2$): the lowest public parent concept node between concept node $c_1$ and $c2$.
Begin
Step1: let $c_p$ =c
    While $c_p$ doesn't exist in HIC
    Set $c_p$ =parent(CH, $c_p$ )
Step2: Set $c_s$ =children(HIC, $c_p$ )
    For each node $c_i$ in $c_s$
        Set $c_p$ =minparent(c, $c_i$ )
        If c= =$c_p$ then goto step4
        If $c_p$ <> $c_q$ then goto step 5
    Next
Step3: add the c as the $c_p$ 's children node into HIC
    Exit
Step4: add the c as the $c_p$ 's the children node into HIC
    And link $c_i$ into c
    Exit
Step5: add the $c_q$ as the $c_p$ 's the children node into HIC
    add the c as the $c_q$ 's the children node into HIC
    And link $c_i$ into $c_q$
    Exit
End
```

Let c as the new concepts, in the first stage, we can find the c's parent nodes, which has exist in those hierarchal nodes, and choose the shortest path then mark as c_p ; In the second stage, check the node c to find whether the total hierarchal will be imbalance or not, when add c node as the the c_p 's child node. If not, then execute the algorithm's third stage, system will add the c node, which as the c_p 's child node, into the hierarchal. If the hierarchal will be imbalance, according to the "balance definition", c_p has only a child node c_i , and the lowest public parent concept node between c and c_i is different from c_p . In this situation, if c_q and c is an identical concept, system go to Step4, and put c into between the c_p and c_i ; else system put c_q as the c_p 's child node into subclass, and then let c and c_i as the c_q 's child nodes, put they into subclass.

In summary, we observe that in developing an efficient parallel image feature compute algorithm, we have distributed the N-1 communication and computation operations required to perform the ICB procedure and have modified the order in which these operations are performed so that they can proceed concurrently. The result is a regular communication structure in which each task communicates with a small set of neighbors.

The algorithm meet the all ideal image classification criterion, this class include those relevant image concepts, and little adjective concepts to maintain the hierarchal balance at first. Secondly, the hierarchal will be constantly updated to make the structure to meet the image database feature. At last, the algorithm guarantees the HIC structure's balance.

4. EXPERIMENTAL RESULTS

This system has been implemented with a commercial image database including about 116,000 pictures, which are stored in JPEG format. We use about 5 kinds of image data set as training sample image database.

Table 1. Five kinds of the image data set

Data Set	Number of Images	Number of Classes	Class Size	Average Class Size
Cloth	**7804**	**5**	**1023-3452**	**1874**
Autos	**2421**	**7**	**115-783**	**384**
Camera	**3243**	**23**	**23-656**	**214**
Cell phone	**8234**	**73**	**213-32423**	**234**
Computer	**1239**	**39**	**5-234**	

We can measure the system performance from two aspects:(1)The learning strategy efficiency of the commercial images and the ISNF correlation;(2)The performance of the images retrieve, the evaluative parameter use Precision

$$\text{Precision} = \frac{|\,\mathrm{Re}\mathit{levant} \cap \mathrm{Re}\mathit{trieved}\,|}{|\,\mathrm{Re}\mathit{trieved}\,|}$$

Where Relevant shows relevant pictures, Retrieved shows the pictures searched out, | x | represents the number of picture x.

The first performance index can come to metric through the study velocity of the picture and the concept correlation, define $\mathrm{Re}\mathit{lation}_i$ as the related degree of semantic characteristic of the ith picture:

$$\mathrm{Re}\mathit{lation}_i = \frac{|N_i|}{|C_i|}$$

Where $|N_i|$ is the number of correlation between concept entities and images, $|C_i|$ is the number of images in subclass C_i .

In the experiment, choose "The digital camera " ," The desktop computer ","The clothing ","The laser printer" and "Bread " 5 concept entities, carry on the search 30 times. Each procedure of searching carries on the information feedback cycle 10 times again, and marks the relative degree of semantic characteristic after each information relevance feedback.

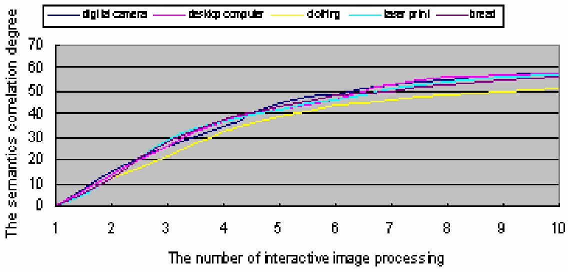


Fig.4. The results ISNF structure dynamically correlative with the image low-level features

From the Fig.4, we can see the algorithm about the ISNF and the low-level feature of image having a quicker adaptability. The semantics correlation degree between the concept entities in the ISNF and the image database comes to a 50% after 10 feedback cycle. But because the image contents variety of "clothing" class compares greatly, the low-level features discrepancy compares greatly, it make learning curve lower.

The second performance index, the experiment adopts 10 concept entities, in the query 200 times. The average precision of the first 20 pictures can find out in the Fig.5.The ISNF-IF method and traditional CBIR [12] method are compared in the commercial image retrieval, the ISNF-IF method has better precision.

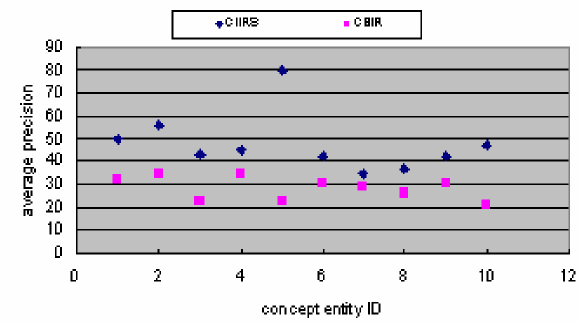


Fig.5. ISNF-based IR compares with traditional CBIR

We compare those four image retrieval approaches, Tradition CBIR(TCBIR), Tradition Parallel CBIR(TPCBIR), ISNF-IR as well as ISNF Parallel IR(ISNF-PIR).

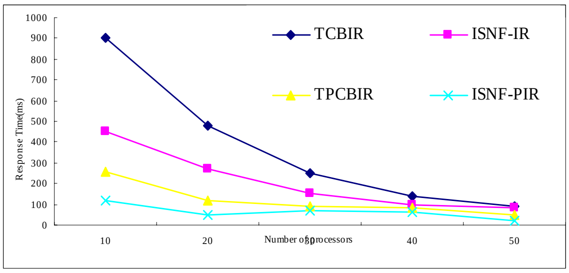


Fig.6. The results in ISNF－IF method after submitting the "digital camera" example image

The results can be seen from above Fig.6. After classifying through ISNF semantic on the multi-level, the ISNF-IR's response time is improved greatly through parallel algorithm.

5. CONCLUSIONS

Through ISNF-PIF algorithm, combining grayscale histogram and Gabor wave filter, the color and texture features can draw from every image. By Euclid distance computing similarity, system retrieve the images according to the commercial semantic content. The search result and the person's subjective perception have a certain consistency.

The result of the test indicates that ISNF cannot totally express the semantic information of the commercial picture yet at the same time. We will work from entity's structure of thinning ISNF concept further, and the comprehensive characteristic of constructing the color, texture and shape will be set about.

REFERENCES

[1] A. Pentland, R.W. Picard, and S. Sclaroff, “Photobook: Tools for Content-Based Manipulation of Image Databases”, International Journal of Computer Vision,18(3),pp.233-254,1996.

[2] A. Gupta and R. Jain, “Visual Information Retrieval,”Comm. ACM.vol. 40, no. 5, pp.70-79, May 1997.

[3] M. Flickner, H. Sawhney,W. Niblack, J. Ashley, Q. Huang, B. Domet al. “Query by Image and Video Content: The QBIC System” ,IEEE Computer, vol. 28, no. 9, 1995.

[4] J.R. Smith and S.F. Chang, “VisualSEEk: A Fully Automated Content-Based Image Query System,” Proc. ACM Multimedia,pp. 87-98, Nov. 1996.

[5] E.G.M. Petrakis and A. Faloutsos, ”Similarity Searching in Medical Image Databases,” IEEE Trans. Knowledge and Data Eng,vol. 9, no. 3, pp. 435-447, May/June 1997.

[6] J.Z. Wang, G. Wiederhold, O.Firschein, and X.W. Sha,”Content-Based Image Indexing and Searching Using Daubechies' Wavelets,” *Int'l J. Digital Libraries*, vol. 1, no. 4, pp. 311-328, 1998.

[7] M.Stricker and M.Orengo,”Similarity of color images,” *SPIE Storage and Retrieval for Image and Video Databases III*, vol 2185,pp.381-392,Feb,1995.

[8] John R. Smith and Shih-Fu Chang. Tools and techniques for color image retrieval. In Proc of SPIE: Storage and Retrieval for Image and Video Database. vol 2670,1995.

[9] G..Pass and R.Zabih,”Histogram refinement for content-based image retrieval,” IEEE Workshop on Application of Computer Vision, pp.96-102,1996.

[10] Niblack, et al., “The QBIC project: querying images by content using color, texture, and shape,” Proc. Of SPIE, Storage and Retrieval for Image and Video Database, vol.1908,San Jose,pp.173-187, February 1993.

[11] Ling-Yun, Ouyang-Yi,Li-Biwei. The ECommerce Information Model Driven Semantic Searching Alogrithm. 2006 Inernational Symposium on Distributed Computing and Applications to Business, Engineering and Science,2:840-844,2006.

[12] Y.Rui,T.S.Huang. “A Novel Relevance Feedback Technique in Image Retrieval,” Proc.Int.ACM Conf.on Mutimedia, pp.67-70, 1999. 1989.

HongXia Shi is a Full Professor in College of Computer and Information Engineering, Zhejiang Gongshang University. She graduated from Zhejiang Gongshang University in 1983; She preside over several project of the science and technology of Zhejiang Province, and participating in many National and province natural science funds projects ,etc. She obtain outstanding teaching achievement second prize of Zhejiang Province. Her research interests are in distributed parallel processing, intellectual information processing, e-commerce and soft project.

A Novel Modular PCA Method Based on Phase Congruency Images * for Face Recognition

Zhanting Yuan[1], Yanfeng Jin[1,2], Qiuyu Zhang[1], Jiawen Hu[1], Lei Sun[1], Wenjing Li[3]
[1]School of Computer and Communication,Lanzhou University of Technology,Lanzhou,730050,China
[2]Direct Mail Research&Consulting Center, China Post Group, Shijiazhuang, 050021, China
[3]College of Mathematics and Information science, Northwest Normal University, Lanzhou, 730050, China
Email: jinyanf@126.com

ABSTRACT

A novel modular PCA algorithm for face recognition based on phase congruency is presented in this paper. Phase congruency features in an image are defined as the points where the Fourier components of that image are maximally in- phase. These features are invariant to brightness and contrast of the image under consideration. This property allows to achieve the goal of lighting invariant face recognition. Firstly phase congruency maps of the training samples are generated. Then the phase congruency face images are divided into smaller sub-images and the PCA approach is applied to each of these sub-images. Since some of the local facial features of an individual do not vary even when the pose, lighting direction and facial expression vary, we expect the proposed method to be able to cope with these variations. The accuracy of the modular PCA method and the proposed method are evaluated under the conditions of varying expression, illumination and pose using standard face databases. The results indicate high improvement in the classification performance compared to the conventional modular PCA method.

Keywords: Face Recognition, Principal Component Analysis, Modular PCA, Phase Congruency

1. INTRODUCTION

Face recognition is a difficult problem because of the generally similar shape of faces combined with the numerous variations between images of the same face. The image of a face changes with facial expression, age, viewpoint, illumination conditions, noise etc. The task of a face recognition system is to recognize a face in a manner that is as independent as possible of these image variations. Automatic recognition of faces is considered as one of the fundamental problems in computer vision and pattern analysis, and many scientists from different areas have addressed it. A survey on several statistical-based, neural network-based and feature-based methods for face recognition was presented[1]. Currently, one of the methods that yields promising results on frontal face recognition is the principal component analysis (PCA), which is a statistical approach where face images are expressed as a subset of their eigenvectors, and hence called eigenfaces[2,5]. PCA has also been used for handprint recognition[6], human-made object recognition [7], industrial robotics[8], and mobile robotics[9].But results show that the recognition rate is not satisfactory for pose variations exceeding 30 and extreme changes in illumination. The main objective of this research is to improve the accuracy of face recognition subjected to varying facial expression, illumination and head pose. As stated before, PCA method and modular PCA method have been popular technique in facial image recognition. But both are not highly accurate when the illumination and pose of the facial images vary considerably. In this research work an attempt is made to improve the accuracy of this technique under the conditions of varying facial expression, illumination and pose. We propose the novel modular PCA method, which is an extension of the conventional modular PCA method. In the novel modular PCA method, before the face images are divided into smaller images, we pre-process the face images by using phase congruency firstly and then the PCA method is applied on each of the divided smaller images. Whereas in the traditional PCA method the entire face image is considered, hence large variation in pose or illumination will affect the recognition rate profoundly. Since in the case of our novel modular PCA method the phase congruency face image is divided into sub-images the variations in illumination in the image will affect only some of the subimages, hence we expect this method to have better recognition rate than both the conventional PCA method and the modular PCA method. This paper is organized as follows: Section 2 describes the technique of implementation to obtain phase congruency features from the corresponding intensity images. Section 3 explains the novel modular PCA method. Section 4 presents simulation results obtained by applying the PCA method and the conventional modular PCA method and the proposed novel modular PCA method to the face image sets with large light and pose variations. Finally, a conclusion is drawn in Section 5.

2. PHASE CONGRUENCY FEATURES

Gradient-based operators, which look for points of maximum intensity gradient, will fail to correctly detect and localize a large proportion of features within images. Unlike the edge detectors, which identify the sharp changes in intensity, the phase congruency model detects points of order in the phase spectrum. According to Opeinheim and Lim [10], phase component is more important than the magnitude component in the reconstruction process of an image from its Fourier domain. There is also physiological evidence, indicating that human visual system responds strongly to the points in an image where the phase information is highly ordered. Phase congruency provides a measure that is independent of the overall magnitude of the signal making it invariant to variations in image illumination and/or contrast. Fig.1 shows phase congruency image and the corresponding intensity image. The phase congruency technique used in this paper is based on the one developed by Peter Kovesi [11].

Phase congruency function in terms of the Fourier series expansion of a signal at some location x is given by:

$$PC(x) = \frac{\sum_n A_n \cos(\phi_n(x) - \phi(x))}{\sum_n A_n} \tag{1}$$

Where A_n represents the amplitude of the nth Fourier component, and $\phi_n(x)$ represent the local phase of the Fourier

*The Natural Science Foundation of Gansu (3ZS062-B25-037)

component at position x . $\phi(x)$ is the weighted mean of all the frequency components at x . Phase congruency can be approximated to finding where the weighted variance of local phase angles relative to the weighted average local phase, is minimum. An alternative and easier interpretation of phase congruency is proposed in [11]. It is proposed that energy is equal to phase congruency scaled by the sum of the Fourier amplitudes as shown in equation 2.

$$E(x) = PC(x)\sum_n A_n \tag{2}$$

Hence phase congruency is stated as the ratio of $E(x)$ to the overall path length taken by the local Fourier components in reaching the end point. This makes the phase congruency independent of the overall magnitude of the signal. This provides invariance to variations in image illumination and contrast. $E(x)$ can be expressed as $E(x) = \sqrt{F(x)^2 + H(x)^2}$. If $I(x)$ is the input signal then $F(x)$ is the signal with its DC component removed and $H(x)$ is the Hilbert transform of $F(x)$ which is a 900 phase shift of $F(x)$. Approximations to the components $F(x)$ and $H(x)$ are obtained by convolving the signal with a quadrature pair of filters. In order to calculate the local frequency and phase information in the signal, logarithmic Gabor functions are used. If $I(x)$ is the signal and M_n^e and M_n^o denote the even symmetric and odd-symmetric wavelets at a scale n . The amplitude and phase of the transform at a given wavelet scale is given by equation 3 and equation 4 respectively.

$$A_n = \sqrt{e_n(x)^2 + o_n(x)^2} \tag{3}$$

$$\phi_n = tag^{-1}(o_n(x)/e_n(x)) \tag{4}$$

Where $e_n(x)$ and $o_n(x)$ are the responses of each quadrature pair of filters. Equation 5 illustrates the response vector.

$$[e_n(x), o_n(x)] = [I(x) * M_n^e, I(x) * M_n^o] \tag{5}$$

$F(x)$ and $H(x)$ can be obtained from the equations 6 and 7.

$$F(x) = \sum_n e_n(x) \tag{6}$$

$$H(x) = \sum_n e_n(x) \tag{7}$$

And $\sum_n A_n$ at x is given by equation 8.

$$\sum_n A_n(x) = \sqrt{e_n(x)^2 + o_n(x)^2} \tag{8}$$

If all the fourier amplitudes at x are very small then the problem of phase congruency becomes ill conditioned. To overcome the problem a small positive constant ε is added to the denominator. The final phase congruency equation is given by equation 9.

$$PC(x) = \frac{E(x)}{\varepsilon + \sum_n A_n} \tag{9}$$

One-dimensional analysis is carried out over several orientations, and the results are combined to analyze a two dimensional signal (image)[11].

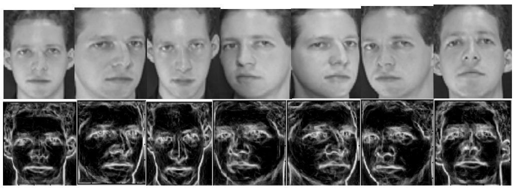


Fig.1. Phase congruency map obtained from the intensity image

3. PROPOSED NOVEL MPCA METHOD

The PCA based face recognition method is not very effective under the conditions of varying pose and illumination, since it considers the global information of each face image and represents them with a set of weights. Under these conditions the weight vectors will vary considerably from the weight vectors of the images with normal pose and illumination, hence it is difficult to identify them correctly. On the other hand if the phase congruency face images were divided into smaller regions and the weight vectors are computed for each of these regions, then the weights will be more representative of the local information of the face. When there is a variation in illumination, only some of the face regions will vary and rest of the regions will remain the same as the face regions of a normal image. Hence weights of the face regions not affected by varying pose and illumination will closely match with the weights of the same individual's face regions under normal conditions. Therefore it is expected that improved recognition rates can be obtained by following the novel modular PCA approach. We expect that if the phase congruency face images are divided into very small regions the global information of the face may be lost and the accuracy of this method may deteriorate. In this method, each preprocessed image in the training set is divided into N smaller images. Hence the size of each sub-image will be L^2/N .These sub-images can be represented mathematically as:

$$I_{ij}(m,n) = I_i(\frac{L}{\sqrt{N}}(j-1)+m, \frac{L}{\sqrt{N}}(j-1)+n) \tag{10}$$

Where i varies from 1 to M , M being the number of images in the training set, j varies from 1 to N , N being the number of sub-images and m and n vary from 1 to $L/\sqrt{N}$ .Fig.2 shows the result of dividing a phase congruency face image into four smaller images using Eq.(5) for N = 4. The average image of all the training sub-images is computed as:

$$A = \frac{1}{M \cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij} \tag{11}$$

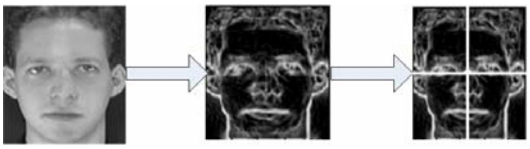


Fig.2. A face image divided into N smaller images, where N = 4

The next step is to normalize each training subimage by subtracting it from the mean as:

$$Y_{ij} = I_{ij} - A \tag{12}$$

From the normalized sub-images the covariance matrix is computed as:

$$C = \frac{1}{M \cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N} Y_{ij} \cdot Y_{ij}^T \tag{13}$$

Next we find the eigenvectors of C that are associated with the M' largest eigenvalues. We represent the eigenvectors as $E_1, E_2, \cdots, E_{M'}$ .The weights are computed from the eigenvectors as shown below:

$$W_{pnjK} = E_K^T \cdot (I_{pnj} - A) \qquad \forall p, n, j, K \tag{14}$$

Where K takes the values $1, 2, ..., M'$, n varies from 1 to Γ, Γ being the number of images per individual, and p varies from 1 to p, p being the number of individuals in the training set. Weights are also computed for the test sub-images using the eigenvectors as shown in the next equation:

$$W_{test\ jK} = E_K^T \cdot (I_{test\ j} - A) \qquad \forall j, K \tag{15}$$

Mean weight set of each class in the training set is computed from the weight sets of the class as shown below:

$$T_{pjK} = \frac{1}{\Gamma} \sum_{K=1}^{M'} \sum_{n=1}^{\Gamma} W_{pnjK} \qquad \forall p, j \tag{16}$$

Next the minimum distance is computed as shown below:

$$D_{pj} = \frac{1}{M'} \sum_{K=1}^{M'} \left| W_{test\ jK} - T_{pjK} \right| \tag{17}$$

$$D_p = \frac{1}{N} \sum_{j=1}^{N} D_{pj} \tag{18}$$

$\min(D_p) < \theta_i$ for a particular value of p, the corresponding face class in the training set is the closest one to the test image. Hence the test image is recognized as belonging to the p th face class.

4. EXPERIMENTS

We carry out the experiments on ORL face database[12]. The ORL database has 400 images of 40 adults, 10 images per person. The face images vary with respect to facial expression and illumination. The images have normal, sad, happy, sleepy, surprised, and winking expressions. There are also images where the position of the light source is at the center, left and right. Out of the 10 images of a person, only seven were used for training and the remaining three were used to test the recognition rates. Fig.3 a and b show the set of images of a person used for training and testing respectively. The choice of the training and test images was made to facilitate comparison of performance of the three methods for test images with uneven illumination and partial occlusion. We also conducted experiments by leaving out one image from each individual's set of 10 images during training and testing the recognition rate with the images left out. This was repeated 10 times by leaving out a different image each time. This kind of testing is referred to as leave out one testing in the remainder of the paper.

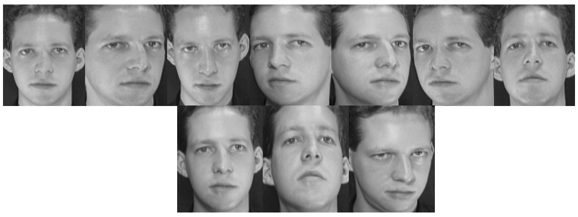


Fig.3. Show the set of images of a person used for training and testing

We tested the performance of PCA and modular PCA and our novel modular PCA algorithms for varying number of eigen-vectors. Considering more eigenvectors results in increased recognition rates, however the increase in computational cost is linear with the number of eigenvectors. Fig.4 shows the recognition rates of PCA and modular PCA and our novel modular PCA for varying number of eigenvectors. The results shown in Fig.4 were obtained using the ORL face database by leaving out one testing. Threshold was not used for this testing; hence there are no rejections, only correct recognition or false recognition. It can also be observed from Fig.4 that the recognition rate is increasing in both PCA and modular PCA methods as we increase the value of M', and there is not much improvement for $M' > 30$. Similar results have been observed for values N = 4, 16, 64, 256 and 1024. It has been observed that the novel modular PCA algorithm provides better recognition rate with the preprocessed face images.

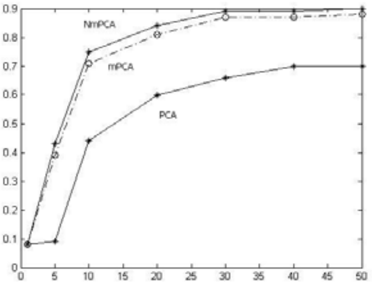


Fig.4. The recognition rates of PCA, mPCA and NmPCA for varying number of eigenvectors

In the next experiment we compared the recognition rate, false recognition rate and false rejection rate of the two methods for large expression and illumination variations using the images in the ORL database. The training and test images were chosen as described in next Section. As before we vary the value of N from 4 to 4096 to observe the effect it has on face recognition. Fig.5shows the recognition rate, false recognition rate and false rejection rate for the novel modular PCA method with varying N .In the case of mPCA the recognition rate was 0.44, false recognition rate was 0.31 and false rejection rate was 0.24.

A second set of experiments were performed by leaving out one testing. The results obtained for novel modular PCA are shown in Fig.6. For mPCA, recognition rate was 0.48, false recognition rate was 0.36 and false rejection rate was 0.16.

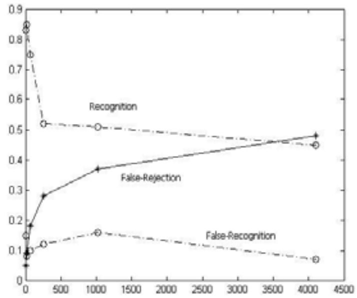


Fig.5. (a) Shows the recognition rate, false recognition rate and false rejection rate of NmPCA when mPCA are 0.44, 0.31, 0.24

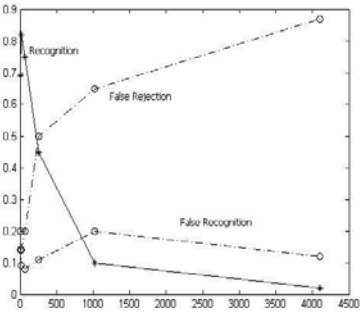


Fig.5. (b) Shows the recognition rate, false recognition rate and false rejection rate of NmPCA when mPCA are 0.48, 0.36, 0.16

The PCA based method was not very effective under the conditions of varying illumination, since it considers the global information of each face image and represents them with a set of weights. Under this condition the weight vectors of the test image will vary considerably from the weight vectors of the training images with normal illumination, hence it is difficult to identify them correctly. The huge improvement in the case of novel modular PCA was observed since the preprocessed face images were divided into smaller regions and the weight vectors were computed for each of these regions, hence weight vectors will be more representative of the local information of the face. Therefore for variations in illumination, the weights of the face regions not affected by varying illumination closely match with the weights of the same individual's face regions under normal conditions. This leads to better recognition results using novel modular PCA as observed in the experimental results.

5. CONCLUSIONS

A novel modular PCA method, which is an extension of the PCA method and also the improvement of the conventional modular PCA method for face recognition has been proposed. The novel modular PCA method performs better than the PCA method and the conventional modular PCA method under the conditions of large variations in expression and illumination. For large variations in pose there is no significant improvement in the performance of novel modular PCA. But for large variations in illumination it performs much better. For face recognition, the novel modular PCA method can be used as an alternative to the PCA method. In particular, the novel modular PCA method will be useful for identification systems subjected to large variations in illumination and facial expression.

REFERENCES

[1] Chellappa, R., Wilson, C.L., Sirohey, S., 1995, "Human and machine recognition of faces: A survey," in *Proc. IEEE83 (5)*, 705-740.

[2] Kirby, M., Sirovich, L., 1990, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. PatternAnal. MachineIntell*, 12(1),103-108.

[3] Graham, D.B., Allinson, N.M., 1998, "Characte-rizing virtual eigen signatures for general purpose face recognition," in: F*ace Recognition: From Theory to Applications*, NATO ASI Series F, Computer and Systems Sciences, vol.163, pp.446-456.

[4] Moghaddam,B.,Pentland,A,1997, "Probabilistic visual learning for object representation," *IEEE. Trans. Pattern Anal. MachineIntell*, PAMI-19(7), 696-710.

[5] Martinez, A.M., 2000, "Recognition of partially occluded and/or imprecisely localized faces using a probabilistic approach," in *Proc of Computer Vision and Pattern Recognition*, vol. 1, pp.712–717.

[6] Murase, H., Kimura, F., Yoshimura, M., Miyake, Y.,1981, An improvement of the auto-correlation matrix in pattern matching method and its applications to handprinted 'HIRAGANA'.Transactions on IECE J64-D(3).

[7] Murase, H., Nayar, S., 1995, "Visual learning and recognition of 3-D objects from appearance," *Int.J. Computer Vision* 14, 5–24.

[8] Nayar, S.K., Nene, N.A., Murase, H., 1996, "Subspace methods for Robot vision," *IEEE Trans.Robot. Automat*, RA-12 (5), 750–758.

[9] Weng, J.J., 1996, "Crescepton and SHOSLIF: towards comprehensive visual learning," in Nayar, S.K., Poggio, T. (Eds.), *Early Visual Learning*, Oxford University Press, pp.183– 214.

[10] A.V. Oppenheim, J.S. Lim, "The importance of phase in signals," *IEEE Proceedings*, v 69, May 1981, pp 529-541.

[11] [11] P. Kovesi, "Edges Are Not Just Steps," *The 5th Asian Conference on Computer Vision*, pp. 23-25, January 2002.

[12] Ferdinando Samaria, "Andy Harter. Parameterisation of a Stochastic Model for Human Face Identification," in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota FL, December 1994.

Zhanting Yuan: Professor and doctor tutor in the School of Computer and Communication Engineering, Mr Yuan received his MASc in June 1989 from Artificial Intelligent and Robot Research Center of Xi'an Jiaotong University. Now Professor Yuan is the Science Leader of computer science and technology, director of Gansu manufacturing information engineering research center, administrative director of Chinese electrical higher education, and the first batch of Chinese new century "Bai qian wan" person with ability. The research interests of professor Yuan include: image processing and pattern recognition, computer vision, Software engineering, information security etc.

Yanfeng Jin: Graduate student. Born in Shijiazhuang Hebei province in 1981, he has published many academic papers in domestic core magazine and international conference His research interests include: image processing and pattern recognition, computer vision, information security.

On a Class of Graphs with Disjoint Cycles

Fugui Liu[1], Kaisheng Lu[2]
[1]Science College, Energy and [2]Power Engineering College,
Wuhan University of Technology, Wuhan, China 430063
Email: lufugui620@163.com

ABSTRACT

This paper gives a group of graphs with three disjoint cycles: If G is a graph, and $\varepsilon(G)$ $\varepsilon(G) \ge v(G)+6, v(G) \le 7$, or $\varepsilon(G) \ge v(G)+7v(G) \le 9$, then G has three disjoint cycles. This result is considered to be optimal.

Key words: Cycle, Girth, Connected Graph, Isomorphism

1. INTRODUCTION

The graphs discussed in this paper are all non-directional simple graphs, the graph with vertex set *V* (*G*) and side set *E* (*G*) is noted as *G* = (*V* (*G*) ,*E* (*G*)). And the numbers of vertexes and sides for Graph *G* are respectively *V(G)*, $\varepsilon(G)$. If any subgraph of *G* is connected and each vertex is 2 degree, then such subgraph is regarded as a cycle of *G* and noted as *C*. The number of sides for *C* is the cycle length of *C*. If the two cycles C_1 and C_2 of *G* have no common side, then C_1 and C_2 are called two disjoint cycles. The shortest cycle length in *G* is regarded as the girth of G. Other symbols and terms refer to literatures[1-3].

L. Pósa once proved that: when the side number and vertex number of G satisfy $\varepsilon(G)$ $\ge v(G)+4$, then G has disjoint cycles[2]. Mr. Zhu Ruijun proved that if the numbers of sides and vertexes of G satisfy $\varepsilon(G) \ge v(G)+8$ and $v(G) \le 12$, then G has at least three disjoint cycles[4]. The conclusion of this paper is given as follow, and this conclusion is going to answer an issue, which has put forward in Literature [5].

2. MAIN RESULTS AND PROOFS

Theorem If side number $\varepsilon(G)$ and vertex number $v(G)$ of Graph G satisfy one of the following conditions:

(1) $\varepsilon(G) \ge v(G)+6 \,\&\, v(G) \le 7$;

(2) $\varepsilon(G) \ge v(G)+7 \,\&\, v(G) \le 9$.

Then Graph G has at least three disjoint cycles.

Proof This theorem actually has two independent conclusions, which will be proved respectively as follow.

(I). Firstly to prove when Graph *G* satisfies $\varepsilon(G) \ge$ $v(G)+7 \,\&\, v(G) \le 9$,G has at least three disjoint cycles.

Suppose *G* is a graph satisfying $\varepsilon(G) \ge$ $v(G)+7 \,\&\, v(G) \le 9$, However, it does not have three disjoint cycles and has minimum vertex numbers. Then G has the following properties:

(i) The girth of *G* is $g \ge 4$.

In fact, from $\varepsilon(G) \ge v(G)+7$, we can know *G* has cycles. Assuming *G* has a cycle with the length equal or less than 3. Here we only to prove the case with the cycle length is 3, the case for the cycle length of 2 can be proven in the same way. Assuming the cycle with the length of 3 in *G* is C_1,and draw Graph $G_1 = G - E(C_1)$,then G_1 satisfying $\varepsilon(G_1) \ge$ $v(G)+4 = v(G_1)+4$. From the conclusion proven by L. Pósa, we can know that G_1 has two disjoint cycles C_2, C_3 . Apparently, C_1, C_2, C_3 are disjointed. This contradicts the assumption of G .

(ii)The minimum degree of G : $\delta(G) \ge 3$

In fact, if $\delta(G) \le 1$,assuming $u \in V(G)$ and $d(u) \le 1$, draw Graph $G' = G - \{u\}$,then G' satisfy $\varepsilon(G')$ $\ge v(G')+7$,which is contrary to the definition of G ;If $\delta(G) = 2$,assuming $v \in V(G)$ and $d(v) = 2$,note the two vertexes adjacent to Vertex *V* as x and y,draw Graph G'' $= G - \{v\} + \{xy\}$,then G'' satisfy $\varepsilon(G'') \ge v(G'')+7$,which is also contrary to the assuming of *G*.

The following will prove: When $v(G) \le 9$, graphs satisfy the conditions(i), (ii)and $\varepsilon(G) \ge v(G)+7$, and such graphs without three disjoint cycles do not exist.

Knowing from (i) that only simple graphs need to be considered. In addition, from (i), (ii) and $\varepsilon(G) \ge v(G)+7$, we know that it's only necessary to discuss the case of $v(G) \ge 8$. Otherwise, it's easy to know that *G* will be contrary to one of the conditions (i),(ii) or $\varepsilon(G) \ge v(G)+7$. Here we only to prove the case of $v(G) = 7$, the case of $v(G) < 7$ can be proven in the same way. In fact, we take cycle C with the length of 4 from *G*, i.e. $C: v_1v_2v_3v_4v_1$, shown as Figure 1. From (ii) we can know: in connection to each vertex $v_i (i = 1, 2, 3, 4)$,outside *C*, there must be a point adjacent to it. When *G* satisfy the conditions (i),(ii), under the meaning of isomorphism, the adjacent relationship of its vertexes can have the only shape shown as Figure 1. However, obviously, Figure 1 does not meet the condition of $\varepsilon(G) \ge v(G)+7$.

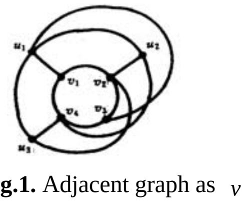


Fig.1. Adjacent graph as $v(G) = 7$

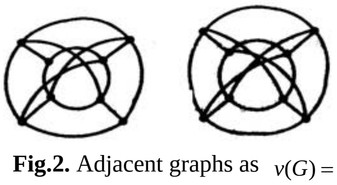


Fig.2. Adjacent graphs as $v(G) = 8$

Under the following two cases, from condition (ii)we can know: for each case,the girth of *G* can not be bigger than 5. In the meanwhile, it's easy to know that the girth of *G* cannot be equal to 5 either. Here we only to prove the case with $v(G)=8$, the case with $v(G)=9$ can be proven in the same way. In fact, if the girth of *G* is equal to 5,we take a cycle with the length of 5 from *G*, i.e. Cycle $C: v_1v_2v_3v_4v_5v_1$. From(ii)we can know, in regard to each $v_i(i=1,2,3,4,5)$,outside *C*, there must be a point adjacent to it; From $v(G)=8$,we can know that there are only the three points u_1, u_2, u_3 available to be adjacent to $v_i(i=1,2,3,4,5)$, which are contrary to *G* girth of 5. Therefore, when $v(G)=8$, the girth of G cannot be equal to 5. Further more, from condition (i) we can know, the girth of G can only be 4. Thus, in the following two cases, *G* girth can only be 4.

1) Suppose $v(G)=8$.

Here, the case of $\varepsilon(G)<15$ does not need to be considered, because *G* does not satisfy $\varepsilon(G)\geq v(G)+7$ in this case, and the case of $\varepsilon(G)>16$ does not need to be considered either. Because in such case *G* does not satisfy the condition(i), therefore, it's only need to investigate the graph in the two cases of $\varepsilon(G)=15$ and $\varepsilon(G)=16$. Under the meaning of isomorphism, adjacent graphs satisfying conditions(i),(ii)and $\varepsilon(G)\geq v(G)+7$ can only be the two cases shown as Figure 2. In the two graphs, there are three disjoint cycles for each. Thus, when $v(G)=8$, graphs without three disjoint cycles do not exist.

2) Suppose $v(G)=9$.

For the same reasons, the cases of $\varepsilon(G)<16$ and $\varepsilon(G)>20$ do not need to be considered. It's only need to investigate the adjacent graphs in the cases of $16\leq\varepsilon(G)\leq 20$. Under the meaning of isomorphism, adjacent graphs satisfying conditions(i), (ii) and $\varepsilon(G)\geq v(G)+7$ can only be the 8 cases shown as Figure 3. However, in the 8 adjacent graphs, there are three disjoint cycles for each. Thus, when $v(G)=9$, graphs without three disjoint cycles do not exist.

Analyzing Figure 4, this graph $v(G)=10$, which is easy to know that satisfy the conditions(i), (ii)and $\varepsilon(G)\geq v(G)+7$, however, this graph does not have three disjoint cycles.

(II). Next to prove when *G* satisfy $\varepsilon(G)\geq v(G)+6$ and $v(G)\leq 7$, it has at least three disjoint cycles.

Suppose *G* is a graph satisfy the conditions of $\varepsilon(G)\geq v(G)+6$ and $v(G)\leq 7$, but it does not have three disjoint cycles and it has min. vertexes. Similar to Conclusion (I), *G* can be proven to have the following properties:

(i)*G* girth: $g\geq 3$;

(ii)Minimum degree of *G*: $\delta(G)\geq 3$.

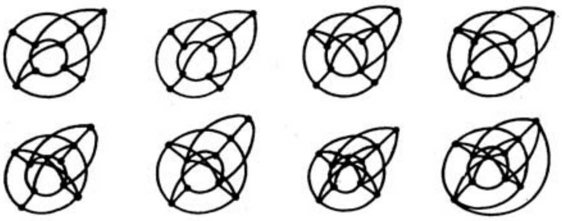


Fig.3. Adjacent graphs as $v(G)=9$

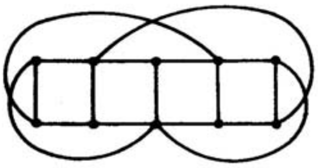


Fig.4. $\varepsilon(G)\geq v(G)+7$ & $v(G)=10$,
The graph without three disjoint cycles

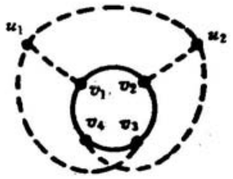


Fig. 5. Adjacent graph with girth of 6 as $v(G)=6$

Similar to the Conclusion (I), it is only necessary to investigate various adjacent graphs in the two cases of $v(G)=6$ and $v(G)=7$.

Next to prove when $v(G)=6$ and $v(G)=7$, graphs without three disjoint cycles, satisfying the conditions of (i),(ii)and $\varepsilon(G)\geq v(G)+6$, do not exist. Therefore, firstly to prove: in the two cases, the girth of G can only be 3.

In fact, from the condition (ii)we can know, *G* girth in each case can not be bigger than 4. In addition, it's easy to know that the girth of *G* cannot be equal to 4 either. Here we only to prove the case of $v(G)=6$, the case of $v(G)=7$ can be proven in the same way. In fact, we take cycle C with the length of 4, i.e. $C: v_1v_2v_3v_4v_1$,known from(ii), in connection to each $v_i(i=1,2,3,4)$, outside *C*, there must be a point adjacent to it. And known from $v(G)=6$, there are only the two points u_1 and u_2 available to be adjacent to $v_i(i=1,2,3,4)$. Under the meaning of isomorphism, there is the only adjacent shape shown as dot line in Figure 5. However, Figure 5 does not satisfy $\varepsilon(G)\geq v(G)+6$. Therefore, when $v(G)=6$, the girth of *G* cannot be equal to 4. Furthermore, known from the condition(i), the girth of *G* can only be 3. Thus, in the following two cases, *G* girth can only be 3.

1) Suppose $v(G)=6$.

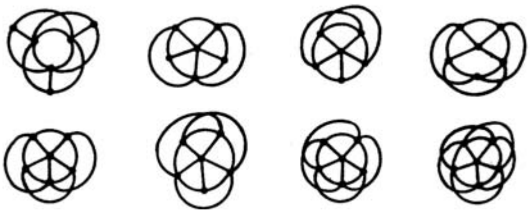


Fig.6. Adjacent graph as $v(G)=6$

Similar to (I), it can be proven that it is only necessary to investigate adjacent graphs in various cases of $12 \le \varepsilon(G) \le 15$. Under the meaning of isomorphism, adjacent graphs satisfying the conditions of (i),(ii)and $\varepsilon(G) \ge v(G)+6$ are only the 8 cases shown as Figure 6. But in the 8 adjacent graphs, all contain three disjoint cycles. Thus, when $v(G)=6$, graphs without three disjoint cycles do not exist.

2) Suppose $v(G)=7$

For the same reasons, it is only necessary to investigate various adjacent graphs in the cases of $13 \le \varepsilon(G) \le 21$. Under the meaning of isomorphism, there are only 72 adjacent graphs satisfying the conditions of (i), (ii)and $\varepsilon(G) \ge v(G)+6$, but the 72 adjacent graphs all have three disjoint cycles. Thus, when $v(G)=7$, graphs without three disjoint cycles do not exist either.

Analyzing Figure 7, for this graph $v(G)=8$, it is easy to know that it satisfies the conditions of (i),(ii)and $\varepsilon(G) \ge v(G)+6$,but it does not have three disjoint cycles.

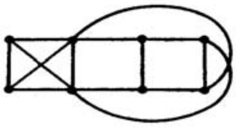


Fig.7. $\varepsilon(G) \ge v(G)+6 \,\&\, v(G)=8$
Graphs without three disjoint cycles

3. CONCLUSIONS

Studying graphs with disjoint cycles has a comprehensive practical application in some fields such as traffic, network circuit, etc. In addition, it is important to theoretical study in the graph theory. The issue proposed in[5] has been answered in the paper, which has some impacts on famous Hamilton problems and great applications in Computer Network.

REFERENCES

[1] N. Biggs, *Algedbraic Graph Theory*. Cambridge University Press, 1974.

[2] J. A. Bondy, U. S. R. Murty, *Graph Theory with Applications*, MacMillan Press LTD, 1976.

[3] Weixuan Li, *Graph Theory*. Changsha: Hunan Publishing Company of Science and Technology, 1980.

[4] Ruijun Zhu,”Graphs with Three Disjoint Circles,” *Journals of Xinjiang University (Edition of Natural Science)*, Vol. 5, No.1, pp.35~39, Jan.1988.

[5] Fugui Liu, Luchen Xie, “Graphs with Three Disjoint Cycles,”*Journals of Wuhan Transportation University*, Vol.20, No.5, pp.618~620, Oct. 1996.

Boot Loader Design of Video Surveillance System in Windows CE 5.0

Xiaofeng Wan, Wenli Huang
Information Engineering School, Nanchang University
Nanchang, Jiangxi 330031, China
Email: xfwan@ncu.edu.cn

ABSTRACT

The video surveillance systems are applied to industry and living widely. A video surveillance system based on embedded processor and embedded operating system is developed. In this system, S3C2410X is adopted as CPU and Windows CE 5.0 is selected as RTOS. This paper gives a brief introduction of the hardware platform and BSP. Designing an appropriate boot loader, which is a key step to develop an embedded system, is given in detail. The boot loader, the hinge connects hardware and OS, is made up of OEM startup code and main code mostly. At the same time, some configuration files should be edited, such as sources, makefile, Boot.bib and Dirs etc. Validated by experiment, author can run the boot loader in the video surveillance system successfully, it can help to debug the system and download the OS image, and the method described in this text to develop a boot loader is available.

Keywords: Video Surveillance, ARM920T, S3C2410X, WinCE 5.0, BSP, Boot Loader

1. INTRODUCTION

With the rapid development in storage capacity, network bandwidth, processor performance and various video compression technology, video surveillance system cyberizes and digitizes day by day. According to Mark Kirstein, the vice president of Multimedia Content and Services for iSuppli, revenue from shipments of video surveillance nearly doubled in 2005, and will continue to grow at a compound annual growth rate (CAGR) of 87.9 percent from 2004 to 2010, to reach $3.9 billion [1].

Windows CE 5.0, which is an outstanding RTOS, has more market share in wireless communication, industrial control, and consumer electronics. Microsoft releases the Platform Builder development tool and CETK test tool for Windows CE, and provides different reference modules for different CPUs. Now numerous third-part developers provide the PDK (Peripheral Development Kit) for Windows CE. As a result, it is easy and quick to develop with the help of Windows CE. And designing a boot loader with perfect function is the first step and a decisive step to develop a system. Combining an example, this paper will introduce how to develop a boot loader suitable for the hardware provided.

2. HARDWARE PLATFORM

The hardware platform design is the first step to develop the whole video surveillance system. And the main function of the system is image acquisition, video compression, digital image output, give an alarm and control the pan. The hardware platform includes some modules to carry out these function, such as CPU module, video input module, audio input module, audio video output module, pan control module, FLASH, RAM, and debug module. The hardware frame is shown in Fig.1.

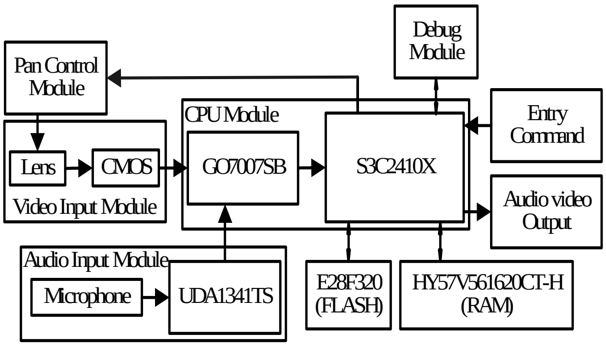


Fig.1. Hardware Frame

A brief introduction of the CPU module is as follows:
CPU module which is made up of S3C2410X and GO7007SB is the core of the video surveillance system. Both of them connect with each other by HPI (Host Peripheral Interface). The main function is controlling and managing the whole system, and processing image. S3C2410X is a chip based on the ARM920T core, its frame is shown in Fig.2 [2].

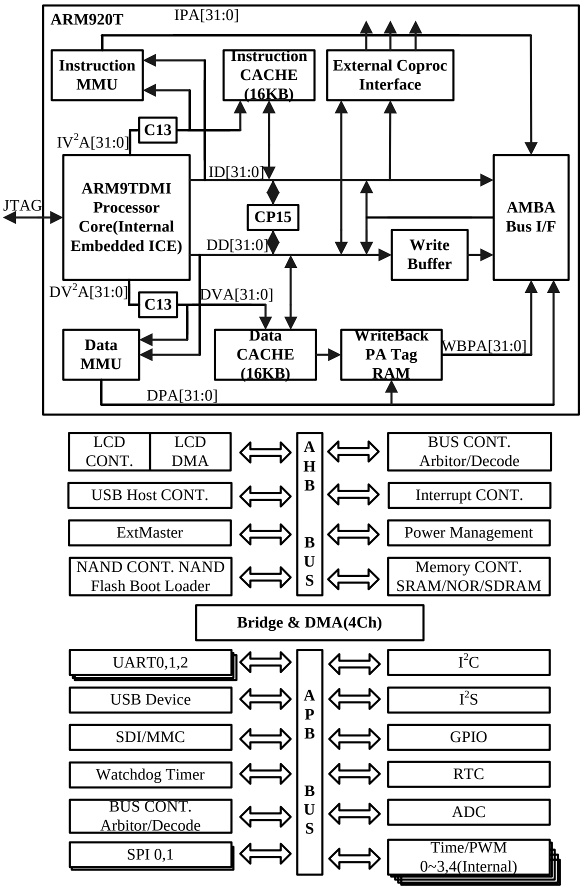


Fig.2. S3C2410X Block Diagram

3. BOARD SUPPORT PACKAGE

BSP (Board Support Package) is a software system, abstracts the interface between hardware and operating system, strictly speaking, it should be one part of operating system.

In Windows CE, BSP is made up of 4 parts as follows [3]:

(1) OAL (OEM Abstraction Layer) whose code depends on the hardware much is the interface between hardware and operating system kernel.

(2) Boot loader is implemented in hardware develop board. Its main function is to initialize hardware, to load OS image into memory, and to jump to implement operating system.

(3) Configuration file contains some configuration information which usually has something to do with the OS image or the source code, such as .bib, .db, .reg and .dat files etc.

(4) Device driver is a software component that permits operating system to communicate with a device. In most cases, the driver also manipulates the hardware in order to transmit the data to the device..

In general, developing the BSP can be divided into 7 steps, as shown in Fig.3.

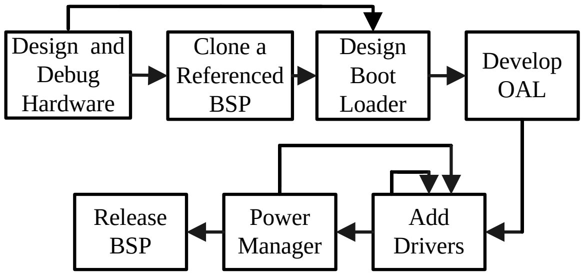


Fig.3. Steps of Develop BSP

4. BOOT LOADER

In terms of static, the boot loader is made up of Blcommon, OEM code, Eboot, memory management and EDBG drivers etc, which is shown in Fig.4. The Blcommon is a universal frame of boot loader; The OEM code depending hardware is an initialization program, the code contains some functions with the capital OEM initially, and this part is the most difficult part of the boot loader. The Eboot is the Ethernet function, such as UDP, DHCP, TFTP program etc. The memory management is the distribution management program that used for permanently storage, EDBG is a driver to debug network card [4].

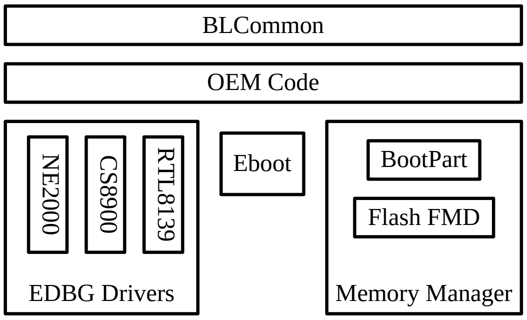


Fig.4. Boot Loader Structure

The boot loader of Windows CE has 3 functions:

(1) Initialization hardware: include initialization memory, interrupt controller, clock and MMU etc.

(2) Control startup: the boot loader usually provides a simple alternate menu for user, letting the user choice startup process.

(3) Download and execute OS image.

Its executable sequence is shown in Fig.5:

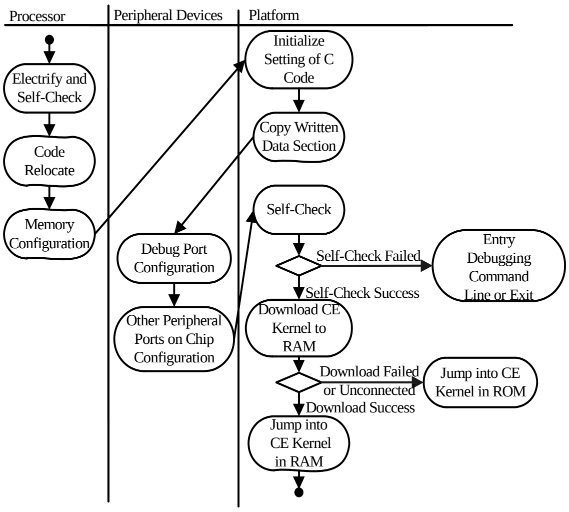

Processor
Peripheral Devices
Platform
Electrify and Self-Check
Code Relocate
Memory Configuration
Initialize Setting of C Code
Copy Written Data Section
Debug Port Configuration
Other Peripheral Ports on Chip Configuration
Self-Check
Self-Check Failed
Entry Debugging Command Line or Exit
Self-Check Success
Download CE Kernel to RAM
Download Failed or Unconnected
Jump into CE Kernel in ROM
Download Success
Jump into CE Kernel in RAM

Fig.5. Boot Loader Executable Sequence

At first, after system is electrified, the boot loader usually does code relocation. The boot loader will make itself move from a place to another place to carry out a more convenient accessing in this step. Then, the boot loader configures memory to implement the boot loader. Immediately, after setting environment variables for executing C code, the boot loader jumps to start implementing the main function. Immediately the boot loader initializes debug ports and others, makes debug ports can be used as early as possible, helps debug boot loader itself too. Then the boot loader implements self-check, if self-check passes, OS image will be downloaded. At last, after downloading OS image successfully, system jumps to execute OS image.

Development of the Boot Loade

After system is electrified, the first Eboot function implemented is Startup( ) that is a system initialization, this function is edited using assembly language, locating on the Startup.s. The Startup ( ) function initializes CPU and other kernel logical devices. The work what Startup.s does mostly is shown as follows [5]:

1) Set CPU in supervisor mode, in this mode memory and hardware can be accessed without limits.
2) Shield CPU against all interruptions.
3) Close the MMU and TLB.
4) Make cache and write buffer invalidated.
5) Initialize memory controller.
6) Initialize other devices on chip, such as clock.
7) Set stack pointer.
8) Set and open MMU to map physical address to logical address, and open cache.
9) Copy Eboot code to RAM, jump to Eboot code in RAM.
10)Jump to Main ( ) function.

Boot loader control chart is shown in Fig.6. These functionsin the figure will be called when the boot loader is implementing.

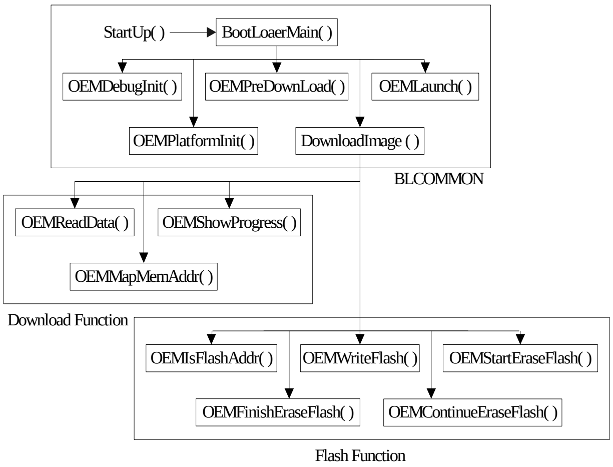


Fig.6. Boot Loader Control Chart

The next is the description of these functions:

1) OEMDebugInit: this function is the first called by the BLCOMMON framework when a boot loader starts. It initializes the debug transport, usually just initializing the debug serial universal asynchronous receiver-transmitter (UART).
2) OEMPlatformInit: use the OEMPlatformInit function to perform platform-specific initialization, such as clock, driver, and transport initialization.
3) OEMPreDownLoad: this function is called by the BLCOMMON framework prior to download and can be customized to prompt for user feedback, such as obtaining a static IP address or skipping the download and jumping to a flash-resident run-time image.
4) DownloadImage: this method starts the download of a run-time image to a target device.
5) OEMLaunch: the OEMLaunch function collects post-download connection information from Platform Builder and jumps to the newly downloaded boot loader image
6) OEMReadData: this function reads data from the transport during the download process.
7) OEMShowProgress: this function shows visual information, on a LED, for example, to let users know that the download is in progress. It is called as the download progresses.
8) OEMMapMemAddr: this function handles downloads that are destined for flash. OEMMapMemAddr remaps a flash-resident address to a unique RAM-based address so that flash memory OS images can be temporarily cached in RAM while the download occurs. This provides enough time to handle the flash memory update while not stalling the download process because the flash memory operation typically takes more time than the download process.
9) OEMIsFlashAddr: this function determines whether the address provided lies in a platform's flash memory or RAM address range. Based on the results of this call, the OS image being downloaded might be destined for flash memory. In that case, it might need to be handled differently, for example, by providing a RAM-based file cache area to support downloading while a flash memory update takes place.
10) OEMWriteFlash: this function writes to flash memory the OS image that might be stored in a RAM file cache area.
11) OEMStartEraseFlash: called when the download process begins, and provides the overall run-time image start address and total run-time image length. The run-time image start address and total run-time image length specify the overall flash memory address range to be erased for the new run-time image.
12) OEMFinishEraseFlash: called after the run-time image is completely downloaded and allows the boot loader to finish all flash memory erase operations.
13) OEMContinueEraseFlash: called on every run-time image .bin record download to allow the boot loader to continue the flash memory erase operation.

This is an important part of boot loader, but it depends on the hardware performance much, as a result I will not give further description.

In debugging stage, the menu provided by boot loader basically implement corresponding operation. Providing more menus can simplify operation and test the platform better. The menus, which my boot loader provides, are shown as follows.

1) IP address, subnet mask
2) Boot delay
3) DHCP
4) Reset TOC to default
5) Startup image
6) Program RAM image into Boot Media
7) Mac address
8) Kernel Debugger
9) Format Boot Media for BinFS
B) Support BinFS
D) Download image now
F) Low-level format boot media
L) Launch existing boot media image
R) Read configuration
U) Set UDID
W) Write configuration right now
X) Download image to boot media, and then launch it off the media.

4.2 Configure and Create Boot Loader

The source code of the boot loader is made up of code editing in C language and assembly language, Platform Builder can be used to compile these code. Therefore, edit the sources file, makefile file, Boot.bib file and Dirs file.

Both the sources file and makefile file control the compiler of the boot loader, makefile file drives compiler process, and compiler information comes from the sources file.

The sources file defines the method of compiling code. Compiler decides how to compile and link through the sources file. The sources file code mainly includes some macros shown as follows:

```
TARGETNAME=EBOOT
TARGETTYPE=PROGRAM
WINCECPU=1
RELEASETYPE=PLATFORM
WINCETARGETFILES= VideoSurveillance
EXEENTRY=StartUp
EXEBASE=0x80000000

INCLUDES=$(INCLUDES);$(_PUBLICROOT)\common\oak
\drivers\block\msflashfmd\inc
INCLUDES=$(_TARGETPLATROOT)\Drivers\NandFlsh\FM
D;$(INCLUDES)
ADEFINES=-pd "ALLOCATE_TABLE SETS \"FALSE\""
$(ADEFINES)
LDEFINES=-subsystem:native /DEBUG /DEBUGTYPE:CV
```

```
/FIXED:NO
CDEFINES= $(CDEFINES)
-DPPSH_PROTOCOL_NOTIMEOUT -DCOREDLL
-DPLAT_ONBOARDEDBG=1
-DBOOT_LOADER=1
-DNOSYSCALL=1

TARGETLIBS=\
$(_COMMONOAKROOT)\lib\$(_CPUINDPATH)\fulllibc.lib
\
$(_COMMONOAKROOT)\lib\$(_CPUDEPPATH)\
oal_blcommon.lib\
$(_PLATCOMMONLIB)\$(_CPUINDPATH)\oal_memory_s3
c2410x.lib\
$(_PLATCOMMONLIB)\$(_CPUINDPATH)\oal_cache_s3c2
410x.lib\
$(_PLATCOMMONLIB)\$(_CPUINDPATH)\oal_rtc_s3c2410
x.lib\
$(_COMMONOAKROOT)\lib\$(_CPUDEPPATH)\eboot.lib
\
$(_COMMONOAKROOT)\lib\$(_CPUINDPATH)\cs8900dbg
.lib\

SOURCES=\
       startup.s          \
       util.s             \
       fw.s               \
       mmu.s              \
       main.c             \
       debug.c            \
       ether.c            \
       flash.c            \
       bitmap.c           \
       e28f320.c          \
       hy57v561620ct      \
       ecc.c              \
       fmd.cpp            \
       oemboot.c          \
       time.c
```

1) TARGETNAME: this macro definition specifies the name of the .exe or .lib file being built, excluding the file name extension. In this example, Build.exe creates a file named EBOOT.extension, where extension can be .lib, .dll, or .exe.
2) TARGETTYPE: this macro definition specifies the type of file being built. Hence, the extension in that file's name. PROGRAM generates an .exe file.
3) WINCECPU: this macro definition specifies whether your target is dependent on a specific CPU. Build.exe uses WINCECPU to locate the release target and determine whether it is out of date. Build.exe also deletes the release target if it is performing a clean build. In this example, WINCECPU is set to 1; Build.exe builds the target and places it in the CPU subdirectory specified by _TGTCPU. This results in multiple binaries being built for different CPUs even if the binaries are of the same CPU type. This setting should only be used by low-level software like the kernel, or portions of the HAL.
4) RELEASETYPE: this macro definition sets two flags: RELEASEDIR and RELEASELIBDIR, which specify which output directory to place binaries and libraries in after they are built. In this example, PLATFORM is for code inside platform-specific projects.
5) WINCETARGETFILES: this macro definition specifies nonstandard target files that Build.exe should build after Build.exe links all other targets in the current directory. The example shows a sources file that adds the nonstandard target file VideoSurveillance to the list of dependencies.
6) EXEENTRY: this macro definition specifies the function that is used as an entry point for the .exe file if TARGETTYPE is set equal to PROGRAM. In this example, Startup means the entry point for a boot application.
7) EXEBASE: this macro definition specifies the base address for an .exe file in memory, which is the location in memory that the executable must load. In this example, it is 0x80000000.
8) INCLUDES: this macro definition specifies additional paths used to find included .h files.
9) ADEFINES: defines the assembly language files to be built.
10)LDEFINES: defines the flags specific to the linker.
11) CDEFINES: this macro definition specifies compiler define statements.
13) TARGETLIBS: this macro definition specifies import or .lib files that must be linked.
14) SOURCES: this macro contains a list of sources files with extensions. These sources files become the contents of the .lib file or .dll file.

The Boot.bib details the memory layout that the boot loader will use and defines the structure of the boot loader. Romimage.exe uses this file to convert the boot loader executable (.exe) file into .bin and .nb0 files. It mainly includes several parts shown as follows:

1) MEMORY: reserves MEMORY regions by name.
2) CONFIG: specifies attribute
3) MODULES: specifies which Windows CE-based modules are included in the run-time image, and how they are loaded into the memory table as established in the MEMORY section of the Boot.bib file.

The dirs file is a text file that specifies the subdirectories that contain source code to be built.

The boot loader is compiled in command line, it can be divided into 3 steps:

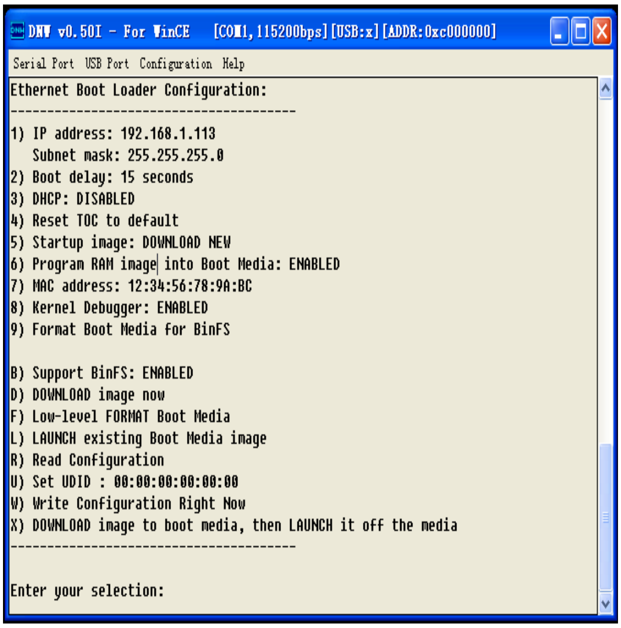
1) Set command line parameters. In this step, execute Wince.bat to confirm the target CPU and a corresponding BSP.
2) Run sysgen. This step builds static libraries for compiling boot loader.
3) Compile the code of boot loader

Two image files, Eboot.nb0 and Eboot.bin, will be gotten.

5. CONCLUSIONS

Just as shown in Fig.7, basing on Windows CE 5.0 embedded operating system and SUMSUNG S3C2410x processor boot loader have already worked in the video surveillance system successfully.

The boot loader depends on CPU performance, the peripherals devices and the operating system adopted much, as a result, the design process above isn't fixed and unchangeable, appropriate modification is necessary in other practical application.


DNW v0.50I - For WinCE [COM1,115200bps][USB:x][ADDR:0xc000000]
Serial Port USB Port Configuration Help

```
Ethernet Boot Loader Configuration:
----------------------------------------
1) IP address: 192.168.1.113
   Subnet mask: 255.255.255.0
2) Boot delay: 15 seconds
3) DHCP: DISABLED
4) Reset TOC to default
5) Startup image: DOWNLOAD NEW
6) Program RAM image into Boot Media: ENABLED
7) MAC address: 12:34:56:78:9A:BC
8) Kernel Debugger: ENABLED
9) Format Boot Media for BinFS

B) Support BinFS: ENABLED
D) DOWNLOAD image now
F) Low-level FORMAT Boot Media
L) LAUNCH existing Boot Media image
R) Read Configuration
U) Set UDID : 00:00:00:00:00:00
W) Write Configuration Right Now
X) DOWNLOAD image to boot media, then LAUNCH it off the media
----------------------------------------

Enter your selection:
```

Fig.7. Boot Loader by UART

REFERENCES

[1] Mark Kirstein, “Video Surveillance: Migrating to IP Cameras and Networked Systems,” An iSuppli Report, January 2006.

[2] *Samsung Electronics.21.2-S3-C2410X-052003 S3C2410X 32-Bit RISC Microprocessor User's Manual*, Revision 1.2. Korea: Samsung Electronics Co., Ltd, 2003.

[3] He Zongjian, *Windows CE Embedded System*, Beijing: Beihang University Press, 2006.

[4] Zhang Dongquan, Tan Nanlin et al, *Applied developable technology of Windows CE*, Beijing: Publishing House of Electronics Industry, 2006.

[5] Microsoft Corporation. *Platform Builder for Microsoft Windows CE 5.0 Help.*

Xiaofeng Wan, female, is a professor and vice director of Electrical and Automation Department in School of Information Engineering, Nanchang University. She graduated from Zhejiang University in 1994 with Master's degree. From Sep 2002 to Jul 2003 she went to the Huazhong University of Science and Technology as a senior visitorial scholar, during the next semester, she was a visiting scholar of the Institute Senior Electronics Paris. She has published two books, over 20 Journal papers. She undertakes teaching and scientific research, and her research interests are in computer control, embedded control, and system integration.

Wenli Huang, born in 1983, is a graduate student in Nanchang University, his major is control theory and control engineering, and his research interests are computer control and embedded intelligent instrument.

Isomorphism of the Graph

Huaan Wu
School of Sciences, Wuhan University of Technology
Wuhan, Hubei Province, Peoples Republic of China
Email: huaan_wu@yahoo.com.cn

ABSTRACT

This paper describes two isomorphic invariants for graph theory. One of them is the adjacent matrix, symmetric matrix, defined on given graph. We have shown that any given graphs having the dual congruent adjacent matrices must be isomorphism. Some properties about adjacent matrix are obtained in the paper that is for any graph *G* with vertices $\{v_1, v_2, \cdots, v_n\}$, the *i*th row sum of the adjacent matrix defined on *G* is equal to $\deg(v_i)$. If two adjacent matrices are dual congruent, then they have the same eigenvalues.
The fundamental group is used as another invariant. The net is first defined as a connected graph with $\deg(v) \geq 2$ for any vertex of the graph. It is proved in the paper that any connected graph can be collapses simplicially to a net, the graph and the net have the same simple homotopy-type, therefore they have the same fundamental group, free group. by the means above, we got a necessary condition that isomorphic graphs have the same fundamental group.

Keywords: Graph, Isomorphism, Simplicial, Complex, Fundamental Group

1. MOTIVATION

In the istributed parallel algorithms and program design, we are frequently get into the graph with respect to the nodes and paths. gram design of computer, it is known that one important omission in graph theory is to distinguish two given graph. For given two graphs, if we can deformation from one to other preserving the numbers veritices and edges that is they are not essentially different, we say that the graphs are isomorphic. Because any graph is concerned only with the numbers of its veritices and edges, besides the location of veritices and the length of edge. Intuitively, two graphs are isomorphic if and only if one graph is a copy of the other. Unfortunately, when given two graphs, it is hard to decide whether they are isomorphic. That is to say that it was not an easy matter to find give an entirely satisfactory and precise way of treating this problem. It is difficult to see that whether two graphs are same in geometric method.

Consider revering partly, some isomorphic invariants are introduced, the properties of graph that are preserved under isomorphic, such as connectivity, the numbers of vertices etc. The isomorphic graphs must have the same invariants, if one graph has the invariant in question and another does not, then the two graphs cannot be isomorphic. But the converse is false, because the invariant is just only a necessary condition for isomorphic graphs.

2. PREPARATION OF MANUSCRIPTS

Definition 3 Let *A* be a $n \times n$ -matrix, taking integers i, j ($1 \leq i, j \leq n,\ i \neq j$), interchanging the *i*th, *j*th rows in *A* and the same manipulation is performed for *i*th, *j*th columns in *A* at the same time , we call the operation by dual congruent.

A natural question to ask is what happens if we interchanging any given two vertices.

Let $V = \{v_1, \cdots, v_i, \cdots, v_j, \cdots, v_n\}$ be the set of vertices of graph *G*, and the set $V' = \{v'_1, \cdots, v'_i, \cdots, v'_j, \cdots, v'_n\}$ is obtained by reordering elements of *V*, such that

$$v'_k = \begin{cases} v_k, & k \neq i, j, \\ v_j, & k = i, \\ v_i, & k = j. \end{cases}$$

Assume that *A* and A' be the adjacent matrices of *G* with respect to *V* and V' respectively,

Theorem 2 Suppose that *A* and A' be defined as above, then $A' = E(i,j)AE(i,j)$, Where $n \times n$ matrix $E(i,j)$ is obtained by interchanging rows *i* and *j* of the identity matrix, i.e. A' can be achieved by forming a dual congruent to *A*.

Proof Let $A = (a_{st})$, $A' = (a'_{st})$. Compare the elements of matrix *A* with the those of A'. By using the definition of adjacent, for the element of *s*th row, *t*th column in *A*, we have that

If $s \neq i, t \neq j$, then $a_{st} = |[v_s, v_t]| = |[v'_s, v'_t]| = a'_{st}$, when $s \neq t$, or $a_{ss} = 2|[v_s, v_s]| = 2|[v'_s, v'_s]| = a'_{ss}$, when $s = t$.

If $s = i, t \neq i, j$, then

$$a_{st} = a_{it} = |[v_i, v_t]| = |[v'_j, v'_t]| = a'_{jt}.$$

If $s = j, t \neq i, j$, then

$$a_{st} = a_{jt} = |[v_j, v_t]| = |[v'_i, v'_t]| = a'_{it}.$$

If $s = t = i$, then $a_{st} = a_{ii} = 2|[v_i, v_i]| = 2|[v'_j, v'_j]| = a'_{jj}$, similarly, $a_{jj} = a'_{ii}$

If $s = i, t = j$ or $s = j, t = i$, we deduced that $a_{ii} = a'_{jj}$ or $a_{jj} = a'_{ii}$ respectively.

It is obviously that $A' = E(i,j)AE(i,j)$, this completes the theorem.

Theorem 3 Let G_1 and G_2 be graphs, A_1 and A_2 are adjacent matrices of G_1 and G_2 respectively, then $G_1 \cong G_2$ if and only if any one matrix of the two matrices can be achieved by other via a finite dual congruent.

Proof The necessity of the condition has shown in the theorem 2. It suffices to show the sufficient. Without of a generality,say A_2 is obtained interchanging 1th and 2th in A_1 via one dual congruent . Letting the vertex sets V_1 and V_2 of G_1 and G_2 are expressed as respectively as follow

$$V_1 = \{v_i \mid 1 \leq i \leq n\}, \qquad V_2 = \{u_i \mid 1 \leq i \leq n\}$$

adjacent matrices with respect to the vertex sets are

$A_1=\left(a_{ij}\right)_n$ and $A_2=\left(b_{ij}\right)_n$ respectively. It is clear that we can reorder the vertices of G_2 such that $V_2^*=\{u_i' \mid 1\le i\le n\}$, where

$$u_i'=\begin{cases}u_2, & i=1,\\ u_1, & i=2,\\ u_i, & \text{otherwise}.\end{cases}$$

then the mapping $f:\ V_1\to V_2^*$ defined by $f(v_i)=u_i',\ i=1,\ 2,$ $\cdots,n$, must be an isomorphic from G_1 to G_2. The theorem is proved.

So far for given two graphs, we have transformed the question that adjust whether they are isomorphic to the linear algebra one. It turns out discuss the isomorphic problem by using the algebraic method. However, sometime it is a rigorous process provided by the theorem 3, since the cardinality vertex set which we face to is usually large. Hence it is useful to write a evaluate processing. To avoid a rigorous evaluate, we could use the property given by the following.

Theorem 4 If two matrices *A* and *B* are dual congruent, then their characteristic polynomials coincide.

Proof Since the relation of dual congruent between graphs is symmetrical, it is sufficient to show that the conclusion hold when *A* is dual congruent to *B*. Let *A*, *B* be $n\times n$ matrices, $B=E(i,j)AE(i,j)$, $Bx=\lambda x$, where nonzero vector *x* be an eigenvector of *B*, corresponding to the eigenvalue λ, then we have $E(i,j)AE(i,j)x=\lambda x$. Note that the matrix $E(i,j)$ is nonsingular and $E^{-1}(i,j)=E(i,j)$, so $AE(i,j)x=\lambda E(i,j)x$. $E(i,j)x\ne 0$
since $x\ne 0$. The result showed that λ is an eigenvalue of *A*.
Example The two graphs given below arenonisomorphic.

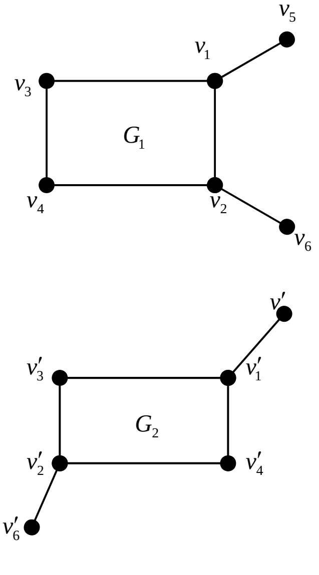


Fig.1. Nonisomorphism graphs

Solution The adjacent matrices of G_1 and G_2 are

$$A_1=\begin{pmatrix}0&1&1&0&1&0\\1&0&0&1&0&1\\1&0&0&1&0&0\\0&1&1&0&0&0\\1&0&0&0&0&0\\0&1&0&0&0&0\end{pmatrix}$$

$$A_2=\begin{pmatrix}0&0&1&1&1&0\\0&0&1&1&0&1\\1&1&0&0&0&0\\1&1&0&0&0&0\\1&0&0&0&0&0\\0&1&0&0&0&0\end{pmatrix}$$

respectively. It is easily verify that the characteristic polynomials of A_1 and A_2 is

$$p_{A_1}(\lambda)=\lambda^6-6\lambda^5+5\lambda^2+1,$$
$$p_{A_2}(\lambda)=\lambda^2(\lambda^2-1)(\lambda^2-5).$$

G_1 is not isomorphic to G_2, since $p_{A_1}(\lambda)\ne p_{A_2}(\lambda)$,

If we regard any graph as a finite simplicial complex in R^3, then some invariants about topology Space can be used in the graph theory. In this way, it is easily to decide whether the given graphs isomorphism.

Definition 4 Let $G=<V,E>$be a graph, *v* an vertex of *G* with $\deg(v)=1$, $[u,v]\in E$. Assume that $G'=G-\{(u,v]\}$, where the set $(u,v]$ denotes all points on the curve linking the points *u* and *v* but *u* is wiped out, then we say *G* elementary simplicial collapses to G'. If a graph G' is obtained by a finite sequence of elementary simplicial collapses from graph *G*, then we say that *G* collapses simplicially to G', *G* and G' have the same simple-homotopy type.

Any graph that is connected and $\deg(v)>1$ (*v* any vertex of the graph) is called net. Obviously, any graph and the net obtained via simplicial collapses from *G* have the same simple-homotopy type. We known that two graphs have the same simple homotopy- type must have the same homotopy group, specially fundamental group. It is easy to evaluate the fundamental group of the net, so we got an isomorphic invariant for net as below.

Suppose *G* be a net, *v* is any given vertex of *G*, *n* is Euler characteristic of *G*, the $\pi_1([G],v)$ is isomorphic to free group with *n* generators. Where $[G]$ indicate the union of the sets ofall edges and vertices.

3. CONCLUSIONS

We have got two main results in the paper.One is the theorem 3. It is said that two given graphs are isomorphic to each other, sufficient and necessity condition is their adjacent matrices isomorphism, by a finite interchanging vertices pair.

The other is about the net, two graphs are isomorphic if they can be simplicial collapses to a net, so their fundament group isomorphism. Although the problem discussing here is pure mathematic, but it is useful obviously in distributed parallel algorithms, since the node and path in the algorithms could be evidently due to the graph.

REFERENCES

[1] W.T.Tutte,*Graph Theory*. California: Addison-Wesley Publishing Company Pub., 1984.

[2] M.M. Cohen, *A Course in Simple Homotopy Theory*, New York: Springer-Verlag. Pub., 1973.

[3] Liu yuzhen, Liu yong mei, D*iscrete Mathematics*, Wuhan: Wuhan University Press.Pub., 2003.

[4] I. M. Singer, J.A. Thorpe, *Lecture Notes on Elementary Topology and Geometry*, New York: Spinger-Verlag. Pub., 1967.

The Application of Fuzzy Theory to Color Image Filtering *

Ruihua Lu' Li Deng
School of Electronics and Information Engineering, Southwest University
Chongqing 400715, China
Email: chenlr@swu.edu.cn

ABSTRACT

Aiming at multichannel (and color in particular) image filtering, two fuzzy filtering techniques are proposed: one of them is crude fuzzy, and the other is fuzzy inferential. The cardinal idea of the techniques is to assign weights that are decreasing with respect to the distance from each noise affected vector to the desired correct value. The crude fuzzy filtering uses no fuzzy rules and the actual weights are computed based on a "membership function strengths" approach. The particular nature of the approach is given by the choice of a specific function, i.e. membership function that maps some statistical measure of the vectors (colors) within the filtering window. The fuzzy inferential filters are those from the fuzzy inferential ruled by Else-Action (FIRE) family. Basically, all of them depend on the use of luminance difference between various pixel pairs within the filtering window as input variables. This mechanism assigns given input data (or variable) set and activation and superposition of each fuzzy rule. The input of system is obtained by inverse fuzzification. The fuzzy techniques can process effectively non-deterministic and fuzzy information .Simulation experiments show that above-mentioned two fuzzy techniques are efficient and the performance of fuzzy inferential filters is better than that of crude fuzzy filters.

Keywords: Color Image Filtering, Fuzzy Theory, Crude Fuzzy, Fuzzy Inferential, Inverse Fuzzyfication, Inherent Color Space Fuzziness, Grey Level Approach.

1. INTRODUCTION

The term "multichannel image" denotes that every sample of the image signal is characterized via a vector set, each item of which is considered as a component of the sample. The simple "stack of scalars" model of the multichannel sample is not proper. In particular, the independent component processing with the mentioned method can not take into account the existing correlation between the signal components, which causes artifacts, i.e. false colors, in the case of color images [1].

As the correlation between the signal components is the reason for producing artifacts through the independent component processing, scientists come up with an idea of removing the dependency by classical decorrelation techniques, such as the Karhumen-Loève transform[2]. However, this method is not widely applied because of some problems related with the computational complexity of the decorrelation transform and its intense dependencies on images from a certain class. Thus , the vector processing attracts extensively scientists' attention.

The most common filtering method of both scalar and multichannel images is the sliding (or moving) window technique. A planar shape (subsequently called filtering mask) scans the whole 2D structure; in each position it selects some pixel values which will be combined to turn out the new value of the same spatial location in the filtered image. One of the most generic processing paradigms is the weighted linear combination of the selected values and their order statistics. The use of the linear combination of pixel values is equal to a frequency domain filtering [3]. This method is effective only if the weights are modified at each spatial location in accordance with the specific local vector values. The use of the order statistics yields a class of nonlinear filters, based on ordering, i.e. L-filters [2]. These filters are very effective and versatile in scalar image processing. But it is rather difficult to extend them to the case of multichannel images due to The hardness of introducing a simple, topology-preserving ordering relation for vectors [4]. However, the median statistics has been widely applied and there are several multichannel extensions, based on sub-ordering principles [5]. These extensions all start from an important paper that present the Vector Median Filtering [1].

Under the circumstances of the color images, the use of ordering is indirect, thus it is reasonable to pay close attention to the adaptive locally linear filtering. Presuming that each pixel value is expressed by a p-dimensional vector, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and that the filtering mask selects n vectors, $x_{i1}, x_{i2}, \cdots, x_{in}$, related to the pixels within the mask, then the local operation which yields the output y is expressed (at each spatial location within the image) by

$$y = \sum_{j=1}^{n} w_j x_j \ . \tag{1}$$

The weighting factors w_j in Eq.(1) are, as a rule, positive scalars which have to sum to 1 (in order to perform a smoothing, uniformity enhancing filtering operation) [5, 6]:

$$\sum_{j=1}^{n} w_j = 1 \ . \tag{2}$$

The choice of the weighting factors is made on the basis of the distribution of the selected vectors x_i in the sample space. Here the fundamental thought is to assign weights decreasing with respect to the distance from each noise-affected vector to the hoped correct value. Digital images are mappings of natural scenes (sampled and quantized slices of the 3D reality) and therefore they embed an important amount of uncertainty, in both value and location (spatial support).This uncertainty is due to the imprecise nature of pixel values and to the indetermination existing along the border regions of the image. So the color image filtering based on the fuzzy theory is proposed in the paper which consists of 3 sections except the introduction and references. Section2 presents crude-fuzzy color image filtering

Section3 describes fuzzy inferential color image filtering. In Section 4 conclusions and comments are given.

2. CRUDE-FUZZY COLOR IMAGE FILTERING

* Supported by the fund of University Science and Technology under Grant SWNU2004006.

The Crude-fuzzy method consists simply of determining some weights w_i which satisfy Eq. (2) and are "fuzzy numbers", i.e. $w_i \in [0,1]$. This normalization is gained in two steps: for each selected vector x_i some positive scalar a_i is computed according to some rules reflecting the space of the vectors x_i in the sample space, and then each a_i coefficient is normalized to their sum:

$$w_i = \frac{a_i}{\sum_{j=1}^{n} a_j} \qquad (3)$$

It is obvious that the weights computed by Eq.(3) meet Eq. (2) and are within [0, 1]. This type of filtering uses no fuzzy rules and the actual weights are calculated on the basis of the "membership function strengths" approach [5].This approach is widely applied for several classes of filters, distinguishing features of which are given by the choice of the membership function that maps some statistical measure of the vectors (colors) within the filtering window. Two main statistical measures are employed to characterize the position of a color vector with respect to a set of color vectors. They are measures based on magnitudes and measures based on angles. The use of magnitude-based measures produces various classes of filters, for example, Multichannel Distance Filter (MDF) [7], Adaptive Nonlinear Filters (ANF) [8] or Distance Dependent Multichannel Filters (DDMF) [9].

The functions in accordance with which the weight is defined have to be drably decreasing, i.e. assigning more important weights to the vectors that are closer to the center of the vector cluster. For this purpose two most popular approaches, namely the polynomial and exponential ones, are used.

2.1 The Polynomial Approach

The polynomial approach is based on the a_i coefficients calculated by the following [7,9]:

$$a_i = d_i^{-r} \qquad (4)$$

Where d_i is the statistical measure related to the color vector x_i. The d_i measure used is the aggregate Euclidean distance (the sum of Euclidean distances from the current sample x_i to all the other samples within the filtering window, given by Eq.(5)), or the distance to some fixed point (marginal median, as in the example presented in Fig.1(c)) [7], or the sum of distances to some fixed points (marginal median, average, current vector) [9] .

$$d_i = \sum_{j=1}^{n} (x_i - x_j)(x_i - x_j)^T . \qquad (5)$$

2.2 The Exponential Approach

The exponential approach mainly proposes a negative exponential function Eq. (6) and sigmoidal function Eq. (7):

$$a_i = \exp(-\frac{d_i \ln a}{\beta d_{\max}}) \qquad (6)$$

$$a_i = (1 + \exp(-d_i))^{-r} \qquad (7)$$

The Distance Dependent Multichannel Filter uses the aggregate Euclidean distance and the function given in Eq.(6) (a DDMF filtered image is shown in Fig.1 (d)). The Fuzzy Vector Directional Filter uses the sum of angles between vectors (given in Eq. (8)) as a measure of space distribution and the sigmoidal function Eq. (7) (a FVDF filtered image is presented in Fig.2 (e)):

$$d_i = \sum_{j=1}^{n} x_i \hat{x}_j . \qquad (8)$$

The trimming factors, namely the power r , the constants a, β and $d_{\max}$, are calculated in accordance with experience, empiric deduction and extensive testing ($a = 2, \beta = 0.05$ in [13]), specific data constraints ($d_{\max}$, which equals the maximum inter-color distance in [13]). These factors can be computed by adaptation also. The power r is connected to the underlying distribution of the noise superimposed on the image. The tests show that the best outcomes are gotten with $r = 1$ for uniformly distributed noise, $r = 0$ for Gaussian noise and $r = -2$ for any "long tailed" distributed noise.

There is no doubt that the filters labeled as crude fuzzy measure, in some way, the membership of each color vector into the set "correct filter output". Otherwise, these filters are unable to be considered more fuzzy than any other adaptive filter.

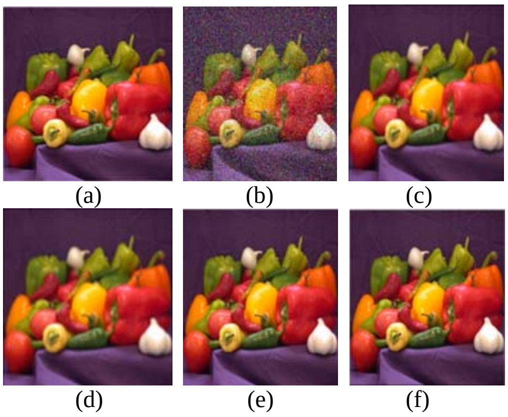

(a) (b) (c)
(d) (e) (f)

Fig.1. Results of filtering by different approaches

(a) Original true color test image;
(b) Impulse noise polluted image from Fig.(a);
(c) MDF filtered image from Fig.(b); the filter coefficients are computed by Eq.(4) with $r = 2$;
(d) DDMF filtered image from Fig.(b); the filter coefficients are calculated by Eq.(6) with $a = 2, \beta = 0.05$ and $d_{\max} = 255\sqrt{3}$;
(e) FVDF filtered image from Fig.(b);
(f) Median-like filtering of the image from Fig.(b) by using perceptual information; the color is represented in the HIS (Hue, Saturation, Intensity) color space and fuzzy rules are used for a soft decision regarding the relative importance of the components.

3. FUZZY INFERENTIAL COLOR IMAGE FILTERING

A fuzzy inferential filter combines several fuzzy associations concerning relational definitions of the objects of the universe with respect to some given linguistic notions:

$$R_i : \ if \ (v_1 \ is \ A_{1i}) \ and \ (v_2 \ is \ A_{2i}) \ and \cdots (v_n \ is \ A_{ni}) \ then \ (O \ is \ B_i).$$

Each association represents a linguistic rule (R_i). A_i and B_i are fuzzy sets mapping linguistic concepts (e.g. important, irrelevant, big, small) to each input and to the output variable in the i-th rule, respectively. The information included in the set of fuzzy rules (i.e. rule base) is numerically processed with the inference mechanism, which evaluates, for a given set of input data (or variables) v_i , the activation of each fuzzy rule and then their superposition. The output of the system (O) is gotten by defuzzification.

3.1 Direct Expansion of the Grey Level Approach

It is known that existing fuzzy inferential filtering approaches for processing grey level images are widely used [10].The fuzzy inferential filters are those from the FIRE (Fuzzy Inference Ruled by Else-action) family [11]. All of them depend on the use of luminance difference between various pixel pairs within the filtering window as input variables v_i . In [11] these differences are calculated between each pixel of the filtering window and its center, i.e. the pixel being processed. In [12] the differences are computed within linear subset of the filtering window with respect to the median. The differences are expressed, as a rule, with linguistic description such as positive Zero and Negative, or by their absolute values labeled as Small, Medium, Big. However, more detailed descriptions such as Positive Small, Positive Medium, Positive Big, Zero, Negative Small, Negative Medium, Negative Big can be used if necessary.

As a representative example of this approach, in [12] the rules that depict the credibility (how appropriate a value is as a filter outcome) are used for the values within linear-shaped sub-windows W_i (horizontal, vertical and diagonal) centered in the currently processed location.

(1) if the absolute difference between the median value z_i and the other points from W_i is very big, then the credibility of z_o is low.

(2) if the absolute difference between the median value z_i and the other points from W_i is very small, then the credibility of z_i is low.

(3) if the absolute difference between the median value z_i and the other points from W_i is medium, then the credibility of z_i is high.

Finally, the median values with the highest credibilities are considered as candidates for the outcome and a further median is performed upon this set. The membership function which measures the credibility has the classical trapezoidal, triangular or parabolic shape.

A simple and direct way of expanding such scalar filters to the multichannel case is to restore the linguistic sintagm luminance difference to inter-vector distance [13]. Thus, any fuzzy, rule-based, scalar filter can be directly translated for multichannel images (including color images).

3.2 Employing the Inherent Color Space Fuzziness

Another effective way of dealing with fuzzy rules in the color case is to study the specific properties of the colors and mainly their characterization in a more fitted space than the primary RGB space. Here it means HIS space. The Hue (H) is a depiction of the color type (for example, the color, is blue, or orange, or red, or green etc.). The Saturation (S) measures the pureness of the color, i.e. the degree of mixing with uniform white. A very low saturation (O, at the limit) denotes that the color is a shade of grey and the RGB components are all the same. The Intensity is a metric of the perceived color luminance and is related to a vertical axis of rotational symmetry of the new color space. The Hue is understanded as an angle that divides the hull of the space in areas corresponding to pure colors. The HIS color space is gained from the RGB color space via the rotation (see Eq. (9)) and the nonlinear transform (see Eq. (10)) [6], [14].

$$\begin{pmatrix} I \\ Y \\ X \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{9}$$

$$\begin{cases} I = K_2(R+G+B) \\ S = \sqrt{X^2+Y^2} \\ H = \arctan(X/Y) \quad , \end{cases} \tag{10}$$

where $X = K_1(G-B), Y = K_4B - K_3(R-G), K_1 = \frac{1}{\sqrt{2}}, K_2 = \frac{1}{\sqrt{3}}$, $K_3 = \frac{1}{\sqrt{6}}$ *and* $K_4 = \frac{2}{\sqrt{6}}$ In a case of HIS transform, Hue varies between $0^\circ \sim 360^\circ$: Blue is 0° , Cyan is 60° , Green is 120°, Yellow is 180° , Red is 240°, Purple is 300°, etc. In order to evaluate exactly Hue, in computing close attention should be paid to positive/negative relation of x and y for deciding the quadrant in which x and y exist. The relation of the cymbols and the quadrants is shown in the following table.

Table 1. The relation of the cymbols and the quadrants

Quadrant	x	y
1	+	+
2	+	-
3	-	-
4	-	+

Some significant properties of the HIS color representation have been noticed and applied. For example, in natural images the Saturation is relatively low and is proportional to the degree of importance of the Hue; the independent noise components acquired on the R, G, B channels are lessened in the Intensity component (a linear combination of the three original channels). So, fuzzy rules are presented for measuring the relevance of each of the three HIS components [14]:

(1) if Saturation is low , then the Hue is irrelevant.
(2) if Saturation is medium ,then the Hue is weakly relevant.
(3) if Saturation is high , then the Hue is very relevant.

This basic set of rules distinguishes a particular of Hue which can be applied to the further processing by a cooperation between Hue and Intensity, relying on the Saturation. This new set of fuzzy rules founds a soft decision for the relative significance of the three color components and their further application to filtering.

(1) if Saturation is low, then the Intensity is used for further processing.

(2) if Saturation is medium, then the Hue and Intensity are jointly used for further processing.

(3) if Saturation is high, then the Hue is used for further processing.

A representative result of a median filtering based on such a fuzzy rule set is illustrated in Fig.1 (f). This method opens up noticeable broad prospects for using the specific inter-color relations in filtering. It is clear that the use of the RGB color space, although attractive for some reasons, has its limitation in terms of measuring the inter-color distances. Perception experiments show that the human eye is unable to properly differentiate certain colors which are just noticeable different and are placed enough close in the color space, under the term Just Noticeable Difference (JND) [15]. The JND equals the Euclidean distance between colors expressed as CIELAB triples and provides that it is smaller than 2.3. It is obvious that the JND finds a natural way for integrating visual uncertainty. It firms the possibility of defining color multisets—all colors that are just noticeable different with respect to some given colors—hence allowing the construction of larger sample populations and thus providing more robust estimates without increasing the size of the filtering window. The JND can perform as a threshold for the perceived error measurement, which can be reflected in some quality measures of the type of Normalized Color Difference (NCD), and thus gain some perceptual support for the objective quality measures.

4. CONCLUSIONS AND COMMENTS

In this paper the crude-fuzzy and the fuzzy inferential filters are proposed. Both these filtering approaches are based on the fuzzy theory. The filters labeled as crude-fuzzy measure, in some way, the membership of each color vector into the set "correct filter output" by either an analytical expression or an iteractive calculation. The performance of the fuzzy inferential filters determined by classical objective quality measures, provides some obvious improvements in comparison with the crude-fuzzy filters. The existing literature shows that the applications of fuzzy theory to color imaging are rare and they are concentrated chiefly in the area of image segmentation due to the direct use of the fuzzy clustering algorithms. Further applications of the fuzzy theory in more fields, in particular, applications of JND in color image filtering remain to us to research.

REFERENCES

[1] Abbas J. and Domanski M., Vector Nonlinear Recursive Filters for Color Images, in "Proceedings of IWSSIP'99—6th International Workshop on Systems, Signals and Image Processing" (Bratislava, Slovakia), pp.30—33, 1999.

[2] Pitas I. and Venetsanopoulos A.N., "Nonlinear Digital Filters—Principles and Applications" Kluwer Academic Publ., Norwell MA, 1990.

[3] Sangwine J.S. and Thornton A.L., Frequency Domain Methods, in: "The Colour Image Processing Handbook (Sangwine J.S. and Horne R.E.N., eds.)", Chapman & Hall, pp.228-241, 1998.

[4] Barnett V., The ordering of Multivariate Data, Journal of Royal Stat. Soc. A, Vol.139, No.3, pp.318—354,1976.

[5] Plataniotis K. N. and Venetsanopoulos A.N., Vector Filtering, in: "The Colour Image Processing Handbook (Sangwine J. S. and Horne R. E. N., eds)", (Chapman & Hall, pp.188—209, 1998.

[6] Castleman K. R., "Digital Image Processing (2nd edition)", Prentice Hall, Englewood Cliffs NJ, 1996.

[7] Economou G. and Fotopoulos S., "A Family of Adaptive Nonlinear Lox Complexity Filters," in: *Proceedings of ECCTD'93—European Conference on Circuit Theory and Design (Davos Switzerland),* pp.521—524, 1993.

[8] Buchowicz A.and Pitas I., "Multichannel Distance Filters," in: *Proceedings of ICIP'94—IEEE Conference on Image Processing,* (Austin, TX), pp.575—578, 1994.

[9] Fotopoulos S. and Economou G., "Multichannel Filters Using Composite Distance Metrics," in: *Proceedings of the IEEE Workshop on Nonlinear Signal and Image Processing* (Neos Marmara, Halkidiki, Greece), Vol.2, pp.503—506, 1995.

[10] Russo F., "Nonlinear Fuzzy Filters: An Overview," in *Proceedings of EUSIPCO'96—8th European Signal Processing Conference* (Trieste, Italy), Vol.1, pp.257—260, 1996.

[11] Russo F., and Ramponi G., "A Fuzzy Operator for the Enhancement of Blurred and Noisy Images," *IEEE Trans.* on Image Processing, Vol.4, No.8, pp.1169—1174, 1995.

[12] Yang X. and Toh P.S, "Adaptive Fuzzy Multilevel Median Filter," IEEE Trans. on Image Processing, Vol.4, No.8, pp.680—682,1995.

[13] Vertan C., Vertan C.I. and Buzuloiu V., "Fuzzy Developments of Multichannel Filters," in: P*roceedings of KES'97 — First International Conference on Conventional and Knowledge-based Intelligent Electronic Systems* (Adelaide, Australia), 1997.

[14] Carron T. and Lambert P., "Symbolic Fusion of Hue-Chroma-Intensity Features for Region Segmentation," in: *Proceedings of ICIP'96—IEEE Conference on Image Processing* (Lausanne, Switzerland), pp.971—974, 1996.

[15] Sharma G. and Trusell H.G., "Digital Color Imaging," *IEEE Trans.* on Image Processing, Vol.6, No.7, pp.901—932,1997.

Ruihua Lu is a full Professor, head of Microcomputer Teaching and Research Section in School of Electronics and Information Engineering, Southwest University. She graduated from Southwest Normal University in 1982. She was a visiting scholar of Tsinghua University (1997-1998).She is a senior member of CCF. She has published two books, over 40 Journal papers. Her research interests are in signal and information processing.

Li Deng is a graduate-student in school of Electronics and Information Engineering, Southwest University.

Kalman Filter and MeanShift Based Occlusion Object Tracking

Shunyan Wang, Shuangzhong Qiu, Luo Zhong
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: wang_syan@126.com

ABSTRACT

Complete occlusion is a difficult problem in object tracking, and it is also one of the most difficult problems in video image processing. This paper describes a method to solve the complete occlusion problem that using MeanShift algorithm, integrating with Kalman filter, employing simply Case-based Reasoning. We also present a new method to update the template while occlusion completely with case base. The experiment results show that, via the method, the object tracking problems while complete occlusions significantly improved.

Keywords: Object Tracking, MeanShift, Kalman Filter, Case-Based Reasoning

1. INTRODUCTION

Research and application of object tracking methods is an important branch in the field of computer vision. Tracking accurately is the base of object tracking, whiling considering the real-time of algorithms. MeanShift, as an efficient pattern-matching algorithm, was introduced into computer vision in 1995 by Cheng[1], and it was firstly applied to the problem of object tracking by Comaniciu and Meer[2]. For MeanShift is simple and efficient, it has been successfully applied to real-time tracking field[2]. MeanShift algorithm can track object that partial occlusions, But it will fail when occluded completely.

This paper describes a method that using MeanShift algorithm, integrating with Kalman filter [3], employing simply Case-based Reasoning method to solve the complete occlusion problem. The experiment results show that, via the method, the accuracy probability of tracking with completely occluded is significantly improved, and can track object robustly.

2. TRACKING ALGORITHM BASED ON MEANSHIFT [4]

MeanShift tracking algorithm initializes the target interactively. At the first frame, choose a feature area by hand, and the area is regarded as the target area. Calculation every eigenvalue of the feature space for all pixel in the initial area, as the target model. Next, calculate every eigenvalue in the feature space for every candidate regions in the following frames, as the candidate model. Bhattacharyya coefficient is generally chosen as the similarity function, through MeanShift iteration, we can get the optimal candidate region as the target.

2.1 Target Representation

For RGB images, we divide each R, G, B subspace into *k* equal intervals, each interval is a bin, and these compose the feature space. The number of features (the number of *bins*) is $m=k^3$. Let x_0 be the spatial center of target model, x_i is the pixel *i* locations in the target region. We define the probability Density of Eigenvalue *u* in target model as:

$$q_u = c\sum_{i=1}^{n} k\left(\left\|\frac{x_0 - x_i}{h}\right\|^2\right)\delta[b(x_i)-u] \qquad (1)$$

Where *u=1, 2, 3……m*，*k(x)* is the *kernel* function, *h* is the bandwidth of *kernel* function, $\{x_i\}_{i=1\ldots n}$ is coordinate of the pixel i. $b(x_i)$ is the eigenvalue of pixel located at X_i . $\delta(x)$ is *Kronecher delta* function. $\delta[b(xi)-u]$ is to find out that whether the pixel at X_i is in the bin *u*. The normalization constant c is derived by imposing the condition

$$\sum_{u=1}^{m} qu=1$$

from where

$$c=\frac{1}{\sum_{i=1}^{n} k\left(\|\frac{x_i - x_0}{h}\|^2\right)} \qquad (2)$$

2.2 Target Candidates

From the second frame and the following frames, let the region that might include the target be candidate target. Its center is y_0, $\{x_i\}_{i=1\ldots.n_h}$ is the coordinate of pixel in candidate region.

So the probability of eigenvalue of the target candidate is computed as:

$$p_u(y_0)=c\sum_{i=1}^{n_h} k\left(\left\|\frac{y_0-x_i}{h}\right\|^2\right)\delta[b(x_i)-u] \qquad (3)$$

Where

$$c_h=\frac{1}{\sum_{i=1}^{n_h} k\left(\|\frac{y-x_0}{h}\|^2\right)} \qquad (4)$$

2.3 Similarity Function

The Bhattacharyya coefficient is a divergence-type measure, which has a straightforward geometric interpretation; it is the cosine of the angle between two m-dimensional unit vectors [4]. [4] has described that Bhattacharyya coefficient is superior then other similarity function in MeanShift algorithm.

Here we use the Bhattacharyya coefficient as the similarity function, it is described as:

$$\rho(y) = \rho(p(y), q) = \sum_{u=1}^{m}\sqrt{p_u(y)q_u} \qquad (5)$$

Where $\rho(y)\in[0,1]$ .and we can easily get that the bigger the $\rho(y)$ is, the more similar between two models.

2.4 Target Localization

To get the maximum of $\rho(y)$, the localization procedure starts from the position y_0 of the target in the previous frame and search in its neighborhood, to find out the location y_1 , where let $\rho(y_1)$ be maximum. Using Taylor expansion around the value $p(y_0)$, then Eq.(5) equals to the following:

$$\rho(p(y),q)=\frac{1}{2}\sum_{u=1}^{m}\sqrt{p_u(y_0)q_u}+\frac{C_h}{2}\sum_{i=1}^{n_h}w_i k(\left\|\frac{y-x_i}{h}\right\|^2) \tag{6}$$

Where

$$w_i=\sum_{u=1}^{m}\sqrt{\frac{q_u}{p_u(y_0)}}\delta[b(x_i)-u] \tag{7}$$

when got the maximum of the similarity function, we can also get the center $\hat{y}_1$ of the candidate, where

$$\hat{y}_1=\frac{\sum_{i=1}^{n_h}x_i w_i}{\sum_{i=1}^{n_h}w_i} \tag{8}$$

Through iteration, we can get an optimal center location y_1 of the target regional. The theory and others about MeanShift algorithm can find out in reference [3][4][5].

3. KALMAN FILTER

The base assumption of MeanShift algorithm is that the object in next frame and current frame is close. In the deduce of Eq.(6), we use Taylor expansion on Eq.(5), when there is no significant changes between the target candidate model and the target model, MeanShift can estimate the candidate center accurately. But when the object moving too fast, or getting occluded, Eq.(6) can not estimate Eq.(5) correctly. And in this case, the result of MeanShift tracking is not very well. These problems can be regarded as the target has not been initialized well. Kalman filter is introduced in this section to solve this problem.

3.1 Kalman Tracking Model

The essence of MeanShift tracking algorithm is tracking the center of the target, which trajectory is a series of points. here the kinematics of the target are described in Cartesian coordinates ,on the x-coordinate and y-coordinate, the target moves linearly ,with a white Gaussian acceleration noise, the acceleration a is a random variable, $a(t)\sim N(0,\sigma_w^2)$.

The system model takes the form:

$$X(k)=[x(k) \quad y(k) \quad x'(k) \quad y'(k)]^T \tag{9}$$

Where $x(k)$, $y(k)$ are the target center on the x-coordinate and y-coordinate, $x'(k)$, $y'(k)$ are the velocities on x-coordinate and y-coordinate.

The measurement model takes the form:

$$Y(k)=[x_c(k) \quad y_c(k)]^T \tag{10}$$

Where $x_c(k)$ 、 $y_c(k)$ are the measurement target center on x-coordinate and y-coordinate.

On the x-axis, according to Newton's Theorem, we have the following function:

$$x(k)=x(k-1)+x(k-1)t+\frac{1}{2}W(k)t^2 \tag{11}$$

$$x'(k)=x'(k-1)+W(k)t \tag{12}$$

Where *t* is a time variable, here we take it as the frames.

According the Kalman two models, system model comes as:

$$X(k)=A(k-1)X(k-1)+B(k)W(k) \tag{13}$$

Measurement model:

$$Y(k)=C(k)X(k)+V(k) \tag{14}$$

then the system model and the measurement model for x-axis are become [6]:

$$\begin{bmatrix} x(k) \\ y(k) \\ x'(k) \\ y'(k) \end{bmatrix}=\begin{bmatrix} 1 & 0 & t & 0 \\ 0 & 1 & 0 & t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x(k-1) \\ y(k-1) \\ x'(k-1) \\ y'(k-1) \end{bmatrix}+\begin{bmatrix} t^2/2 \\ t^2/2 \\ t \\ t \end{bmatrix}W(k) \tag{15}$$

$$\begin{bmatrix} x_c(k) \\ y_c(k) \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}\begin{bmatrix} x(k) \\ y(k) \\ x'(k) \\ y'(k) \end{bmatrix}+\begin{bmatrix} 1 \\ 1 \end{bmatrix}V(k) \tag{16}$$

Where initialized with: $t=1$; $X(-1)=(xs,ys,0,0)$, xs,ys are the target center in the first frame.

4. COMPLETE OCCLUSION HANDING

MeanShift can deal with partial occlusion problem robustly, but when completely, the tracking will be lost. This paper presents a solution on this problem.

4.1 Case Base Establishment

When the object gets occluded completely, we use the information in the case base to estimate the location. This paper present a new case base based on [7] [8].

Case base1: location information. Which is stored the index of each frame and the target location in every frame.

Case base2: eigenvalues, it composed the target feature in frames which are tracked accurately.

When the target get occluded completely, using the case base1 to predict the target in the next frame. And then, from the point we just get, using case base2 to get more optimal locations.

For reasons of the initial area usually is not accurate, and lighting, non-grid object transformation, or target rolling, and some other reasons, in each frame, the initial target model may be not optimal model. So in tracking process, we recalculate initial model as Eq.(15), here we set a threshold Tc . Frames whose correlation coefficient with first frame are bigger then Tc will be add into case base2. For current frame, using the information in case base2, and the correlation coefficient with first frame, using Eq.(15),get a current optimal model. Target model update comes as:

$$q'=\frac{\sum \rho qi}{\sum \rho_i} \tag{17}$$

Where q' is the updated model. ρ is the correlation coefficient with first frame. q_i is the ith correlation coefficient with first frame.

4.2 Search Strategy

In the current frame, it is noted that the residual of $X(k)$ and the $Y(k)$ come as:

$$r(k)=\sqrt{(x_C(k)-\hat{x}(k))^2+(y_C(k)-\hat{y}(k))^2} \tag{18}$$

If $r(k)$ is far greater then normal value, it means that target may get occluded. With (5), here we set another two threshold $Tr,T\rho$,when $\rho(y)>T\rho$ and $r(k)>Tr$, let Kalman predict stopped, and start our search strategy.

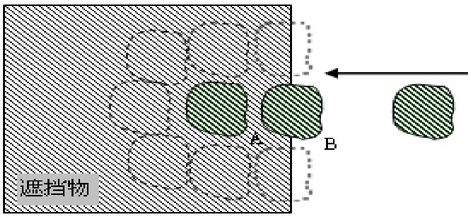


Fig.1. Search Strategy

As the Fig.1 present, assume that at time *t*, the target is at B, and begin get occluded, at *t+1*, target is arrived A , then we can get a conclusion by the threshold that the target get occluded completely, then start the search strategy.

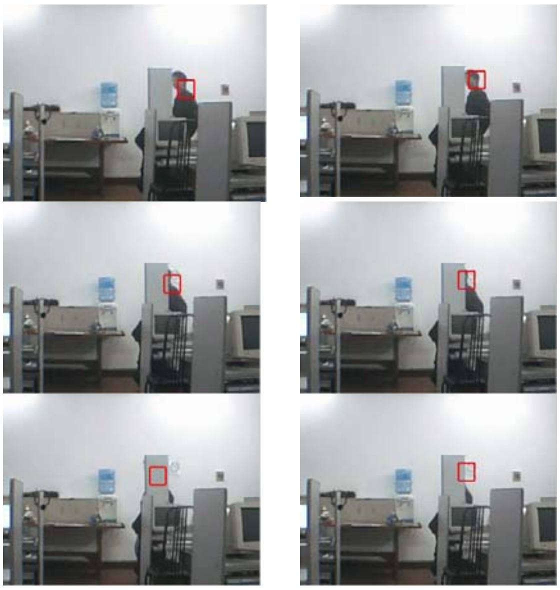
Employing the location information from case base1, using Kalman filter to predict and get the location A at time *t+1*. Then using A as a new center, the kernel bandwidth as radius, search in the 8 direction of A. Following with the information of case base2, employed Eq.(5) , get 8 similarity coefficients, and the biggest is the optimal, and also the optimal location. Now, this process on this frame is over, and get into next frame.

5. EXPERIMENTAL RESULTS

The algorithm runs on a P4 2.66GHZ, 512M memory, Windows XP SP2 operating system, Matlab7.0 environment. Experimental data is of 67 video sequences. From the ninth frame, the target was occluded; the 10th frame occluded completely, 17th frames, partially out of the occlusion, from 18th, completely got out. As shown in Fig.2, when using Kalman filter, the tracking efficiency improved. Using the original MeanShift algorithm, target would be lost when get complete occluded.

6. CONCLUSIONS

In this paper, we improve the accuracy of tracking of occlusion, using the predict function of Kalman filter combined with MeanShift algorithm. Through importing the method of case based, the algorithm can track object of complete occlusion with confidence. The algorithm of object tracking, this paper presented is efficiency, and more robustness.



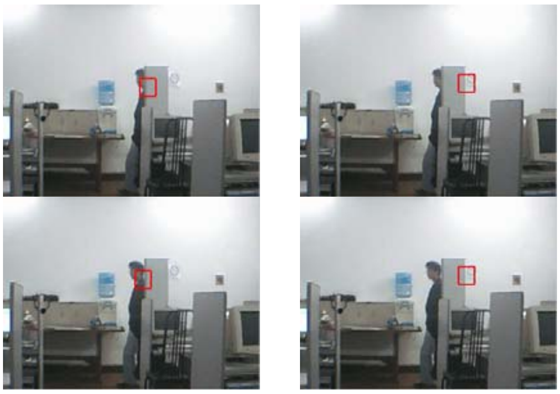


Fig.2. Compare of the two.Left one is employed Kalman filter, right one is without. From up to down are the 9th, 11th, 14th, 16th, 17th frame's tracking results.

REFERENCES

[1] Y. CHENG, “Mean shift, mode seeking and clustering [J],” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1995, 17(8): 790-799.

[2] Comaniciu Dorin, Ramesh Visvanathan, Meer Peter, “Real-time tracking of non-rigid objects using mean shift [A],” in:*Proceedings of the IEEE Conforence on Computer Vision and Pattern Recognition[C]*,Hilton Head Island, Sounth Carolina, USA,2000,2:142~149.

[3] SALMOND D, Target tracking, introduction and Kalman tracking filters[M], QinetiQ Itd, 2001.

[4] Comaniciu D, Ramesh V, Meer P, “Kernel-Based object tracking,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003, 25(5): 564-577.

[5] D.COMANICIU, V.RAMESH, P.MEER, “The variable bandwidth mean shift and data-driven scale selection[A],” in *Proc.8th Intl.Conf.on Computer Vision[C]*, Vancouver, Canada: IEEE, 2001,1:438-445.

[6] .Zhu Sheng-li, Zhu Li-an, “Algorithm for tracking of fast motion objects with Mean shift,” *Opto-Electronic Engineering*, 2006.

[7] Wenhui Liao, Yan Tong, Zhiwei Zhu, Qiang Ji, “Robust Object Tracking with a Case-Base Updating Strategy,” *IJCAI* 2007: 925-930.

[8] Zhiwei Zhu, Wenhui Liao, Qiang Ji, “Robust Visual Tracking Using Case-Based Reasoning with Confidence,” CVPR *(1)* 2006: 806-816.

An Improved Adaptively Weighted Sub-pattern PCA Approach for Face Recognition*

Qiuyu Zhang[1], Yanfeng Jin[1,3], Zhanting Yuan[1], Jiawen Hu[1], Lei Sun[1], Wenjing Li[2]
[1]School of Computer and Communication,Lanzhou University of Technology,Lanzhou,730050,China
[2]College of Mathematics and Information science,Northwest Normal University,Lanzhou,730050,China
[3]Direct Mail Research&Consulting Center,China Post Group,Shijiazhuang,050021,China
Email: jinyanf@126.com

ABSTRACT

A face recognition algorithm based on adaptively weighted Sub-pattern PCA approach is presented in this paper. The proposed algorithm when compared with conventional PCA algorithm has an improved recognition rate for face images with large variations in lighting direction and facial expression. Unlike PCA based on a whole image pattern, the improved adaptively weighted Sub-pattern PCA (IAw-SpPCA) operates directly on its sub-patterns partitioned from an original whole pattern and separately extracts features from both same and different persons'face image, unlike mPCA that neglect different contributions made by different parts of the human face in face recognition, IAw-SpPCA can adaptively compute the contributions of each part and then endows them to a classification task in order to enhance the robustness to both expression and illumination variations. In the process of classification, we use face image set of both same and different person. Experiments on two standard face databases show that the proposed method is effective.

Keywords: PCA, Face Recognition, Modular PCA, Iaw-Sppca Eigenfaces

1. INTRODUCTION

Face recognition is a difficult problem because of the generally similar shape of faces combined with the numerous variations between images of the same face. The image of a face changes with facial expression, age, viewpoint, illumination conditions, noise etc. The task of a face recognition system is to recognize a face in a manner that is as independent as possible of these image variations. Automatic recognition of faces is considered as one of the fundamental problems in computer vision and pattern analysis, and many scientists from different areas have addressed it. As a classical self-organized learning method, principle component analysis (PCA) is widely used in data compression and feature extraction. There are two basic approaches to the computation of principal components: batch and adaptive methods. The batch methods include the method of eigen decomposition and the method of singular value decomposition (SVD), while the adaptive methods are exemplified by Hebbian-based neural networks, such as generalized Hebbian algorithm (GHA) and adaptive principal components extraction (APEX) etc. [1,4]. Despite these different implementations of PCA, their essences are the same, namely, to explain the variance–covariance structure of the data through a few linear combinations of the original variables. So in this paper, we have adopted the batch methods for PCA implementation. Currently, PCA has also become one of the most popular appearance-based algorithms applied to face recognition [6,9]. However, due to utilizing only the global information of face images, this method is not very effective under different facial expression, illumination condition and pose, etc. The recently proposed mPCA method [3] is one of the methods which try to overcome such ineffectiveness by exploring the face's local structure. In this method, a face image is first partitioned into several smaller sub-images, and then a single conventional PCA is applied to each of them. Consequently, variations in expression or illumination in the image will only affect some sub-images in mPCA rather than the whole image in PCA, and thus the local information of a face image may be better represented. However, such a local representation in mPCA ignores the mutual spatial relationship among sub-images partitioned from the original face image, so some spatial information in the original face image is more likely lost and the different contributions made by different parts of face are de-emphasized. In the previous work [5], the usefulness of sub-pattern PCA (SpPCA) has been demonstrated. In the first step of this method, an original whole pattern denoted by a vector is partitioned into a set of equally sized sub-patterns in a non-overlapping way and then all those sub-patterns sharing the same original feature components are respectively collected from the training set to compose a corresponding sub-pattern's training set. In the second step, PCA is performed on each of such sub-pattern's training set to extract its features. At last, a single global feature for the original whole pattern is obtained by concatenating each sub-pattern's PCA projected features together. Although such a concatenation can formally generate a global feature for the original whole pattern and avoids the problems with mPCA, the SpPCA algorithm only utilizes the separately extracted local information from each sub-pattern set, but it does not concern different contributions made by different sub-patterns, in other words, it endows equal importance to different parts of a pattern in classification. As a result, the global vector more likely contains redundant or even useless local information, which will degrade final classification performance as shown in the face recognition experiments in Section 3. IAw-SpPCA proposed here aims to compensate for the shortcomings of the above mentioned algorithms and focuses on the application of face recognition. Not only is the spatially related information in a face image considered and preserved in each sub-pattern, but also the different contributions made by different parts of the face are emphasized. Moreover, these different contributions make classification accuracy improved. The PCA approach is reviewed in section 2. The algorithm is detailed in Section 3. Experiments are carried out in Section 4 to evaluate IAw-SpPCA, SpPCA, PCA and mPCA methods using three standard face databases. Finally, Section 5 concludes this paper.

2. REVIEW THE BASIC THRORY ABOUT PCA

The PCA method has been extensively applied for the task of face recognition. Approximate reconstruction of faces in the

* The Natural Science Foundation of Gansu (3ZS062-B25-037)

ensemble was performed using a weighted combination of eigenvectors (eigenpictures), obtained from that ensemble (Sirovich and Kirby, 1987).The weights that characterize the expansion of the given image in terms of eigenpictures are seen as global facial features. In an extension of that work, Kirby and Sirovich (1990) included the inherent symmetry of faces in the eigenpictures. All the face images in the face database are represented as very long vectors, instead of the usual matrix representation. This makes up the entire image space where each image is a point. Since the faces have a similar structure, the vectors representing them will be correlated. We will see that faces of the same class will group at a certain location in the image space. Hence the face images are represented by a set of eigenvectors developed from a covariance matrix formed by the training of face images[11].The idea behind eigenimages is to find a lower dimensional space in which shorter vectors will describe face images.

2.1 Computing Eigenfaces

Consider the face images in the face database to be of size L by L. These images can be represented as a vector of dimension L^2 , or a point in L^2 -dimensional space. A set of images therefore corresponds to a set of points in this high dimensional space. Since facial images are similar in structure, these points will not be randomly distributed, and therefore can be described by a lower dimensional sub-pattern PCA gives the basis vectors for this subspace (which is called the ‘‘face space’’).Each basis vector is of length L^2 , and is the eigenvector of the covariance matrix corresponding to the original face images.

Let $I_1, I_2, ..., I_M$ be the training set of face images. The average face is defined by:

$$A = \frac{1}{M}\sum_{i=1}^{M} I_i \tag{1}$$

Each face differs from the average face by the vector $Y_i = I_i - A$. The covariance matrix *C* is obtained:

$$C = \frac{1}{M}\sum_{i=1}^{M} Y_i \cdot Y_i^T \tag{2}$$

The eigenvectors of the covariance matrix are computed and the M' significant eigenvectors are chosen as those with the largest corresponding eigenvalues. From these eigenvectors, the weights for each image in the training set are computed as：

$$W_{iK} = E_K^T \cdot (I_i - A) \qquad \forall i, K \tag{3}$$

Where E_K's are the eigenvectors corresponding to the M' largest eigenvalues of *C* and *K* varies from 1 to M' .

2.2 Classification

A test image I_{test} is projected into face space by the following operation:

$$W_{testK} = E_K^T \cdot (I_{test} - A) \quad \forall K \tag{4}$$

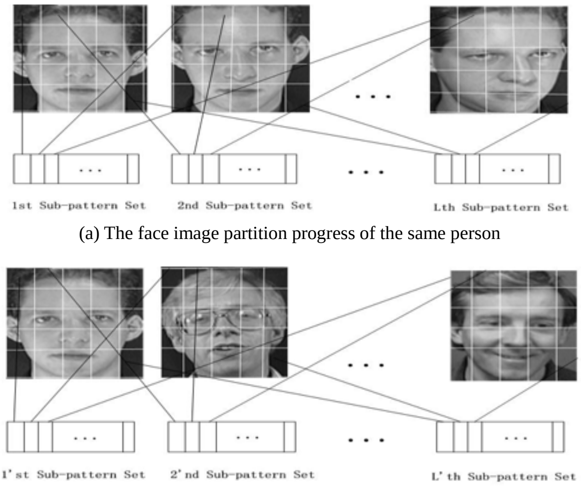
The weights W_{iK} form a vector $T_P^T = [w_1, w_2, ..., w_{M'}]$ which describes the contribution of each eigenface in representing the input face image. This vector can then be used to fit the test image to a predefined face class. A simple technique is to compute distance of W_{testK} from T_P , where T_P is the mean weight vector of the *p*th class. The test image can be classified to be in class *p* when min(D_P) < θ_i , where $D_P = || W_{test} - T_P ||$ and θ_i is the threshold.

3. PROPOSED ALGORITHM

There are three main steps in IAw-SpPCA algorithm: (1) Partition face images into sub-patterns. (2) Compute contributions of each sub-pattern. (3) Classify an unknown image.

3.1 Image Partition

In the IAw-SpPCA algorithm, a face image can be partitioned to a set of equally or unequally sized sub-images, depending on user options, while all sub-images partitioned in the mPCA are strictly confined to equal size due to the mPCA’s inherent limitation. In this paper without loss of generality, we still adopt equally sized partition for a face image. Suppose that there are *N* $W_1 \times W_2$ images belonging to *M* persons in the training set, these persons possess $N_1, N_2, N_3, ..., N_M$ face images, respectively. Each image is first divided into *L* equally sized sub-images in a non-overlapping way which are further concatenated into corresponding column vectors with dimensionality of $W_1 \times W_2 / L$, then we collect these vectors at the same position of all face images to form a specific sub-pattern’s training set, in this way, *L* separate sub-pattern sets are formed. This process is illustrated in Fig. 1.



(a) The face image partition progress of the same person

(b) The face image partition progress of different persons

Fig.1. The progress of face images sub-pattern set construction

3.2 Computing Contributions

We first generate a gallery set and a probe set for each sub-pattern and thus possess corresponding *L* sub-patterns’ gallery and probe sets, respectively. The gallery set is identical to the sub-pattern’s training set, but the probe set is generated by both the ‘‘sub-pattern median face’’ and the ‘‘sub-pattern mean face’’ of each person in this sub-pattern’s training set rather than by one validation set independent from the gallery set as usual. The reason why we select the sub-pattern median and mean faces from the training set is to use these sub-pattern representatives to determine contributions made by different parts to face classification. A process of computing the contributions consists of the two following steps[12]. In the first step, we compute sub-pattern median and mean faces and define a similarity between two samples; the second step is to compute the contributions.

Step 1: For the *j*th sub-pattern, so-called sub-pattern median face of the *i*th person is first computed by:

$$I_{ij\_median} = Median(I_{ij1}, I_{ij2}, ..., I_{ijNi}) \tag{5}$$

and similarly the sub-pattern mean face by:

$$I_{ij\_mean} = \frac{1}{N_i}\sum_{k=1}^{N_i} I_{ijk} \qquad (6)$$

Where I_{ijk} denotes the column vector corresponding to the vectorized *i*th person's *j*th sub-image in the *k*th image of this person. And then the conventional PCA is applied to the *j*th sub-pattern's gallery set, and the respective projection matrix U_j is constructed by selecting first M' eigenvectors associated with the first largest M' eigenvalues. The similarity between sub-pattern samples *x* and *y* is defined as:

$$Similarity(x,y) = -(x-y)^T U_j U_j^T (X-Y) \qquad (7)$$

Step 2: Compute the contribution of a sub-pattern to classification as follows: For a sub-pattern sample from the probe set, the similarities between it and every sample in this sub-pattern's gallery set are first computed, then the gallery samples are ranked in the descending order of the obtained similarities, and the identity of the top 1 sample in the rank list is considered as the recognition result. The result is true if the resulted identity and the probe's identity are matched, else false. After the computation is completed for all probe set samples of the *j*th sub-pattern, we denote by C_j, the number of how many probe set samples of the *j*th sub-pattern are correctly classified. Finally, the contributions made by the *j*th sub-pattern to classification is defined as:

$$W_j = C_j / 2M \qquad （8）$$

3.3 Classification

In this process, in order to classify an unknown face image ***p***, the image is also first partitioned into *L* sub-patterns in the same way previously applied to the training images. Then in this image's each sub-pattern, the unknown sub-pattern sample's identity is determined in a similar way described in Section 3.2, Step 2. Since one classification result for the unknown sample is generated independently in each sub-pattern, there will be total *L* results from *L* sub-patterns. To combine *L* classification results from all sub-patterns of this face image ***p***, two distance matrixes are constructed and denoted by $D(p) = (d_{ij})_{N\times L}$ and $D'(p) = (d'_{ij})_{N\times L}$ the size of $N \times L$, where d_{ij} denotes the distance between the corresponding *j*th sub-patterns of the ***p*** and the *i*th same person, d'_{ij} denotes the distance between the corresponding *j*th sub-patterns of the ***p*** and the *i*th different person and $Min(d'_{ij})\cdot\varepsilon + d_{ij}(1-\varepsilon)$ is set to W_j if the computed identity of the unknown sample and the *i*th person's identity are identical. ε denotes a threshold value. Involving in ε can improved the efficiency of the face recognition greatly than without it. Consequently, a total confidence value that the ***p*** finally belongs to the *i*th person is defined as

$$TC_i(p) = \sum_{j=1}^{L} D_{ij} \qquad （9）$$

And the final identity of this ***p*** is determined by:

$$Identity(p) = \arg\max_i (TC_i(P)) \qquad (10)$$

4. EXPERIMENTS

4.1 Face Image Databases

We carry out the experiments on two face databases: Yale face database [10] and the ORL face database [8]. Fourteen images per person are used. In Yale face database, there are 165 images of 15 adults, 11 images per person while ORL database contains 400 images of 40 adults, and Yale databases feature frontal view faces with different facial expression and illumination condition. Besides these variations, images in ORL database also vary in facial details and head pose. In the preprocessing step, faces images in Yale database are rotated to make eyes horizontal and cropped to size 50×50. Some preprocessed images from Yale database are illustrated in Fig 1. In the ORL database, face images are resized to112×92without any other preprocessing.

4.2 Experimental Results

As noted above, Experiments on ORL database are conducted by each time randomly selecting 5 images per person for training, the rest 5 per person for testing. This experiment is independently repeated 40 times, and the averages of these experiments' results are presented in Fig. 3. The sub-image's sizes in IAw-SpPCA and SpPCA are both set to 7×2, while in mPCA to 16×2. Experiments on the Yale database are carried out by leaving out one image per person each time for testing, the rest 10 images per person for training. This experiment is repeated 11 times by leaving out a different image per person each time. Results listed in Fig. 3. are the average of 11 times results. For the Yale face database, the sub-image's sizes in Aw-SpPCA and SpPCA are both set to be 5×10, while in mPCA to 5×5. In Fig. 3, σ is defined as:

$$\sigma = \left(\frac{number-of-selected-eigenvectors}{number-of-all-the-eigenvectors}\right)\times 100\% \qquad (11)$$

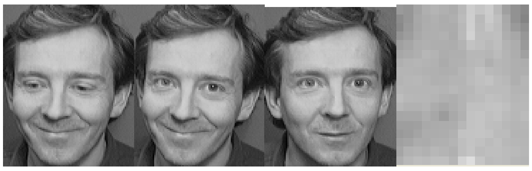
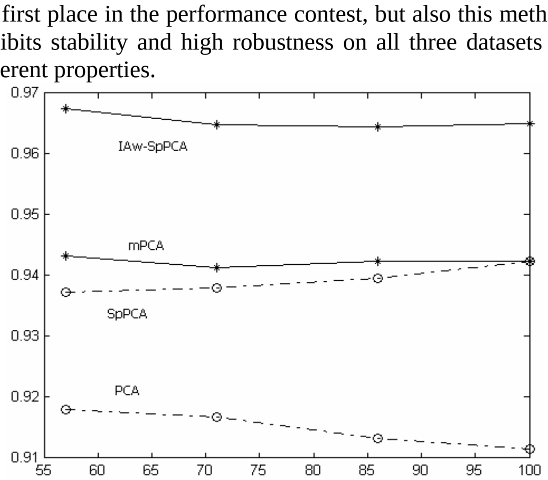
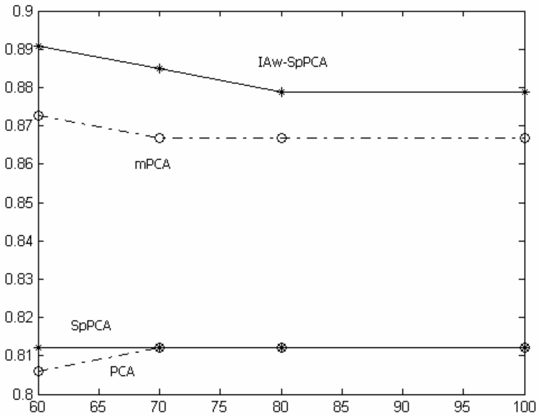


Fig. 2. Images from ORL database [8] and the contribution matrix generated in experiment.

Here all the eigenvectors are sorted in the descending order of their corresponding eigenvalues, and selected eigenvectors are associated with the largest eigenvalues. It can be seen from Fig. 3. that the experiments on ORL and Yale face database show that the performance of proposed IAw-SpPCA is competitive. Although mPCA's performances are impressive on Yale databases, it does even worse than PCA on ORL database. So, we can say that not only does IAw-SpPCA get the first place in the performance contest, but also this method exhibits stability and high robustness on all three datasets of different properties.



(a) Images from ORL database[8]



(b) Images from Yale database[10]

Fig.3. Classification of accuracy comparison of Aw-SpPCA, mPCA, SpPCA and PCA

5. CONCLUSIONS

We propose IAw-SpPCA in this paper and compare it with PCA, mPCA and SpPCA in face recognition. The experimental results indicate that our proposed approach not only is effective but also outperforms them under different facial expression and illumination condition. It is worth to note that although we only adopted the batch methods for PCA computation in our experiments; however, in practice, we can use the neural networks to more effectively implement PCA in order to overcome the batch methods' shortcomings on both storage and computation and thus make IAw-SpPCA much more effective for face recognition.

REFERENCES

[1] S.Becker,M.Plumbley,"Unsupervised neural network learning procedures for feature extraction and classification,"J.Appl. Intell.6(3),(1996),pp.185–205.

[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.* 19 (7),1997,pp711–720.

[3] R.Gottumukkal,V.K.Asari,"An improved face recognition technique based on modular PCA approach,"*Pattern Recogn. Lett.* 25(4),2004,pp.429–436.

[4] S.Haykin,"Neural Networks: A Comprehensive Foundation,second ed.,Prentice-Hall,Englewood Cliffs," NJ,1998.

[5] S.C.Chen,Y.L. Zhu, "Subpattern-based principle component analysis," *Pattern Recogn.* 37 (1) (2004) 1081–1083.

[6] A.M. Martinez,R. Benavente,"The AR face database," *CVC Technical Report* #24,June 1998.

[7] M. Kirby,L. Sirovich,"Application of the KL procedure for the characterization of human faces,"*IEEE Trans. Pattern Anal. Machine Intell,* 12(1),1990,pp.103–108.

[8] Ferdinando Samaria,Andy Harter,"Parameterisation of a Stochastic Model for Human Face Identification,"in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL,* December 1994.

[9] Martinez,A.M, 2000,"Recognition of partially occluded and/or imprecisely localized faces using a probabilistic approach,"in *Proc.of Computer Vision and Pattern Recognition,*vol.1,pp.712–717.

[10] Georghiades,A.S. and Belhumeur,P.N. and Kriegman, D.J. From Few to Many: "Illumination Cone Models for Face Recognition under Variable Lighting and Pose." *IEEE Trans. Pattern Anal. Mach. Intelligence* 23(6),pp.643-660,2001.

[11] Rajkiran Gottumukkal, Vijayan K.Asari. "An improved face recognition technique based on modular PCA approach,"*Pattern Recognition Letters* 25, 2004, pp. 429–436.

[12] Keren Tan,Songcan Chen,"Adaptively weighted sub-pattern PCA for face recognition,"*Neuro computing* 64,2005,pp.505–511.

Qiuyu Zhang: Associate professor and master tutor. Vice dean of School of computer and communication in Lanzhou University of Technology, director of software engineering center, vice dean of Gansu manufacturing information engineering research center, director of "software engineering" characteristic research direction and academic group of Lanzhou University of Technology. His research interests include: image processing and pattern recognition, multimedia information processing, information security, software engineering etc.

Yanfeng Jin: Graduate student. Born in Shijiazhuang Hebei province in 1981, he has published many academic papers in domestic core magazine and international conference His research interests include: image processing and pattern recognition, computer vision, information security.

Comparative Analysis for Probability Modeling of Multi-class SVM *

Xiang Zhang[1,2] ,.Xiaoling Xiao[1,3]
[1]Yangtze University, Jingzhou, Hubei, 434023, China
[2] Key Laboratory of Exploration Technologies for Oil and Gas Resources, Ministry of Education, Jingzhou, 434023
[3]School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430063
Email: zx_jr_xl@163.com

ABSTRACT

The one-against-one method and the one-against-rest method are two popular multi-class classification methods that combine together all results of two-class support vector machine classifiers. The paper presents the probability output of two methods in multi-class SVMs. The binary output and the probability output of two multi-class SVM methods in terms of classification precision and the total times of both training and predicting stages are compared and analyzed in order to evaluate the classification performance of probability output of multi-class SVM. Three experiment results show that the probability output of the one-against-one method exhibits the excellent classification performance in terms of classification precision and computational cost.

Keywords: Support Vector Machine, Probability Modeling, Multi-class Classification, Comparative Analysis

1. INTRODUCTION

Support vector machines (SVMs) have been proved to be a fruitful learning machine, especially for classification [1]. Standard SVMs do not provide probability output. A classifier produces a posterior probability to enable post-processing in practical recognition. Posterior probabilities are also required when a classifier is making a small part of an overall decision, and the classification outputs must be combined for the overall decision, especially for the output of multi-class support vector machines.

Some works are proposed for modeling the probability outputs of SVMs. For the probability modeling of two-class SVM, Hastie and Tishirani used Gaussians to fit the class-conditional densities $p(f|y=1)$ and $p(f|y=-1)$, and compute the posterior probability by the Bayes’ s rule[2]. Plat used a parameter model to fit the posterior $p(y=1|f)$ directly instead of estimating the class-conditional densities $p(f|y=1)$ [3]. Zhang estimate the probabilistic outputs for support vector machines based on the maximum entropy estimation [4]. These received results were promising, but they did not extended their methods to multi-class problem. Since SVMs were originally designed for binary classification, it is not a straightforward issue to extend binary SVM to multi-class problem. There are different strategies to decompose a multi-class problem into a number of binary classification problems. For a N-class classification problem, one method is to use one-against-rest principle to construct N binary classifiers. Each binary classifier distinguishes one class from all the other classes. The other is so called one-against-one. This method constructs all possible $N(N-1)/2$ two-class classifiers, each of which is used to discriminate two of the N classes [5] .

In this paper, we propose probability estimates for two sorts of multi-class support vector machines: one-against-rest classifiers and one-against-one classifiers. We make comparative analysis for probability estimates of two kinds of multi-class support vector machines in terms of classification precision and the total times of both training and predicting stages.

2. POSTERIOR PROBABILITY ESTIMATE IN MULTI-CLASS SVM

The solution of binary classification problems using support vector machines is well developed, but multi-class problems with more than two classes have typically been solved by combining independently produced binary classifiers. The classical and effective approaches to solving N class pattern recognition problems are the one-against-rest method and the one-against-one method. The following section will discuss the probability output of the one-against-rest classifiers and the one-against-one classifiers for the multi-class classification problems.

2.1 One-against-rest Classifiers

The earliest used implementation for SVM multi-class classification is probably the one-against-rest method. In this method, it constructs N different binary SVM classifier where N is the number of classes. The ith SVM is trained with all of the examples in the ith class with positive labels, and all other examples with negative labels. In the classification phase, the classifier defines the estimated posterior probability of the current input vector as:

$$P(i\,|\,\mathbf{x})=\frac{P_{iar}(i\,|\,\mathbf{x})}{\sum_{j=1}^{N}P_{jar}(j\,|\,\mathbf{x})}\qquad i=1,2,...,N \tag{1}$$

where, $P_{iar}(i\,|\,\mathbf{x})$ is the probability output of the binary SVM classifier which separates ith class and the rest. $P(i\,|\,\mathbf{x})$ is the final estimated posterior probability of the current input vector $\mathbf{x}$.

2.2 One-against-one Classifiers

Another major method is called the one-against-one method. For N classes classification problem, this method constructs $N(N-1)/2$ binary SVM classifiers where each one is trained on data from two classes. The posterior probability of each binary SVM classifier is estimated by Zhang’s method. All posterior probabilities of $N(N-1)/2$ binary SVM classifiers are combined to estimate the final posterior probability output for the current input vector $\mathbf{x}$. The general combination method is simple summation, as following:

* This work was funded under Project 60673189 supported by the National Natural Science Foundation of China, Project 2004ABA043 supported by the Natural Science Foundation of Hubei Province and Project D200612002 supported by the Key Scientific Research Project of Hubei Education Department.

$$P(i\mid\mathbf{x})=\frac{\sum\limits_{j=1,j\neq i}^{N}P_{iaj}(i\mid j;\mathbf{x})}{\sum\limits_{k=1}^{N}\sum\limits_{j=1,j\neq k}^{N}P_{kaj}(k\mid j;\mathbf{x})}\qquad i=1,2,...,N\qquad(2)$$

Where, $P_{iaj}(i\mid j;\mathbf{x})$ is the probability output of the binary SVM classifier for between ith class and jth class. $P(i\mid\mathbf{x})$ is the final estimated posterior probability of the current input vector $\mathbf{x}$.

$$\sum_{k=1}^{N}\sum_{j=1,j\neq k}^{N}P_{kaj}(k\mid j;\mathbf{x})=\frac{N(N-1)}{2}\qquad(3)$$

Thus

$$P(i\mid\mathbf{x})=\frac{2}{N(N-1)}\sum_{j=1,j\neq i}^{N}P_{iaj}(i\mid j;\mathbf{x})\qquad i=1,2,...,N\qquad(4)$$

From Eq. (4), it is shown that each classifier plays the same weight in combination of posterior probabilities of $N(N-1)/2$ binary SVM classifiers.

3. COMPARATIVE EXPERIMENTS AND ANALYSIS

In order to evaluate the classification performance of probability output of multi-class support vector machine, we make two groups of comparative experiments in terms of the classification precision and the total times of training and predicting stages respectively. For the purpose of comparison, both the binary output and the probability output methods of two multi-class SVM methods, the one-against-rest method and the one-against-one method, are considered for classification.

The implementation is based on LIBSVM [6], a simple, easy-to-use and efficient software developed by Chih-Chung Chang et. al. and available through:

http://www.csie.ntu.edu.tw/~cjlin/libsvm. The LIBSVM is originally designed for SVM classification and regression on Unix system, so we make some modifications for medical image classification and use Microsoft Visual C++6.0 to rebuild the windows version. The modified software runs on a 2.0GHz P4 processor and the operating system is Windows 2000 Professional. The size of the main memory is about 512Megabytes.

3.1 Experiment 1

In this section we present experimental results on several multi-class problems: Segment, Waveform, Usps and Mnist, which have more classes and higher features [7]. The numbers of classes and features are reported in Table 1. Each feature of all four data is scaled to [-1,1]. In each scaled data, we randomly select 300 training and 500 testing instances from thousands of data points. 20 such selections are generated and the testing error rates are averaged.

Table 1. Experimental dataset1

Dataset	Segment	Waveform	Usps	Mnist
# class	7	3	10	10
#feature	19	21	256	784

Table 2 shows the classification error rates of the above four experimental datasets obtained by both the binary output and the probability output methods of two multi-class SVM methods, the one-against-rest method and the one-against-one method. The RBF kernel was used in binary support vector machine, which parameter is listed in Table 2.

Table 2. The classification error rates of dataset1 for different multi-class SVM methods

Dataset	The classification error rates (%)				Kernel parameter
	one-against-one		one-against- rest		
	Binary output	Probability output	Binary output	Probability output	
Segment	7.0	2.5	8.1	2.9	$\sigma^2=1.0$
Waveform	15.0	7.3	16.7	7.0	$\sigma^2=1.0$
Usps	9.7	4.0	11.7	5.0	$\sigma^2=50.$
Mnist	14.4	5.3	16.0	5.3	$\sigma^2=50.$

From the table 2, we can see that two probability output methods of multi-class support vector machines obtain less classification error rates than two binary output methods of multi-class support vector machines. The situation is especially obvious when the number of feature is very large, for examples, the number of feature for Mnist dataset is 784, its classification error rate decreases from 16.0% to 5.3%. At the same time, we can also see that the classification error rates by the binary output of the one-against-one method are lower than those by the binary output of the one-against-rest method, whereas the probability output of the one-against-one method achieves the similar classification precision as the probability output of the one-against-rest method, such as the classification error rate for Mnist dataset is 5.3%.

3.2 Experiment 2

Another comparative experiment is taken using simulated MR images. We obtain MRI brain data from McConnell Brain Imaging Center, Montréal Neurological Institute, McGill University, which is available on WWW at “http://www.bic.mni.mcgill.ca/brainweb/”. The size of image volume is $181\times217\times60$ pixels. Thickness of each slice is 3 mm. From the BrainWeb phantom, we choose different simulated high resolution T1-weighted 2D scans with 3% noise level and 40% intensity non-uniformity, and the result is compared to the model segmentation, as provided by BrainWeb.

At each location in 2D MR images, we apply SVM classifiers to determine which one of the three brain tissues (WM, GM and CSF) and the background the pixel belongs to. For training data sets, each class is given a respective label (0: background, 1:WM, 2:GM, 3:CSF). In our experiments, the training samples are randomly sub-sampled from a 2D 181×217 image. The testing data sets are chosen both from the rest of the same 2D MR image and from the other 2D MR images. For three brain tissues classes, there are a very large number of background pixels in the labeled image, so that the training set for the “background” class can be impractically large. To solve this problem, we randomly select four classes training samples from 2D image at different random proportion to keep the number of training examples of each class approximate. 27 gray-level and texture features are chosen from 2D MR image [8].

Table 3 shows the classification error rates of the simulated MR images obtained by both the binary output and the probability output methods of two multi-class SVM methods, the one-against-rest method and the one-against-one method. The RBF kernel was used in binary support vector machine, which parameter is listed in Table 3.

Table 3. The classification error rates of dataset2 for different multi-class SVM methods

Silce No.	The classification error rates (%)				Kernel parameter
	one-against-one		one-against- rest		
	Binary output	Probability output	Binary output	Probability output	
22$^{\#}$	9.82	9.47	11.74	10.3	$\sigma^2 = 10.0$
32$^{\#}$	13.8	11.2	18.0	15.3	$\sigma^2 = 0.5$

From the table 3, we can see that two probability output methods of multi-class support vector machines obtain less classification error rates than two binary output methods of multi-class support vector machines respectively. For the binary output and the probability output of multi-class support vector machines, the one-against-one method generally achieves better performance than the one-against-rest method.

3.3 Experiment 3

From the previous two experiments, we can see that the one-against-one method generally achieves better performance than the one-against-rest method in term of classification precision. This experiment will compare the performance for the above four methods of multi-class support vector machines in term of computational times during both training and testing stages. The results in term of the average computational time of simulated MR images in experiment 2 are listed in Table 4. The number of training samples and testing samples is 1500 and 39277, respectively.

Table 4. The average computational times of dataset2 for different multi-class SVM methods

Silce No.	Total times including training and testing(s)			
	one-against-one		one-against- rest	
	Binary output	Probability output	Binary output	Probability output
22$^{\#}$	7.38	6.56	9.52	12.13
32$^{\#}$	7.89	10.3	19.03	34.56

From the table 4, we can see that there exists large difference in computational costs for the two probability output methods of the one-against-one method and the one-against-rest method, although the classification precisions of the two methods are similar. The probability output of the one-against-rest method costs twice the total time than the probability output of the one-against-one method, such as the total time for the probability output of the one-against-one method is 10.3s, compared with 34.56s for the probability output of the one-against-rest method. For N classification problem, the one-against-one method need build $N(N-1)/2$ two-class SVM classifiers, which is more than those needed by the one-against-rest method, but the number of samples in each two-class SVM classifier for the one-against-rest method is far larger than that for the one-against-one method. The total time is related with the number of samples in two-class SVM classifier, and it results in heavy computational cost for the one-against-rest method.

4. CONCLUSIONS

Support vector machines have been introduced as a new technique for solving various pattern recognition problems. The probability output of support vector machines is not only qualitative, but also quantitative. The probability output of the one-against-one method exhibits the excellent classification performance in terms of classification precision and computational cost. Future work will concentrate on the study of combining rules for the one-against-one method in multi-class support vector machines.

REFERENCES

[1] Burges,C., “A tutorial on support vector machines for pattern recognition”. *Data Mining and Knowledge Discovery*, 1998,2:955-974.

[2] T.Hastie, R.Tibshirani. “Classification by Pairwise Coupling”.*The Annals of Statistics*, 1998,26(1):451-471

[3] J.C. Platt.” Probabilities for Support Vector Machines”. In: *Advances in Large Margin Classifiers*. Massachusetts:MIT Press,2000,61-74.

[4] ZHANG Xiang,, XIAO Xiao-Ling, Xu Guang-You. “Probabilistic Outputs for Support Vector Machines based on the Maximum Entropy Estimation”. *Control and Decision*, 2006,21(7): 767-770.

[5] Hsu C W, Lin C J. “A comparison of methods for multi-class support vector machines”. *IEEE* Transactions on Neural Networks, 2002,13(2):415-425.

[6] C.C Chang, C,J, Lin, LIBSVM: a library for support vector machines. 2001. Sofyware available.
http:// www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz

[7] Hsu C W. Department of Computer Science. National Taiwan University, Taipei 106, Taiwan. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.2006.

[8] X.Zhang,X.L.Xiao, J.W.Tian,J.Liu, G.Y.Xu. “Application of Support Vector Machines in Classification of Magnetic Resonance Images”. *International Journal of Computers and Applications*,2006, 28(2):122-128.

Xiang Zhang was born in 1969. He is currently an associate professor. He did Postdoctor research in Department of Computer Science and Technology in Tsinghua University in 2005-2006. He received his Ph.D. in Institute of Pattern Recognition and Artificial Intelligence in Huazhong University of Science and Technology in 2004. His research interests include pattern recognition, and image processing. He has published over 50 technical papers in domestic, international conferences and journals, in which 14 papers were indexed by SCI, EI, and ISTP. He has gained some honors including the first-class award in the science and technology progress prize by China Petroleum and Chemical Industry Association in 2005, and the second-class award in the Science and Technology Progress Prize by the government of Hubei province in 2005.

An Image Blocking Restoration Method Based on the Fuzzy Genetic Algorithm*

Li Deng, Ruihua Lu
School of Electronics and Information Engineering, Southwest University
Chongqing 400715, China
Email: chenlr@swu.edu.cn

ABSTRACT

Aiming at the high computational complexity of the genetic algorithm for image restoration, a kind of blocking restoration method based on the fuzzy genetic algorithm is proposed. It divides averagely a large-scale blurred gray image into blocks and utilizes the genetic algorithm to restore them sequentially. Moreover, the histogram statistics of the blocks are sent to the fuzzy logic controller to adjust the parameters of the genetic algorithm, including the parameters of fitness function, the initial temperature of simulated annealing for fitness scaling, the probabilities of crossover and mutation. Finally, the recovered blocks are recombined to a whole image, and the boundary errors caused by block processing are damped in the meantime. Experiment results show that the presented method can effectively reduce the computational complexity of the genetic algorithm for image restoration, and the quality of the recovered image is obviously improved in comparison with the Wiener filter and the conventional genetic algorithm.

Keywords: Fuzzy genetic algorithm, Image blocking restoration, Histogram, Fuzzy logic controller

1. INTRODUCTION

The Genetic Algorithm (GA) is an adaptive probability searching method for global optimization formed by simulating the evolutionary process of organisms. In searching process, GA only needs the information of the object function in spite of its differentiability and is capable of solving the nonlinear problems which may cause difficulty for gradient-based methods [1]. Especially, when the blurred image is with noise or the reason of degradation is unknown, the image restoration usually becomes an ill-posed problem [2,3], which would result in more constraints or even no resolution for conventional linear methods. The image restoration can be viewed as an optimization problem and the complicated *a priori* constraints can be conveniently incorporated into the evolutionary process by modifying the object function appropriately, so the GA method is suitable for image restoration [3]. However, the quality of restored image is not ideal in view of the large calculation complexity by treating the image restoration as a global problem, together with the poor local performance because of the random search of GA [1,4].

For the above-mentioned reasons, a blocking deblurring method based on the fuzzy genetic algorithm is proposed. This paper consists of three sections except the introduction and references. Section 2 describes the fundamental idea of this method and the design of the fuzzy logic controller. Section 3 presents the results of simulation experiments. In section 4 the conclusions are given.

2. THE IMAGE BLOCKING RESTORATION METHOD BASED ON THE FUZZY GENETIC ALGORITHM

2.1 The Fundamental Idea of the Proposed Method

One of the drawbacks of GA is its high computational complexity [3~5]. In recovering an image by GA, an individual of the population is just an image. If the image size is $U \times V$, and the gray level is 0~L-1, we need to find out $U \times V$ lattice points which most satisfy the object function among the $L^{U\times V}$ lattices in the $U \times V$ dimensional space. What is more, the amount of computation would be growing exponentially with the increasing tendency of the image size. To avoid this trouble, a strategy of “divide and rule” is adopted. Since deblurring blocks respectively can lessen the quantity of data during the genetic evolutionary process, every block can obtain preferable effects in a short time. In addition, the gray-level distribution of each block is diverse, some blocks are simple backgrounds with unitary gray levels, some are details with wide-ranging distributions, the others may be the borderlands between the two instances, having strong edge information. Hence the fuzzy logic controller (FLC) is adopted to adaptively tune the crucial parameters of GA for better recovery performance.

2.2 Design of the Fuzzy Logic Controller

Histograms are the basis of image processing and are lent to various image processing applications due to the inherent statistical characteristics plus the merit of convenient operation. Here, the statistical information of a histogram is used as the input variable of FLC. The histogram of a digital image with gray levels in the range [0, L-1] is a discrete function [6],

$$h(r_k) = n_k, \qquad k = 0,1,\cdots,L-1, \tag{1}$$

where r_k is the $k-th$ gray level and n_k is the number of pixels in the image having gray level r_k, thus the input variable of FLC can be expressed as

$$I_i = \sum_{k=0}^{L-1}(h(r_k) \sim= 0)\Big/L,\ k = 0,1,\cdots L-1,\ i = 0,1,\cdots N, \tag{2}$$

In Eq.(2) N is the total number of blocks, I_i denotes the percentage of gray level numbers in the whole image made up by that of block i. The larger the I_i, the affluenter the gray levels of block i.

The output variables of FLC are parameters r, w in the fitness function, the initial annealing temperature T_0, the probability of crossover P_c and the probability of mutation P_m. According to the energy function of natural image in [7],

* Supported by the fund of University Science and Technology under Grant SWNU2004006.

the fitness function is defined as follows:

$$E(\hat{f},l)=\| \mathrm{g}-\hat{f}*h\|^2+r\sum_i\sum_j(\hat{f}_i-\hat{f}_j)^2(1-l_{ij})+wE_{line}(l), \tag{3}$$

where $\hat{f}$ is an estimate of restored image , l_{ij} is a random variable , the value of which is 1 if the pixel i borders upon the pixel j , or else is 0. The first term of Eq.(3) indicates the coherent measure of energy between the blurring model $\hat{f}*h$ and the degraded image g , the second term refers to the neighborhood correlation of image data, and the last one denotes the punishment for *priori* information of edge property, which is set by some artificial rules in [7] on account of the calculation inconvenience of $E_{line}(l)$. Because every block has distinct intensity of correlation and edge information, the parameters r and w located severally in the second and the third term are adaptively controlled by FLC.

On the other side, when using GA to restore an image, the individuals are generally so similar as to cause the small differences between fitness values. In order to improve the predominance of the excellent individuals, the fitness scaling method based on the simulated annealing in [8] is used to stretch the distance of the fitness values of individuals with the descent of temperature. The scaling function is as follows:

$$E_i=\frac{e^{E_i/T}}{\sum_{i=1}^{K}e^{E_i/T}},\quad T=T_0(0.99^{t-1}) , \tag{4}$$

where E_i is the fitness value of individual i , K is the size of population, t is the generation number, T is the annealing temperature, and T_0 is the initial value of T which also accepts the control of FLC.

The membership function curves and fuzzy domains of aforementioned variables are shown in Fig.1. The membership functions of I_i have the forms of triangle and trapezoid, and those of the rest variables are all Gauss membership functions. The fuzzy linguistic values of the variables are “Big”, “Median”, and “Small” signed as “B”, “M”, and “S” for short.

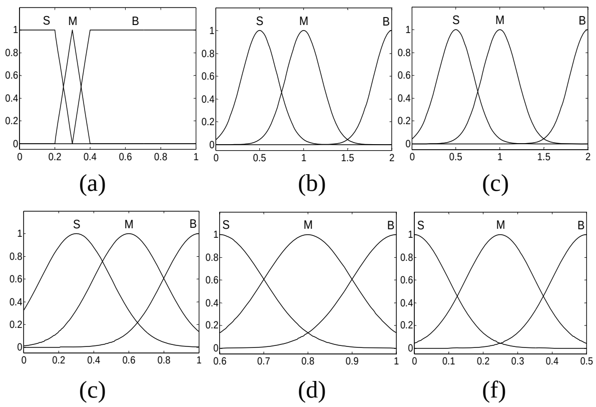
The fuzzy logic rules of FLC are depicted as the following:

Rule1. ***if*** $\{I_i\ is\ S\}$, ***then*** $\{r\ is\ B\},\{w\ is\ B\}$,
$\{T_0\ is\ S\}$, $\{P_c\ is\ B\}$ *and* $\{P_m\ is\ S\}$.

Rule2. ***if*** $\{I_i\ is\ M\}$, ***then*** $\{r\ is\ S\},\{w\ is\ S\}$,
$\{T_0\ is\ M\}$, $\{P_c\ is\ M\}$ *and* $\{P_m\ is\ M\}$.

Rule3. ***if*** $\{I_i\ is\ B\}$, ***then*** $\{r\ is\ M\},\{w\ is\ M\}$,
$\{T_0\ is\ B\}$, $\{P_c\ is\ S\}$ *and* $\{P_m\ is\ B\}$.

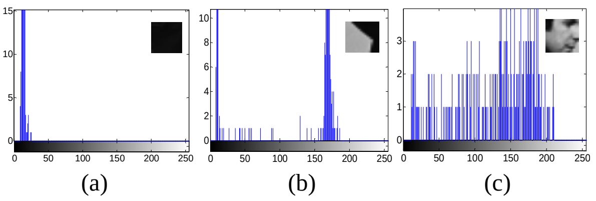
In Rule 1, the fuzzy linguistic value of I_i is “S”, and the histogram of the block i is a narrow district with large amplitude. The block i is the simple background or the region with unitary gray levels (see Fig.2 (a)). The gray values of adjacent pixels are close and have no boundary. This situation may bring on small results of the second and the third term in Eq.(3), so r, w should be set as “B”. Considering the



(a) (b) (c)
(c) (d) (f)

Fig.1. The membership function curves of the input and output variables in FLC: (a) Input variable I_i ; (b) Output variable r ; (c)Output variable w ; (d) Output variable T_0 ; (e) Output variable P_c ; (f) Output variable P_m .

similarity of individuals, T_0 can be set as “S” for stronger effects of fitness scaling, and P_c as “B” to exchange the good parts of individuals to recombine the better ones, yet P_m as “S” to avoid the bad influence on the whole block for the mutation point which is unlike its neighbors. In Rule 3, I_i is “B”. The distribution of the histogram is relatively continuous and concentrative, and the amplitude changes gently also. Here the block i is the details of the whole image (see Fig.2 (c)) which has no obvious edge, and the values of the second and the third term in Eq.(3) is moderate. For this case, r, w should be set as “M”, and T_0 as “B” to slow down the descending speed of T for discovering the excellent individuals more accurately, but P_c as “S” to protect the existing available information of individuals, and P_m as “B” to give the population a little change which would be good for the evolution. Last, in Rule 2, I_i is “M”. The distribution of gray levels is dispersive with visible peaks and valleys. The block i has clear edge and evident distinctness of neighboring pixels, which probably locates at the border of the above instances or between the districts with great difference of gray levels (see Fig.2 (b)). At this time, the second and the third term in Eq. (3) are both big, so r, w should be set as “S”, and the cases of the other parameters are between that of Rule 1 and Rule 3, so all of them can be set as “M”.



(a) (b) (c)

Fig.2. Blocks of different gray-level distributions and their histograms:
(a) The block and its histogram mentioned in Rule1;
(b) The block and its histogram mentioned in Rule2;
(c) The block and its histogram mentioned in Rule3.

3. SIMULATION EXPERIMENTS

An 256×256 image of the cameraman with distinct contrast between the background and the subject is selected as the experimental object , the gray levels of which are in the range [0, 255]. First, the original image is blurred sequentially by the Gaussian noise with variance of 0.02, the horizontal shift of 30 points and the clockwise rotation of 15 degrees, then the degraded image is divided into 256 blocks with the size of 16×16, which are restored by the fuzzy genetic algorithm in turn. The initial population is made up of the deblurred images of the Wiener filter with stochastic disturbance. The two-point crossover operation is adopted, whose cross coefficient is limited in [0, 1] and decreases with the increasing of the generation number[9]. The mutation value of mutation operator is the neighborhood mean of the mutation point. Besides, the population size is reduced with the evolution for the sake of improving the algorithm speed. Once all the blocks are completely restored, returning to a whole image, the noises arranging in row of the image are substituted by the average values of four neighbors in the same array. In the similar fashion, the noises in array are replaced by the average values of four neighbors in the same row. Finally, the degraded image can return to the original by the Wiener filter and the conventional genetic algorithms with different initial populations. The three genetic algorithms have the same fitness function defined as Eq.(3), the population size of 300 and the maximum generation number of 200. The probabilities of crossover and mutation are fixed at 0.8 and 0.1, and the cross coefficient at 0.5 in the two conventional genetic algorithms. Moreover, the initial populations named as Group 1 and Group 2 are formed by stochastic generation and the way mentioned above. The experimental results are shown in Figure 3 and Table 1. Table 1 involves three kinds of objective evolution criterions including the corresponding Peak of the Signal-to-Noise Ratio (PSNR), the elapsed time (Time) and the maximum memory usage (Memory).

From Fig.3 and Table 1 it can be concluded that with the same conditions of fitness function, population size and maximum generation number, the conventional genetic algorithm with Group 1 has the lowest PSNR and the biggest memory usage, which would cost more time to achieve the similar result gotten by the other methods. The conventional genetic algorithm with Group 2 has the smaller memory usage than that of the former, but the PSNR has little improvement than that of the Wiener filter. However, the memory usage and the elapsed time of the proposed method have obvious reduction compared with the two conventional genetic algorithms, and the restoration quality by the proposed method is much better than that by the other methods both on PSNR and visual effect.

Table 1 The objective evolution criterions of the results with different methods

	PSNR (dB)	Time (s)	Memory (MB)
The Wiener filter	20.2358	1.7313	18
The conventional GA with Group1	11.5779	10228.4172	609
The conventional GA with Group2	20.3345	10525.2674	317
The proposed method	22.0197	7644.8677	16

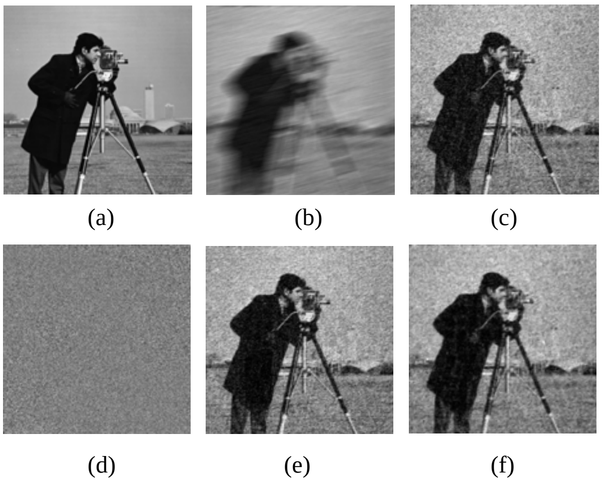


Fig.3. The results of different methods:
(a) The original image;
(b) The degraded image;
(c) The result of the Wiener filter;
(d) The result of the conventional GA with Group1;
(e) The result of the conventional GA with Group2;
(f) The result of the proposed method.

4. CONCLUSIONS

The above experiments indicate that the blocking restoration method based on the fuzzy genetic algorithm has larger predominance on the operation speed, the memory cost and the restoration quality than the conventional genetic algorithm. However, some respects remain to be improved. For example, the ways of blocking and edge noise removing are too simple and rough. The relations between the block size, the memory usage and the operating speed should be weighed more carefully. By the way, if the block processing method is able to be combined with the parallel algorithm, its efficiency would have further advancement [5, 10].

REFERENCES

[1] Tianzi Jiang, Frithjof Kruggel, “3D MR Image Restoration by Combining Local Genetic Algorithm with Adaptive Pre-conditioning,” in *Proceedings of IEEE Conference on Pattern Recognition*, Vol. 3, Sept. 2000, pp.298~301.

[2] Zhang Buyun, Xu Dinghua et al, “Stabilized Algorithms for Ill-posed Problems in Signal Processing,” in *Proceedings of IEEE Conference on Info-tech and Info-net*, Vol.1, Nov. 2001, pp.375~380.

[3] Yen-Wei Chen, Zensho Nakao et al., “Restoration of Gray Images Based on a Genetic Algorithm with Laplacian Constraint,” *Fuzzy Sets and Systems*, Vol.103, 1999, pp.285~293.

[4] Taturo Enokura, Yen-Wei Chen et al, “A Fast Image Algorithm for Image Restoration Based on a Hybrid GA and SA,” in *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*, Vol.4, Oct. 1999, pp.891~ 894.

[5] Yen-Wei Chen, Zensho Nakao et al, “Parallelization of a Genetic Algorithm for Image Restoration and Its Performance Analysis,” in *Proceedings of IEEE Conference on Evolutionary Computation*, May 1996, pp.463 ~ 468.

[6] Rafael C. Gonalez, Richard E.Woods, *Digital Image Processing*, 2nd ed., BeiJing: Publishing House of Electronics Industry, 2002.

[7] Geman Stuart, Geman Donald, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 1984, pp.721~741.

[8] Stoffa Paul, Sen Mrinal K., "Nonlinear Multiparameter Optimization Using Genetic Algorithms," *Inversion of Plane-wave Seismograms Geophysics*, Vol.56, No.11, 1991, pp.1794~1810.

[9] Z.Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, 2nd ed., Berlin: Springer-Verlag, 1994.

[10] Wongyong Sung, Sanjit K. Mitra et al, "Multiprocessor Implementation of Digital Filtering Algorithms Using a Parallel Block Processing Method," *IEEE Trans. on Parallel and Distributed Systems*, Vol.3, No.1, Jan. 1992, pp.110~120.

Li Deng is a graduate-student in school of Electronics and Information Engineering, Southwest University.

Ruihua Lu is a full Professor, head of Microcomputer Teaching and Research Section in School of Electronics and Information Engineering, Southwest University. She graduated from Southwest Normal University in 1982, was a visiting scholar of Tsinghua University (1997-1998). She is a senior member of CCF and has published two books, over 40 Journal papers. Her research interests are in signal and information processing.

Driver's Face Image Recognition for Somber Surroundings* Based on Computer Vision

Ying Yang[1], Wei Zhou[1], Guangyao Zhao[1]
[1]School of Mechanical Engineering and Automation, Northeastern University, Shenyang 110004 China
Email: yangyingsy@163.com

ABSTRACT

A CCD camera was installed in the cab to monitor the driver if he/she fatigue or drowse, especially at night. The technology of driver fatigue monitoring plays an important role in reducing the rate of traffic accidents. It is difficult to recognize the driver's state because the facial image is fainter under somber circumstance. A new method is proposed to detect and location driver's face in order to judge if he/she is drowse. By using the method of combining image corrosion and image expansion get rid of the image noise. Then strengthening facial grey value through nonlinear curve normalized to obtain a proper gray value. Project the facial image to horizontal and vertical to get the driver's face location. It is the key to monitor the driver. The method can be applied in dusky light to locate the face, effectively remove the shadow around interference. Using Visual C + + 6.0 development a program and a test was carry under different environmental and road. The results show that, under the somber light, the driver faces positioning accuracy rate is 87.3% with the property of real-time and high accuracy. This method provides a sound foundation to further identify the driver's fatigue at night.

Keywords: Driver, Fatigue Monitor, Facial Image, Gray Strengthen, Computer Vision

1. INTRODUCTION

In the frequent occurrence of accidents, driver fatigue has become one of the major factors that lead to traffic accidents. Intelligent Transportation System auxiliary driving skills as the key technology increasingly people's concern. The technology of driver fatigue monitoring plays an important role in reducing the rate of traffic accidents. Role in the normal menstrual cycles, fatigue is the most vulnerable people in the night is the most prone to traffic accidents. The number of accidents accounted for 70% of the proportion of the incidents. Release can be seen at night as well as other incidents during the day time is far greater than the release of the accident probability, Research under the driver's face lit light detection is essential and facial image recognition technology is the key, face detection and recognition for the current study limited to the daytime, many people consider face monitoring of the day [1-2]. a more satisfactory results for the identification and treatment of its black face features of small, Role in the normal menstrual cycles, fatigue in the most vulnerable people, are the most prone to accidents. Due to the uniqueness of professional drivers, monitor and advise the night even more important. Black day for daytime image detection algorithm is not applicable, even unable to judge. In this paper a suitable night, lit the pilot light facial positioning algorithm Driver fatigue monitoring technology have expanded the scope of the pilot study at night monitoring system play an important role in reducing the rate of traffic accidents.

2. GET RID OF IMAGE NOISE

The actual image is subjected to random error signal degradation and the impact of the degradation of noise, Image denoise study, this paper, a binary processing and filtering, Corrosion and swelling dispels the image by using some noise.

2.1 The Method of Erosion

The As shown in Fig.1, X as a "十" in the form of binary data sets, left the space coordinates (0,0). B structural elements in (0.0) modules for the B nuclear Bx. X-axis can accommodate the vertical structural elements under B, which came to root axis on the corrosion off. B structural elements are in the nuclear Bx (0. 0) position. So the level of corrosion results of X-axis x-episode left composed of three modules. Corrosion or displacement vector computation algorithm can be achieved.

Vector computing to achieve corrosion

$$X\Theta B=\{x\in E^{n}\mid(x+b)\in X,\ \ b\in B\} \tag{1}$$

Corrosion and displacement achieved by the operation:

$$X\Theta B=\bigcap_{b\in B}X_{-b} \tag{2}$$

In other words, pool B to X structural element to the results of the X-corrosion conducted with all the negative b after the simultaneous displacement. This will eliminate redundant Burr, the result of more tactful handling, a better image.

$$X \quad \Theta \quad B \quad = \quad \{x:B_{x}\subseteq X\}$$

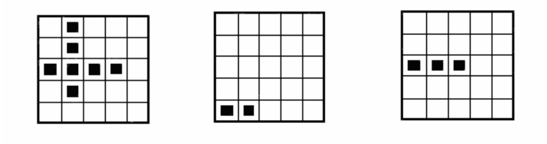


Fig.1. The example of erosion

2.2. The Method of Corrosion

As shown in Fig.2, X as binary data sets, left the space coordinates (0, 0). B is a structural element, here and beyond, unless otherwise stated, are located in (0, 0) elements B nuclear Bx. Bx exchanges with X or B is not empty after two hit X formed by x 4.4 Medium its plans to collect the data. Results are compared with the original X has been the pattern by expanding the scope of a certain Therefore, the operator has been called inflated. Expansion is computational algorithm, and their displacement is the vector operations.

$$X \quad \oplus \quad B \quad = \quad \{x:B_{x}\cap X\neq\phi\}$$

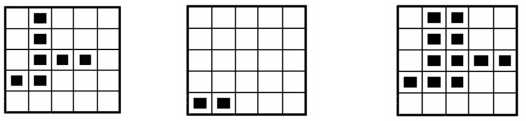
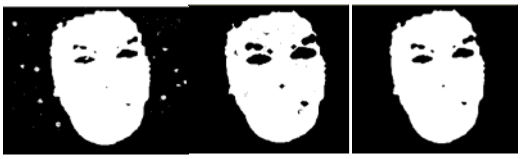


Fig.2. The example of corrosion

* National natural science fund. No 10402008.

Fig.3 shows the results of corrosion image processing, background noise is removed from the region, as illustrated in Fig.3 (b) below, After dilation operator then face regional noise is removed.



(a) (b) (c)

(a) After the binary image (b) Corrosion (c) Expansion

Fig.3. The result of erosion and dilation

3. THE THRESHOLD ALGORITHM

As in the days of black driver monitoring system inadequate lighting, color image to black-and-white image in the process. Because the light that caused the background color change is not so clear and very dark, so the lips color will be dark. If the light increased, As the white light caused by the introduction of color face most of the Gaussian mixture model is no longer meet the color, White therefore not be used to introduce the approach. Ray fainter light of the circumstances, the gray image was enhanced with gray; Binary images obtained the appropriate threshold solution.

After the image is usually transform color space and color selection can be a model of the gray-scale image, Binary images of gray, the gray-scale image can be separated, This paper studies the drivers face in gray-scale image, the greater is the color gray value, the greater the likelihood. But if choose a fixed numerical gray to distinguish, it is difficult to meet different circumstances. Therefore, we must adopt an adaptive method to determine the threshold value of the grey. The threshold differences value was determined by using image-based methods.

The maximum fuzzy information entropy method was to obtain the differences in the threshold method. In Information theory, information accompanied by uncertainty, quantitative characterization of the information is not necessarily tied to the uncertainty of measurement. Information entropy from the definition, we can see that the size of image information entropy represents the image level, the greater the entropy, the more abundant levels. For the whole image, target and background can be split at the junction of the largest entropy. Moreover, its partial information entropy reflects the partial information which contains the size so local information entropy can describe the partial nature of the images, if the local existence brink the dramatic changes were partially gray value, the corresponding information entropy smaller. Image entropy has been widely used in image retrieval. This paper will be applied to the image information entropy to the detection of the target vehicle. This method can be found through the energy of the most informative images of the target area, located on the face. The method has been regarded as the optimal threshold value will automatically choose one. The main advantage is easy-to-read, but not in the quantity of programming under certain conditions from the impact of image contrast and brightness, inadequate lighting and poor image contrast in the treatment system has been widely applied. The high processing speed, are well suited to high real-time performance and poor light conditions. Specific principles and analysis methods are as follows.

A_k Face the driver used a gray-scale image, the subscript representative of the image pixel gray. According to the theory of fuzzy, gray images can be viewed as a vague, by fuzzy gray values can be divided into two pools. One is the black fuzzy sets, a white fuzzy set. Low black fuzzy sets contain the pixel gray value high white fuzzy sets contain the pixel gray value. This is the membership function for the two fuzzy sets, black and white fuzzy $\alpha_d(k)$ and $\alpha_b(k)$ fuzzy sets. The probability of occurrence of two fuzzy sets

$$p_d = \sum_{k=0}^{N} (p(k) \times \alpha_d(k)) \tag{3}$$

$$p_b = \sum_{k=0}^{N} (p(k) \times \alpha_b(k)) \tag{4}$$

$p(k)$ Pixel points in the image.

Fuzzy sets and fuzzy black and white fuzzy entropy can be used formulas (5) and (6) calculated as follows:

$$H_d = -p_d \ln(p_d) - (1 - p_d)\ln(1 - p_d) \tag{5}$$

$$H_b = -p_b \ln(p_b) - (1 - p_b)\ln(1 - p_b) \tag{6}$$

Fuzzy is the total entropy

$$H = H_d + H_b$$

By the information, it found that the greater the entropy, which contains information on the bigger, and in order to achieve objectives and the best background segmentation. Fuzzy entropy will be the largest, Image segmentation is the largest amount of information contained in the original image, and the total achieved maximum fuzzy entropy, solution available through the optimal threshold value.

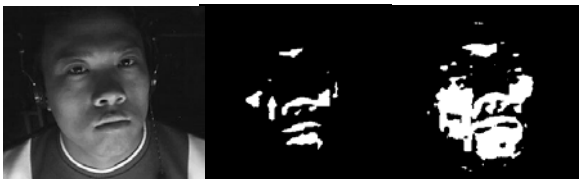


Fig.4. black days under a gray value calculation curve

Only under the conditions in the external environment, more or less, the threshold value can be selected based on differences in image processing method. Obviously in the light of the black days in almost select the threshold selection method is appropriate. Based on the black days of threshold selection method is the most used method threshold value can be inserted in the experimental results shown in Fig.4.

4. GRAY STRENGTHEN ALGORITHM

Gray is certain to enhance the image gray little change between various gray values. the result will not be a good image of the object and the background is inseparable from the good, using some algorithm will expand the value of gray-scale images we all multiples images of the object and the background makes the difference obvious change, the goal has been relatively accurate.

The simplest way to do is to strengthen gray linear expansion [3] [4] gray. In previous image processing method has been widely used in the above specific approach is : First, a set of image and the image of the largest gray value of the minimum gray value $G_{\max}$ and $G_{\min}$, then through a linear transformation of the original gray image gray value of the variable value range between 0 ~ 255. g A new gray value, and it makes the gray values are very similar and assembled into a larger difference between the gray values set. Collective is shown as formula (7).

$$g = \frac{255 - 0}{G_{\max} - G_{\min}} \times (G - G_{\min}) \qquad (7)$$

This method is relatively large value of the pixels under the circumstances turn good results. But even small changes in the gray image some cases we will not be able to solve the problem of binary.

Map fainter light can be observed in the environment, Images of many of the objectives of the regional background of the region that was the wrong computer being removed. Under a human face on this situation is very unfavorable position, therefore, a better approach to the expansion of gray values of the selections. Think of the fainter light in the environment, in the face of regional gray image should have been more obvious characteristics. Gray further expanded its focus on the difference between the gray values. Gray curve equation for a one-time non-use value, the specific curve shown in Fig.5.

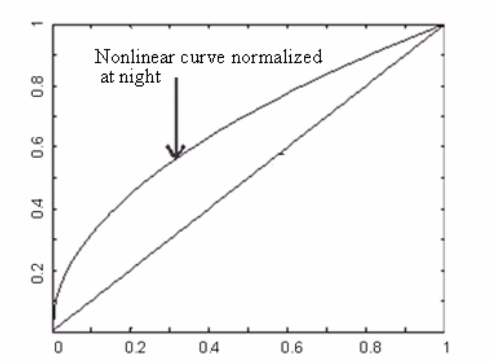


Fig.5. Nonlinear curve normalized method

Fig.5 can be found in the proximal part of the rapid curve close to 1, enables the left part of the rapid rise thus opening up the background and objectives of the size of the gap between the gray values. The following is a function of the gray $y = \sqrt{x}$ for strengthening the use of the specific steps.

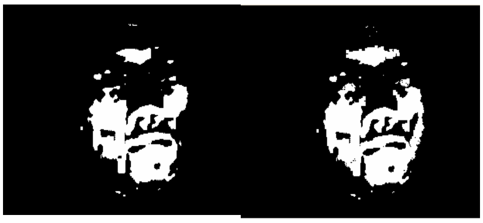


Fig.6. The strengthen picture using one-order gray method

(1) A set of the greatest images of the gray value of $G_{\max}$ and the minimum gray scale images $G_{\min}$;

(2) Under a gray value of G for processing. $g = G / G_{\max}$;

(3) g Will be a new normalized intensity value of $g^{'}$.

$$g' = \sqrt{g}$$

The experimental results were shown in Fig.6. In which we can see the facial region has been completely separated from the region and background, this can be used for facial segmentation on the specific knowledge of the facial map. Gray experiment at the same time strengthen the use of the above mentioned can be found on the situation in poor light [6, 7] image introduced under the white face can detect the precise location of a good phenomenon has been resolved. The experimental results are shown in Fig.7.

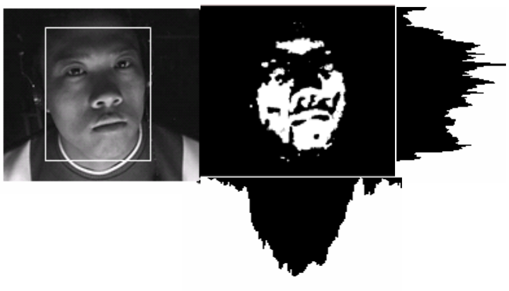


Fig.7. Black days under a gray value calculation curve

5. THE EXPERIMENT OF FACE LOCATION AND RESULTS ANALYSIS

In order to adapt different obscuration light, a CCD camera was fitted in a cab under the following occasions to collect the driver's face images.

(1) Dusk and there are no traffic lights;
(2) Night at the city center, there are traffic lights;
(3) Night at speedway, there are traffic lights;
(4) Dawn, there are traffic lights;
(5) Night at country side, there are no traffic lights.

The results of test results compare to original image, shown as table 1.

Table 1. Compare images of experiment result to original

occasions	Right images/ Total images	Accurate rate%
(1)	18/20	90
(2)	57/62	92
(3)	73/88	83
(4)	46/52	89
(5)	34/71	46

Totally 293 images of drivers' face were collected. Using above facial locating arithmetic to analysis process. The location results was shown as table 1, under condition (1) to (4), the right results is 194 images in 222images, its accurate is 87%, but at occasion (5), because the light is very obscuration, the accurate is only reach 46%.

So the cab's light beam is the main reason to infect recognize accurate.

6. CONCLUSIONS

The In this paper, the uniqueness of professional drivers, and the high incidence of traffic accidents at night characteristics in dusky light, the driver under surveillance, a face positioning algorithm. By corrosion and prevent further expansion of the image noise and enhance an algorithm based on linear gray, quadratic nonlinear equation for strengthening of the gray values, Gray concentrate on improving the image of the face value of the difference between the gray, and under different environmental and traffic control test The results showed dark, Drivers face positioning accuracy rate of 83%, with real-time, high accuracy, To further identify the pilot fatigue at night it provides an important basis for the judgment.

REFERENCES

[1] Hsu. R. L, Mottaleb. M. A, Jain. A. K, “Face detection in color image[J],” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(5):696~706.

[2] H. Greenspan, J. Goldberger, I. Eshet, “Mixture model for face-color modeling and Segmentation [J],” *Pattern Recognition Letters*, 2001, 22(14): 1525-1536.

[3] Shiguang Shan, Bo Cao, Wen Gao, Debin Zhao, “Extended Fisherface for Face Recognition from a Single Example Image per Person [J],” *IEEE International Symposium on Circuits and Systems*, 2002, 2(10): 81-84.

[4] MIAO Jun, YIN Baocai, WANG Kongqiao, “A hierarchical multistage and multiangle system for human face detection in a complex background using gravity-center template [J] ,” *Pattern Recognition*，1999，32(7): 1237-1248.

[5] R. F. eraud, O. J. Bernier, J. E. Viallet, M. Collobert. A Fast and Accurate Face Detection Based on Neural Network [J], IEEE Trans. Pattern Analysis and Machine Intelligence, 2001, 23(1): 42-53.

[6] Hong Liu, Wen Gao, Jun Miao, Jintao Li, “A Novel Method to Compensate Variety of Illumination in Face Detection [C],” *Proc. 6th Joint Conference on Information Sciences*, North Carolina, 2002, 13(3): 692-695.

[7] T. F. Cootes, G. J. Edwards, C. J. Taylor, “Active appearance models [J],” *Proc. European conference on computer vision*, 1998: 484-498.

Vehicle Recognition Based on Support Vector Machine

Tongze Xue [1], Kuihe Yang [1], Xingxia Niu [1]
[1]College of Information, Hebei University of Science and Technology, China
No.505, Xinhua Road, Xinhua District, Shijiazhuang, 050054 China
Email: ykh86198986@126.com

ABSTRACT

In the paper, a vehicle recognition model based on Support Vector Machine (SVM) is presented. SVM can solve the problem of nonlinear well, avoiding some difficulties including high dimensional and local minimum. This paper applies the multi-classification method based on Support Vector Machine to vehicle recognition. Support vector machine is a new theory and technology in the filed of pattern recognition and has shown excellent performance in practice. This method was proposed basing on Structural Risk Minimization (SRM) in place of Experiential Risk Minimization (ERM), thus it has good generalization capability. By mapping input data into a high dimensional characteristic space in which an optimal separating hyperplane is built, SVM presents a lot of advantages for resolving the small samples, nonlinear and high dimensional pattern recognition, as well as other machine-learning problems such as function fitting. The simulation results show the model has strong non-linear solution and anti-jamming ability, and can effectively distinguish vehicle type.

Keywords: Statistical Learning Theory (SLT), Support Vector Machine (SVM), vehicle recognition

1. INTRODUCTION

Along with the fast development of modern traffic, there are number more and more number and kind of vehicle, which needs to identify vehicle type automatically. Vehicle recognition has important meaning for confirming speedway fee, managing large park and watching road traffic and it is important composing part of intelligent traffic system. At present, vehicle recognition research is at original step, and it is a hotspot of study.

In the face of growing traffic congestion, limited resources and financial and environmental pressures, building more infrastructures will be constrained. In the study of this problem, some countries found that the introduction of the pattern recognition and electronic information technology to the system of transport, not only could solve the traffic jams, but also have an impact on traffic safety, traffic accidents and the handling of relief, passenger and freight traffic management and highway toll system, therefore, they continue to expand the research, development and the range of testing, and thus Intelligent Transport System (ITS) emerged.

As the investment of highway construction is large, charging for loan has already been taken. However, highway traffic capacity is greatly reduced by highway toll stations. Abroad, there are examples for charging without parking and it is our direction too. Electronic toll systems embody the ITS, and their applications can solve the "bottlenecks" of traffic charging stations, as well as traffic jams, waiting, environmental pollution and other issues.

At present, the vehicle recognition techniques mainly include the method of outline scanning, axle counting, the magnetic field changing, License plate recognition, as well as the method based on the image processing, the method based on the traffic video and some other methods. Generally speaking, most of the vehicle recognition technique that developed in recent years induct or classify through detecting certain geometry or physical parameters of vehicles, some of which have achieved relatively high accuracy. However, of all those techniques about vehicle recognition and classification, the ones can be used for the moment are very few as a result of various reasons. Because vehicle recognition and tolling built on the type of vehicle are important to highway toll, it is necessary to develop some new methods for vehicle recognition as soon as possible.

Support vector machine is an outstanding method for pattern recognition, which has been taken wide attention to [1]. It is based on Statistical Learning Theory (SLT)[2],and has complete theory and splendid performance. In this paper, we introduce support vector machine to the vehicle recognition.

2. INTRODUCTION ABOUT SVM

The support vector machine is a new machine study method which was established in base of statistical learning theory. The SVM stresses to study statistical learning rules under small sample. Via structural risk minimization principle to enhance extensive ability, the SVM preferably solves many practical problems, such as small sample, non-linear, high dimension number and local minimum points. The SVM has been applied in pattern classification, regression forecasting, probability estimation, control theory and so on.

Support vector machine proposed by Vapnik and his co-workers is a novel technique for classification. The basic principle of SVM is finding the optimal linear hyperplane such that the expected classification error for unseen test samples is minimized. Based on this principle, a linear SVM uses a systematic approach to find a linear function with the lowest VC dimension. For linearly non-separable data, SVM can map the input to a high-dimensional feature space where a linear hyperplane can be found. So, good generalization can be achieved by SVM compared with conventional classifiers. In recent years, SVM has been successfully applied to various tasks in face classification problems.

1.1 The Basic Theory of Support Vector Machine

First let us look at the linear support vector machine. It is based on the idea of hyperplane classifier, or linearly separability. Suppose we have training data points: $T = \{(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \cdots, (\mathrm{x}_l, y_l)\}$, where $\mathrm{x}_i \in R^n$, and $y_i = \{1, -1\}$, $i = 1,2,3 \cdots l$. We would like to learn a linear separating hyperplane classier:

* This work is supported by Science Foundation of Hebei University of Science and Technology (2006021).

$$f(x) = \mathrm{sgn}(\omega \cdot \mathrm{x} + b) \qquad (1)$$

Furthermore, we want this hyperplane to have the maximum separating margin with respect to the two classes. Specifically, we want to find this hyperplane $H: y = \omega \cdot \mathrm{x} + b = 0$ and two hyperplanes parallel to it and with equal distances to it:

$$H_1 : y = \omega \cdot \mathrm{x} + b = +1, \qquad (2)$$

$$H_2 : y = \omega \cdot \mathrm{x} + b = -1 \qquad (3)$$

With the condition that there are no data points between H_1 and H_2, and the distance between H_1 and H_2 is maximized (see Fig.1).

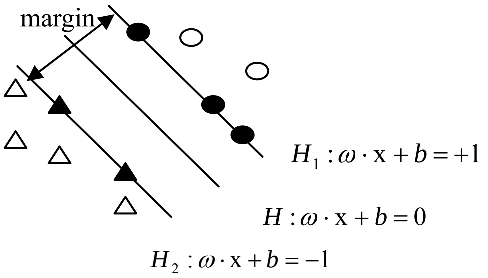

margin
$H_1 : \omega \cdot \mathrm{x} + b = +1$
$H : \omega \cdot \mathrm{x} + b = 0$
$H_2 : \omega \cdot \mathrm{x} + b = -1$

Fig.1. Linear separating hyperplanes for the separable case.

For any separating plane H and the corresponding H_1 and H_2, we can always "normalize" the coefficients vector ω so that H_1 will be $y = \omega \cdot \mathrm{x} + b = +1$ and H_2 will be $y = \omega \cdot \mathrm{x} + b = -1$.

We want to maximize the distance between H_1 and H_2. So there will be some positive examples (black rotundity in Fig.1) on H_1 and some negative examples (black triangle in Fig.1) on H_2. These examples are called support vectors because only they participate in the definition of the separating hyperplane, and other examples can be removed and/or moved around as long as they do not cross the planes H$_1$ and H$_2$. As it is known, the distance between H_1 and H_2 is $\mathrm{margin} = 2/\|\omega\|$. So, in order to maximize the distance, we should minimize $\|\omega\| = \omega^T \omega$ with the condition that there are no data points between H_1 and H_2:

$\omega^T \mathrm{x}_i + b \geq +1$ for positive examples $y_i = +1$,

$\omega^T \mathrm{x}_i + b \leq -1$ for negative examples $y_i = -1$.

These two conditions can be combined into:

$$y_i(\omega^T \mathrm{x}_i + b) \geq 1 \qquad (4)$$

If the two classes are not linear, we can transform the data points to another high dimensional space such that the data points will be linearly separable. Let the transformation be $\phi(\cdot)$ (see Fig.2). Suppose, $\phi(\mathrm{x_i})\phi(\mathrm{x_j}) = K(\mathrm{x_i}, \mathrm{x_j})$. That is, the dot product in that high dimensional space is equivalent to a kernel function of the input space. So we need not be explicit about the transformation $\phi(\cdot)$ as long as we know that the kernel function $K(\mathrm{x_i}, \mathrm{x_j})$ is equivalent to the dot product of some other high dimensional space. Thus the separating hyperplane is replaced by $\omega \cdot \phi(\mathrm{x}_i) + b = 0$.

The other direction to extend SVM is to allow for noise, or imperfect separation. That is, we do not strictly enforce that there be no data points between H$_1$ and H$_2$, but we definitely want to penalize the data points that cross the boundaries. The penalty C will be finite. We introduce non-negative slack variables $\xi_i \geq 0$, so finally, our problem can be formulated as

$$\text{Min} \quad 1/2\omega^T \omega + C\sum \xi_i \qquad (5)$$

$$\text{Subject to } y_i(\omega^T \phi(\mathrm{x}_i) + b) \geq 1 - \xi_i, \; \xi_i \geq 0 \qquad (6)$$

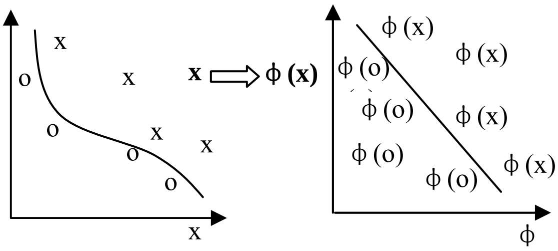


Fig.2. The illustration about mapping from non-linear space to high dimensional linear one.

This is a convex, quadratic programming problem (in ω, b), in a convex set. In order to get the global optimum, we can translate the original problem to a dual problem with the help of Lagrange, and this means that we can equivalently solve the following "dual" problem:

Max

$$Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \phi(\mathrm{x}_i)\phi(\mathrm{x}_j) + \sum_{i=1}^{l}\alpha_i \qquad (7)$$

$$\text{Subject to} \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \qquad (8)$$

$$0 \leq \alpha_i \leq C, i = 1,2,3 \cdots l \qquad (9)$$

When we solve α_i, we can get $\omega = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathrm{x}_i)$, and we can classify a new object x with:

$$f(\mathrm{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i \phi(\mathrm{x}_i)\phi(\mathrm{x}) + b\right)$$
$$= \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathrm{x}_i, \mathrm{x}) + b\right) \qquad (10)$$

In the least-square SVM (LS-SVM)[5], proposed by Suykens and Vandewalb, the optimization criterion adapt quadratic term, and there are only equality constraints, but not inequality constraints, and then a series of equality constraints can be deducted, in place of quadratic programming. In this case, our problem can be formulated as:

$$\text{Min} \quad \frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma\sum_{i=1}^{l}\xi_i^2 \qquad (11)$$

$$\text{Subject to} \quad y_i(\omega^T \phi(x_i) + b) = 1 - \xi_i \qquad (12)$$
$$i = 1,2,3 \cdots l$$

System of linear equations as follow can be gotten:

$$\begin{bmatrix} 0 & y^T \\ y & Q + \gamma^{-1} I \end{bmatrix}_{(l+1)\times(l+1)} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ e \end{bmatrix} \qquad (13)$$

The above formula can be solved by least square. In the method of LS-SVM, the transformation from quadratic programming to system of linear equations simplifies the computational complexity.

2.2 The Problem about Multi-Classification

SVM was proposed for two-class problems originally, and multi-class problems can be obtained by combining two-class SVM. Usually there are two methods: one-against-all [7] and one-against-one [7].

The earliest used implementation for SVM multi- classification is probably the one-against-all method. It constructs M SVM models where M is the number of classes. The ith SVM is trained with all of the examples in the ith class with positive labels, and all other examples with negative labels. Thus given training data $(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \cdots, (\mathrm{x}_l, y_l)$, where $\mathrm{x}_\mathrm{j} \in R^n$, $j = 1,2,3\cdots l$ and $\mathrm{y}_\mathrm{j} \in \{1,2,\cdots M\}$ is the class of x_j, the ith SVM solves the following problem:

$$\text{Min} \quad 1/2(\omega^i)^T \omega^i + C\sum_{j=1}^{l} \xi_j^i \tag{14}$$

$$(\omega^i)^T \phi(\mathrm{x}_j) + b^i \geq 1 - \xi_j^i, \text{ if } y_j = i, \tag{15}$$

$$(\omega^i)^T \phi(\mathrm{x}_j) + b^i \leq -1 + \xi_j^i, \text{ if } y_j \neq i, \tag{16}$$

$$\xi_j \geq 0, \ j = 1,2,3\cdots l \tag{17}$$

Another major method is called the one-against-one method. This method constructs M(M-1)/2 classifiers where each one is trained on data from two classes. For training data from the ith and the jth classes, we solve the following binary classification problem:

$$\text{Min} \quad 1/2(\omega^{ij})^T \omega^{ij} + C\sum_{t=1}^{l} \xi_t^{ij} \tag{18}$$

$$(\omega^{ij})^T \phi(\mathrm{x}_t) + b^{ij} \geq 1 - \xi_t^{ij}, \text{ if } y_t = i, \tag{19}$$

$$(\omega^{ij})^T \phi(\mathrm{x}_t) + b^{ij} \geq -1 + \xi_t^{ij}, \text{ if } y_t \neq j, \tag{20}$$

$$\xi_t \geq 0, \ t = 1,2,3\cdots l \tag{21}$$

There are different methods for doing the future testing after all M(M-1)/2 classifiers are constructed. The voting strategy is described as: if $\left((\omega^{ij})^T \phi(\mathrm{x}_t) + b^{ij}\right)$ says x is in the ith class, then the vote for the ith class is added by one. Otherwise, the jth is increased by one. Then we predict x is in the class with the largest vote. The voting approach described above is also called the "Max Wins" strategy.

3. EXPERMENT

In the electronic toll system, it needs to diagnosis the vehicle automatically according to different vehicles as a result of the automation realizing. Because passing by the toll station directly without stopping, accurate and rapid recognition, classifications are necessary.

The data for this experiment about vehicle recognition come from public dataset http://www.csie.ntu.edu.tw/ ~cjl-in/libsvmtools/datasets/multiclass.html. The original purp- se of the vehicle dataset was to find a method of distinguishing 3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects. The images were acquired by a camera looking downwards at the model vehicle from a fixed angle of elevation (34.2 degrees to the horizontal). All images were captured with a spatial resolution of 128x128 pixels quantized to 64 grey levels. The features were extracted from the silhouettes by the HIPS (Hierarchical Image Processing System) extension BINATTS, which extracts a combination of scale independent features utilizing both classical moments based measures such as scaled variance, skew ness and kurtosis about the major/minor axes and heuristic measures such as hollows, circularity, rectangularity and compactness.

There are four types of vehicles: OPEN, SAAB, BUS and VAN, we sign them separately for 1, 2, 3 and 4. The total of the examples is 846, and 212 for model 1, 217 for model 2, 218 for model 3 and 199 for model 4.

This experiment is built on LS-SVM. We choose the Redial Basis Function (RBF) as the kernel function. In order to make an LS-SVM model, we need 2 extra parameters: gamma (gam) is the regularization parameter, determining the trade-off between the fitting error minimization and smoothness. In the common case of the RBF kernel, sigma^2 (sig2) is the bandwidth. Setting up the parameters: gam=10, sig=0.2.

Training the support vector machine using 846 samples, the model of support vector machine can be obtained. Then, 300 and 500 data extracted separately from the original samples are appended disturbance factor, and we take them for simulating data. The results of this experiment are shown in table 1. Note that "a" is disturbance factor, "perc" is error rate, "n" is the number of mistake and "acc" is simulating accuracy.

From the results we can see that when the disturbance factor is 0, the support vector machine can diagnose the samples completely, and when the disturbance factor is controlled within 10%, we can also obtain better accuracy. The method not only be able to achieve high accuracy but also realized simply.

Table 1. The results of experiment

a	Number:300			Number:500		
	perc	n	acc	perc	n	acc
0	0	0	100%	0	0	100%
0.08	0	0	100%	0	0	100%
0.085	0	0	100%	0.0020	1	99.8%
0.087	0	0	100%	0.0120	6	98.8%
0.09	0.0367	11	96.33%	0.0280	14	97.2%
0.095	0.0667	20	96.33%	0.0500	25	95%
0.10	0.1067	32	89.33%	0.0900	45	91%
0.105	0.1433	43	85.67%	0.1260	63	87.4%
0.11	0.2033	61	79.67%	0.1700	85	83%
0.12	0.3000	90	70%	0.2700	135	73%

4. CONCLUSIONS

There is a very big market for vehicle recognition. This recognition technology may apply in the management of parking charging as well as automatic charging system for each kind of enclosed and open style highway and bridge., It is suitable for each kind of automatic toll, and according to the networking scale, may expand from the road section to the region, the area, trans-region down to the national road network. Therefore, it is imperative to take thorough studies on this technology. In view of its good learning performance, applying support vector machines in the vehicles recognition is a research direction which has a very splendid prospect.

The relatively mature development of SVM is made only in recent years, but it has shown good performance in machine learning and pattern recognition and has won extensive research and application. Thus, this algorithm will provide new way to practical classifications and further studying about SVM has tremendous theory and application value.

REFERENCES

[1] Burges C J C, "A Tutorial on Support Vector Machines for Pattern Recognition" [J], *Knowledge Discovery and Data Mining*, 1998, 2(2):121-167.

[2] Vapnik, *The nature of statistical learning theory*. NewYork: Spinger-Verlag, 1995.

[3] S. R. Gunn, "Support Vector Machines for Classification and Regression," *Technical Report, Image Speech and Intelligent Systems Research Group*, University of Southam- pton, 1997.

[4] Weida Zhou, Li Zhang and Licheng Jiao, "Linear programming support vector machines," *Pattern Recognition*, Volume 35, Issue 12, December 2002, Pages 2927-2936.

[5] Suykens J.A.K, "Vandewalle J.and De Moor B. Optimal Control by Least Squares Support Vector Machines" [J]. *Neural Networks*, 2001, 14 (1): 23-5.

[6] Suykens J, Vandewalle J, Least square support vector machine classifiers, Neural Processing Leters, 1999, 9(3): 293-300.

[7] Hsu Chih-Wei, Lin Chih-Jen. "A comparison of methods for multi-class support vector machine," *IEEE Transactions on Neural Networks* 2002, 13(2):415-425.

[8] Kuh A, "Analysis of detectors for support vector machines and least square support vector machines," *Neural Networks*, 2002. IJCNN '02. *Proceedings of the 2002 International Joint Conference on*, Volume: 2, 2002. Page(s): 1075-1079.

Tongze Xue is an associate professor of Hebei University of Science and Technology.

The Algorithm and Complexity Analyses of Relative Gradient-based Adaptive Image Fractal Compression*

Wenjing Li[1], Qingping Guo[2], Rongwei Huang[1]
[1]Department of Information Technology, Guangxi Teachers Education University, Nanning, 530001, China
Email:liwj@gxtc.edu.cn
[2]Department of Computer Science, Wuhan University of Technology, Wuhan, 430063, China
Email: qpguo@mail.whut.edu.cn

ABSTRACT

This paper introduced the image fractal compression technical basic theories based on Iterated function system. It makes use of self-similarity feature from an original image and its relative gradient image. The method of adaptive quad-tree partitioning is adopted. We give an algorithm of relative gradient-based adaptive image fractal compression. And the algorithmic parallel processing is discussed .The algorithmic complexity analyzed indicates that the method raised compression ratio of image fractal compression, compute quantity is little, and the efficiency is higher.

Keywords: Fractal compression, Iterated function system, adaptive partitioning, relative gradient, algorithmic complexity.

1. INTRODUCTION

In 1988 BANSLEY made fractal compressed encoding on a few particular images, He gained the compression ratio of 10000:1.Although the encoding process demands manual participation, it shows huge potential of fractal technique in image compressed encoding. In 1990, JACQUIN advanced block-based project of Iterated function system (IFS), in which encoding process can progress automatically [1]. But JACQUIN’s fractal encoding had a huge calculation and long time-consumption. Thus its practicality was limited [2]. On his foundation, people put forward many methods of image fractal compressed encoding, to improve the speed of image fractal encoding and resolve the tile effect problem caused by decoding. This paper makes use of an original image and the self-similarity feature [3] of its relative gradient diagram. It adopts the method of quad-tree partitioning, blends two encoding techniques, and advances the algorithm of relative gradient-based adaptive image fractal compression. And the algorithmic parallel processing is discussed .The algorithmic complexity analyzed indicates that the method raises compression ratio of image fractal compression. Thus computation is reduced and efficiency increased.

2. THE BASEIC THEORY OF THE IFS IMAGE FRACTAL COMPRESSION ENCODING

MANDELBROT revealed the essential feature of fractal and defined fractal geometry theoretic frame [2]. BANSLEY and SLOAN found the general course that can simplify image into a series of affine transformation, and defined iterated function system, which lays foundation for image fractal compression.

2.1 Affine Transformation and Compression Commutation

Definition 2.1 Suppose ω is $R^n \to R^n$ affine transformation： To any $x=(x_1,x_2,\cdots,x_n)^T \in R^n$, have:

$$w(x)=Ax+t \qquad (2.1)$$

among, $A=(a_{ij})$ is n × n non-strange matrix , $t=(t_1,t_2,\cdots,t_n)^T \in R^n$ is constant vector.if norm $\|A\|<1$, then affine transformation ω defined by（2.1）is a compress transformation

A gray scale image that posses gray-level two-dimensional array, then $z=f(x,y)$, among (x,y) is space position, Z is the relevant position tonal value. For the sake of adapting the process of gray scale image, expand 2D affine transformation to 3D affine transformation.3D affine transformation that is used for gray scale image ω : $R^3 \to R^3$ be expressed as follows:

$$w\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{bmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & s \end{bmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} e \\ f \\ o \end{pmatrix} \qquad (2.2)$$

Then this affine transformation can be regarded as the combination of 2D affine transformation on (x,y) plane and the gray scale transformation of Z direction. Among, S control the contrast degree of gray scale; O control the offset of gray scale.

Theorem 2.1 (compression transformation Theorem) Suppose (X,d) is a complete metric space, ω :X → X is a compression transformation of X. Then ω has the only fixed point x_w , $\omega(x_W)=x_W$ .And for any x ∈ X, The sequence $\{w^{0n}(x):n=0,1,2,\cdots\}$ converges on x_w .

Theorem 2.2 Suppose $\{w_n,n=1,2,\cdots,N\}$ is a set compression transformation of $(H(X),h(d))$, for every n, compression factor of W_n is c_n , definition $\omega: H(x) \to H(x)$

$$W(B)=w_1(B)\cup w_2(B)\cup\bullet\bullet\bullet\cup w_n(B) \ \forall B\in H \qquad (2.3)$$

Then $\omega:H(x)\to H(x)$ is a compression transformation, compression factor $c=\max\{c_n:n=1,2,\cdots,N\}$.

2.2 Iterated Function System and Collage Theorem

Definition 2.2 Suppose (X,d) is a complete metric space, The B is any non-empty tight subset, also $B\in H(X)$, for any given $\varepsilon>0$, there exists a set of limited compression transformation $\omega_n:X\to X,n=1,2,...,N$, making attractor A in the

* Supported by Department of Education of Guangxi zhuang Autonomous Regional (0626120) and Guangxi Teachers Education University (0604A005).

corresponding iterated function system (IFS): $\{X;\omega_i,i=1,2,...,N\}$ satisfy $h(B,A)<\varepsilon$

Among them, $h(d)$ is distance of HAUSDORFF.

Theorem 2.3 (collage Theorem) Suppose the definite compression transformation in iterated function system $\{X;\omega_i,i=1,2,...,N\}$ is ω, $W(B)=\bigcup_{n=1}^{N}\omega_n(B),\forall B\in H(X)$, its compression factor is c, attractor is A, then for X any not-empty tight subset $B\in H(X)$, the follow formula is tenable.

$$h(B,A)\le\frac{1}{1-s}h(B,W(B)),\forall B\in H(X)\cdot$$

3. THE GENREATION OF IMAGE FRACTAL ITERATED FUNCTION SYSTEM

Suppose the given target image B is a gray scale image, then according to (2.2) formula, x, y is the coordinate of the pixel in the image, a,b,c,d,e,f,s,o are transformation parameters, and all are real numbers. We can use the eight transformation parameters to represent a compression affine transformation coding. If the IFS of an image are composed of the n compression affine transformations, then the IFS encoding of the image can be simplified into the encoding of the $8\times n$ transformation parameters and n transformations of IFS. Image fractal compression encoding process becomes the process of solving $8\times n$ transformation parameters which limits compression transformation of the IFS. But does how to find the n transformation parameters of image fractal IFS. Full searching methods are as follows [4][5]:

1) Divide size $2^N\times2^N$ gray scale image B into a sub-block whose size is $2^R\times2^R$ and does not overlap each other R_i (i=1,2,$\cdots$, $2^{N-R}\times2^{N-R}$),then all block of range value constitutes R block pool. Divide B into father-block $D_i(i=1,2,\cdots,2^{N-D}\times2^{N-D})$ whose size is $2^D\times2^D$ and overlap each other, all definition area blocks constitute father-block pool D.
2) Average from four points of $2^D\times2^D$ pixel of father-block pool D in a gray scale image into one point, converting it into $2^R\times2^R$ pixel image of R pool.
3) Change the averaged gray scale image with 8 different kinds of transformation; get 8 gray scale images of $2^R\times2^R$ pixel. For a father-block in the father-block pool D, gain 8 different gray scale image of $2^R\times2^R$ pixel. Thus there are $8\times2^{N-D}\times2^{N-D}$ gray scale images whose sizes are $2^R\times2^R$ pixel in father-block pool D.
4) Select a certain block of range value R_i,based on the following equations(3.1),(3.2),(3.3) we can compute $8\times2^{N-D}\times2^{N-D}$ parameter set $\{s,o,H\}$ [5], find out of s and o which makes the H take a least value among them, record s , o values and its block of range value D_i and transformation T_i ,completing encoding of a block of range value , gain a corresponding transformation parameter of value range R_i .Supposing R_i gray scale value of each pixel point is $\alpha_1,\alpha_2,\cdots,\alpha_{2^R\times2^R}$. Supposing again gray scale image of $2^R\times2^R$ pixel after its definition domain block was averaged and converted, the corresponding gray scale value is $\beta_1,\beta_2,\cdots,\beta_{2^R\times2^R}$.

$$s_t=\frac{m\sum_{i=1}^{m}a_i\beta_i-\sum_{i=1}^{m}\alpha_i\sum_{i=1}^{m}\beta_i}{m\sum_{i=1}^{m}\beta_i^2-(\sum_{i=1}^{m}\beta_i)^2}\tag{3.1}$$

$$o=\frac{1}{m}(\sum_{i=1}^{m}\alpha_i-s\sum_{i=1}^{m}\beta_i)\tag{3.2}$$

$$H=\frac{1}{m}(\sum_{i=1}^{m}\alpha_i^2-s\ (s\sum_{i=1}^{m}\beta_i^2-2\sum_{i=1}^{m}\alpha_i\beta_i+2o\sum_{i=1}^{m}\beta_i)+o\ (mo\ -2\sum_{i=1}^{m}\alpha_i))\tag{3.3}$$

Among them, m= $2^R\times2^R$.

5) If the value domain sub-blocks number $i>2^{N-R}\times2^{N-R}$, then encoding is over, saving all transformation parameters. Otherwise turn step (4).

Analyzing the above algorithm we can discover, it is by comparing each value domain block R_i with definition domain block father-block pool D $8\times2^{N-D}\times2^{N-D}$ times that the best self-similar father block can be found. That's to say, to find the solution to a transformation parameter in the IFS, $8\times2^{N-D}\times2^{N-D}$ times of matching search and $2^{N-R}\times2^{N-R}$ transformations are needed. In other words, $8\times2^{N-D}\times2^{N-D}$ times are needed to find the solution to all IFS transformation parameters, and gain the encoding of the whole image. The computation is huge and the algorithm complexity is O(2^{4N})[6].

4. THE GENERATION OF ADAPTIVE RELATIVE GRADIENT IMAGES FRACTAL IFS

4.1 The Resemblance between Image and Its Relative Gradient

Theorem 4.1 suppose $f(x,y)$ has consecutive derivative in closed rectangle $[0,1]\times[0,1]$,and $f(0,y)=0$, $f(x,0)=0$,then

$$\int_0^1\int_0^1|f(x,y)|^2dxdy\le\frac{1}{4}\int_0^1\int_0^1|gradf(x,y)|^2dxdy\tag{4.1}$$

See the proving [3]

These theorem shows, if the gradient of the two images $f(x,y)$ and $g(x,y)$ is very small, then when the two sub-image gray scale images are similar; they are also similar to each other.

Definition 4.1 the definition of relative gradient of image $f(x,y)$ is as follows:

$$\hat{g}(x,y)=\frac{g(x,y)}{f(x,y)-\bar{g}(x,y)} \qquad (4.2)$$

Among, $\bar{g}(x,y)$ is the gray scale average of 4 pixel points, when denominator is zero, making it being equal to 1,

$$\bar{g}(x,y)=\frac{1}{4}\{f(x,y)+f(x+1,y)+f(x,y+1)+f(x+1,y+1)\} \qquad (4.3)$$

From the above formula (4.2) and (4.3) we can deduce: if both images are alike, and its relative gradient image is also alike. Therefore, if we know the original image and its relative gradient image are self-alike, we needn't consider the gray compressibility factor s and gray offset factor **o** in matching search. Just search the match piece whose average value and mean-square deviation in the gray scale image are similar, thus greatly reduce time consumption [3].

To reduce the searching time of each range piece in father-block pool and avoid computing compressibility factor and offset in the gray scale image, convert 3D commutation in the gray scale image into 2D commutation to increase the speed of image coding, [3] put forward high-speed algorithm based on fractal image of relative gradient picture, improve encoding efficiency and arithmetic speed .The algorithm adopts the partition method of fixed cent piece ,divides separately the gray scale image B and it's relative gradient image whose size is $2^N\times2^N$ into non-overlapping sub-block R_i （i=1,2,⋯, $2^{N-R}\times2^{N-R}$）, whose size is $2^R\times2^R$ and R_{gi} （i=1,2,⋯, $2^{N-R}\times2^{N-R}$）, get R and R_g block ponds constituted by all the value of original and relative gradient pictures; divide original image B and relative gradient image into overlapping father-block $D_i(i=1,2,\cdots,2^{N-D}\times2^{N-D}$) and $D_{gi}(i=1,2,\cdots,2^{N-D}\times2^{N-D})$ whose size is $2^D\times2^D$, thus all definition domain blocks constitute separately father-block pool D and D_g . Then set a searching radius K and errors threshold value ε of mean-square deviation. All sub–block value of relative gradient picture in father-block pool D_g is in the extent of $(\bar{D}g(m-k),\bar{D}g(m+k))$ and satisfies the errors threshold value ε block constitute search closed together region. As a sub-block R_{gi} finds a matching sub-block in relative gradient image father-block pool D_g , the corresponding sub-block R_i of the original picture also finds a matching sub-block in father-block pool D. The compressibility factor of the IFS image can be ensured. To match each range of value block we only need to search at most $2k$ sub-blocks in the $2^{N-D}\times2^{N-D}$ father-blocks. The average computation times of $2^{N-R}\times2^{N-R}$ sub-blocks are $T(N)=2k\times2^{N-D}\times2^{N-D}$. Suppose the given $k=\frac{1}{\rho}(2^{N-D}\times2^{N-D})(\rho\geq1)$, therefore $T(N)=\frac{2}{\rho}(2^{N-D}\times2^{N-D})\times2^{N-D}\times2^{N-D}$, the complexity of the algorithm is $O(2^{4N})$ .Even under the worst circumstances, its searching speed is eight times faster than the common algorithm, which indicates that the algorithm increases the efficiency of the computation.

4.2 The Algorithm of Adaptive Relative Gradient Iimage Fractal

Adopting fixed partitioning, the amount of value range blocks R_i can not be too few, otherwise, to the complex image the distortion ratio will become much higher after the compression revert, but value range block R_i corresponds with a compression commutation. The more compression commutation, the lower compression ratio [7].In many cases, especially when the background of the picture is monotonous and occupies large frame, we just need several ω_i or more to finish picture compression , gaining higher compression ratio. Therefore, we put forward self-adapting picture fractal compression algorithm based on relative gradient, in which we adopt quad-tree partitioning method, divide the original image and relative gradient picture into sub-blocks, with matching of sub-block processing in the relative gradient picture. The arithmetic is as follows:

(1) Divide image B into 4 non-overlapping sub-blocks $R_1\sim R_4$ serves as preliminary partition. Sub-block R_i is called value range block. Image B is the father-block of sub-block R_i (i = 1～ 4) . Father-block B is named definition domain block, whose size is four times as the sub-block.

(2) According to the formula of (4.2), work out relative gradient picture of the original image. Then according to (1) partition project, carry out partition in the relative gradient picture, gain 4 value range block $R_{g1}\sim R_{g4}$ and definition domain father-blocks B_g . By computing the relative gradient every sub-block R_i of original image always has a corresponding block in the R_g . Definition domain block B and B_g have the same corresponding relation.

(3) Compute mean $\bar{R}_{gi}(i=1,2,3,4)$ and mean-square deviation σ of sub-block $R_{gi}(i=1,2,3,4)$ and B_g
If the mean of sub-block $R_{gi}(i=1,2,3,4)$ is equal to the father-block and mean-square deviation is less than the given error threshold ε , then this sub-block resembles father-block B_g , and the corresponding sub-block R_i of the original image resembles the father-block. Record the coordinate of the sub-block R_i and the coordinate of top left of father-block B ,work out fractal code of the sub-block R_i and preserve in the document, then come to step (5);

(4) If the matching in the step (3) is unsuccessful, that is to say it dose not resemble its father-block, then sub-divide the sub-block of the corresponding original image into four sub-blocks $R_{i1},R_{i2},R_{i3},R_{i4}$ and it's father-block into four sub-blocks $B_{ij}(j=1,2,3,4)$ in the relative gradient image, fractal R_{gi} into 4 sub-blocks $R_{gi1},R_{gi2},R_{gi3},R_{gi4}$ and father-block B_{gi} into 4 sub-blocks $B_{gij}(j=1,2,3,4)$.

(5) For every sub-block R_{gij} , repeat the operations from the

third step to the forth one, till the coding of all sub-blocks can be successfully finished. If the sub-block's size can achieve the regulated minimal sub-block's size but cannot find its matching sub-block, see its minimum error sub-block as its matching block.

(6) All the sub-block's coding makes up image B fractal compress encoding and input and store it in the file.

The algorithm divides image B and its relative gradient images into range of value and definition domain. If we divide initial images and their relative gradient images into minimal regulated range of value $2^R\times2^R$, it needs $N-R$ times division at most. Taking the common complexity into account, we suppose p is probability to find the matching block from farther block in every division [7], and its range of value finishes encoding after finding its match-able sub-block. After the first division, it can get 4 ranges of value and the average number of match-able value is 2^2p, and unmatched sub-block number is $2^2(1-p)$, which needs 2^2 comparative times at most. After the second division of $2^2(1-p)$ ranges of value which haven't matching blocks, it will get $2^4(1-p)$ sub-blocks, $2^4(1-p)p$ matching blocks, and $2^4(1-p)^2$ unmatched blocks. The step needs $2^4(1-p)\times(2^{N-R-k})^2(k=N-R-1,...2,1)$ comparative times. In fact, as for the sub-blocks which can find their match in father-blocks, they can find their match only by comparing with parts of father-blocks, rather than with all the sub-blocks. So it needs $2^4(1-p)\times(2^{N-R-k})^2(k=N-R-1,...2,1)$ comparative times at most. Divide the unmatched ranges of value one by one till the $N-R$ times division and make the ranges of value achieve the minimal regulated size $2^R\times2^R$. Meanwhile, there are still $2^{2(N-R)}(1-p)^{N-R}$ sub-blocks which have no matching blocks, and they need further encoding, and the last division needs $2^{2(N-R-1)}(1-p)^{N-R-1}\times2^{2(N-R-1)}$ times to compare. So the average algorithmic times are:

$$T(N)=2^2+2^4(1-p)\times(2^l)^2+2^6(1-p)^2(2^{l-1})^2+\cdots+2^{2s}(1-p)^s\times2^{2s}$$

(Suppose l=N-R-K, s=N-R-1)

If the probability is $p\in(0,1)$, we suppose $p=0.5$, the self-similarity degree of the image is 50%, the arithmetic average calculating complexity is $T(N)=O(2^{2N})$.

Take the worse complexity algorithm condition into consideration. If we divide image B into $2^R\times2^R$ minimum size, in which no matching can be found between sub-blocks and subs. And in the end, we can get $2^{N-R}\times2^{N-R}$ ranges of value whose sizes are $2^R\times2^R$ and $2^{N-R-1}\times2^{N-R-1}$ subs whose sizes are $2^{R-1}\times2^{R-1}$, and the algorithm's calculating complexity is $O(2^{4N})$. So our algorithm has a higher efficiency.

The algorithm adopts adaptive quad-tree partition method which is more suitable for parallel processing. Suppose parallel environment is PC cluster, and there are $2^{2n}(n\geq1)$ processors. After adaptive quad-tree partitioning, allocate $2^{2(n-1)}$ processors to the initial images and relative gradient images R_1 and R_{g1}. The processor's rank is $1\sim2^{2(n-1)}$. Number $2^{2(n-1)}+1\sim2\times2^{2n}-1$ processor takes charge of the sub-block division of R_2 and R_{g2} and sub-blocks self-similarity search matching of R_{g2}. According to this till all sub-blocks can have the processors to do parallel processing and its average calculating times are

$$T_p(N)=1+(1-p)\times(2^{l+1})^2+(1-p)^2(2^{l+\frac{1}{2}})^2+\cdots+(1-p)^{N-R-1}\times2^{2(l-1)}$$

(Among $l=N-n-R-k$)

Suppose $n=\left\lfloor\frac{N+1}{4}\right\rfloor$, that is to say, the number of processors is $2^{\left\lfloor\frac{N+1}{2}\right\rfloor}$, the average calculating times in parallel processing are:

$$T_p(N)=1+(1-p)\times2^{l+2}+(1-p)^2\times2^{l+1}+\cdots+(1-p)^{N-R-1}\times2^{N-2R-1}$$

(Suppose $l=N-2R-2k$)

If probability $p=0.5$ does not change, then the calculating complexity in parallel processing is const. actually, this algorithm parallel processing is the most suitable for shared memory groups of processors, and its parallel acceleration rate can achieve linearity speedup rate.

5. CONCLUSIONS

By the fixed division algorithm, the number of the divided ranges of value is fixed. So is the number of IFS compressed transformations. So its fractal compression is unchanged. By adaptive relative gradient algorithm we need not to divide range of value because it is dynamic with the matching of range of values and subs. So the number of the divided ranges of value in this way is fewer than by the fixed partition. That is to say, the number of IFS compression transformations will be fewer than by the fixed partition. Thus, the algorithm has improved the compression rate. The average calculating complexity of the algorithm is $O(2^{2N})$, which improves the efficiency of computations.

REFERENCES

[1] A.E.Jacquin. "A Fractal Theory of Iterated Markov Operators with Applications to Digital Image Coding," PhD thesis, Georgia Institute of Technology. Aug，1989.

[2] A.E.Jacquin. "Fractal image coding based on a theory of iterated contractire image transfornzations," In *Proceedings of the SPIE*, Visual Cornrnnications and Image Processing.Oct. 1-4, 1991. Volume 1360: 227-239.

[3] H.J. Jiang.Study on Fractal Image Compression. the degree of master of science from Chongqing UniversitJiang of Science and Technology.2004

[4] Y.Fisher. Fractal image compression theory and application.Springer —Verlag, New York, 1995:1-77.

[5] Z.Sha.H.J.Ruan.Fractal and Fitting (M).HanZhou: ZheJiang University Press，2005

[6] E.W.Jacobs,Y.Fisher, R.D.Boss. Image compression: a study of the iterated transformmethod. Signal Processing 29, 1992: 251—263.

[7] L. G.Xiao, C. Zhong. An improved algorithm for image fractal compression. Journal of Guangxi University (N at Sci Ed) Vol. 27, No. 4 Dec, 2002.

Reconstruction of Video Electromagnetic Leakage from Computer

Bo Hu[1], Hongxin Zhang[2]
[1]Department of Physics and Electronics, Binzhou University, Binzhou, Shandong, 256603, China
[2]E&E, Beijing University of Posts and Telecommunications, 100876, China
Email: hubobz@163.com

ABSTRACT

TEMPEST technology has been concerned much more in recent years. Video Displayer Unit, which is one of the most import parts that may result in computer information leakage, offers the interface for man-machine conversation directly and its video electromagnetic radiation contains the displaying information. The interception and recovery system of video electromagnetic information leakage of the computer is designed. After the electromagnetic leakage is intercepted, the method to recover the word image is presented using the reconstruction techniques such as synchronization, related filtering, and phase lock on etc. This means that the original useful information can be reconstructed by the electromagnetic leakage under certain conditions, and this will menace information security. It is one of the most important factors that should be concerned about in information security and electronic antagonism fields.

Keywords: TEMPEST, Electromagnetic Radiation, Electromagnetic leakage, Filtering, Information Reconstruction

1. INTRODUCTION

Electronic equipments deal with data information through controlling the related changes of electric current (or voltage). But the time-variant electric current will bring electromagnetic radiation which contains abundant frequency spectrum and information that can be unscrambled [1].

The video information radiation from electromagnetic radiation coming with the working computer which can receive the transmitting signals can be monitored. The emission of electromagnetism has two main influences on the information technology equipment: one is the electromagnetic interference and information leakage caused by their own electromagnetic emission, the other is the disturbing and destroying from outside. The two kinds of influence have both the electromagnetic interference problems and the information leakage problems. TEMPEST, advanced during 1980s is a technology based on computer information leakage. It is a new technique developed from the field of electromagnetic annexation bringing great threat to the information safety. Acquiring the intelligence from TEMPEST leakage is one of the important methods used by ELINT to get information .In the Bay-War happened in 1991, America used the most advanced TEMPEST technique to intercept and capture the intelligence in politics, military affairs and economy of Iraq and Bay area. Although the TEMPEST technique is based on the principle of electromagnetic radiation, it pays more attention to abstract useful information handling and identifying to dealing with this useful information. The computer system is the most important component of various information technique devices, so it is the primary research objective among all the TEMPEST technique problems. The computer's make-up circuits are complex and contain many kinds of clock information, all of which has electromagnetic radiation to a certain degree. The sources of radiation can be divided into CPU, communication circuits, transform equipments, and output devices and so on. All of them will lead to information leaking phenomenon. These leakages contain synchronism signal, clock signal, and digit signal, being processing as well as information being displayed on the screen.

In 1995, a Dutch academician professor VanE.W introduced the experience result of information leakage caused by the electromagnetic radiation from displaying and analysis of the information linking theory. Henceforth, people have series of experiments to test data lines, hard disks, kinescope and the phenomenon coming along with CPU while it is transmitting and processing information. Moreover, they have proved that these electromagnetic radiations can lead to information leakage. Peter Smulders and other people have investigated the information leakage caused by electromagnetic radiation from RS-232 and web lines. They have investigated the harmful outcome of it as well. In recent years, Ross J.A and Markus.G have intercepted and reproduced the information, such as words, images, and analyzed the information linking effect resulted from the low frequency and the high frequency among video radioactive information by using Data Safe/ESL Model 400 TEMPEST in the laboratory [3]. Markus G.Kuhn has used photo electricity double tube to receive the electromagnetic radiation (the harm coming from the light leakage) from electric scanning video signals of color display, and then re-appeared the word video leakage via receiver. So he has proposed TEMPEST computer based on the Soft-TEMPEST theory. The computer laboratory of Cambridge University has publicized the control-system structure of TEMPEST mode Hollowman used to do experiments about information leakage from SS7, ATME, UNI/NNI, and RSVP. The key technique of electromagnetic information leakage is strictly kept secret abroad. Compared to foreign countries, we have a long distance to keep up with them [2].

Our country has begun the research on TEMPEST from mid-1980s. In the early 1990s, some key-pointing problems of TEMPEST technique were studied and many important achievements were gained on many subjects. Such as, computer information leakage and its principle of protection, the technique of receiving and restoring micro-mini computer radioactive information, safety evaluation, technical test of product, laboratory and scene testing, distinguishing red and black signals, the technique of electromagnetic leakage protection in micro-mini computer system, etc. In 1980s in the nationwide exhibition of computer application, the department of public security demonstrated that they re-appeared the showing content of micro-mini computer's screen on TV's screen by using TV receiving antenna to aim at the computer. Xi'an Electronic Technique University and other universities used black-and-white television to receive screen information.

Changchun Light Machine Office of Chinese Academy of Science carried out the interception and re-appearance about the video leaking information of the computer whose showing method is CGA. The receiving scope is usually limited to about 3 meters for re-appearing display text information by means of black-and-white television structure. But Van. Eck came up in his thesis that it can receive and deoxidize the video frequency information among 1000 meters. In 1990s, the English reported that they can receive and deoxidize information in 1600m. In comparison, Beijing Postal and label Services University carried out a receiving machine which realized the re-appearance of computer word leakage and got word re-seen picture located at far range. In 2004, the university completed a simulation platform which was used to reappear computer video electro magnetic linking information and realize the ration evaluation about the leaking threshold of information and the reinforcing function of computer [4].

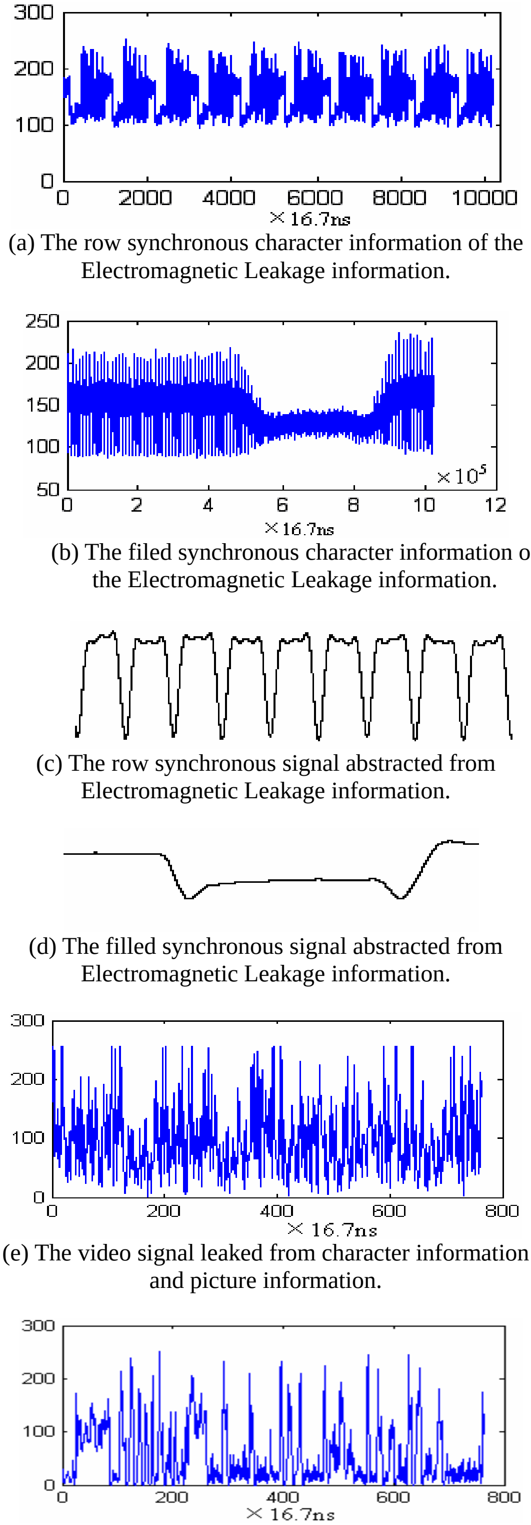
Through the video frequency channel we can realize the direct communication between man and computer [5]. Dealing with the real data use many technique like synchronism, related filtering, phase lock and so on to obtain the word and image reappearance from electromagnetic leakage [6].

From the research we can discover the electromagnetism radiation caused by digital pulse video frequency signal processing a wide frequency spectrum area. In addition, owning to the effect of multi-patches, effect will be neglected under strong noise when transmitting through the wireless channel. At the same time the signal has been deformed. Because the received video signal radiation has low S/N, wide spectrum and will be deformed, so capturing and dealing with the electromagnetism leakage information is different from receiving and doing with the classical communication signal either in theory or in technique. In this article, when the electromagnetic leakage from computer is captured, the technique of abstracting the filed synchronism signal and row synchronism signal, phase clock, related filtering, are used to re-appear the word image of common song-character displaying on the computer screen away from 10 meters. It's a breakthrough to the recorder of the distance researched about the leakage information.

2. THE DISTILLATION OF THE SYNCHRONOUS SIGNAL IN ELECTROMAGNETIC LEAKAGE

The metrical results of the Electromagnetic Leakage were given in the references. In this paper, we design the frame of interception and rebuilding of the Electromagnetic Leakage information. After the interception with the help of the wide-band antenna, the information was transferred into digital. The signal process was finished by software applications. The sample rate of the system is 60M [7].

In the near filed, to output through amplify, demodulation and detector after digital sampling, we could get the character signal of Electromagnetic Leakage when computer display character information near the receiving center frequency. Which Fig.1(a) 、 (b) is field synchronous information and row synchronous information of Electromagnetic Leakage information separately, Fig.1(c)、(d) is field synchronous signals and row synchronous signals which are abstracted by Electromagnetic Leakage information. Fig.1(e) is one row of the video signal of Electromagnetic Leakage information; Fig. 1(f) is the abstracting video signal after signal processing. Corresponding to Fig.1(e) which vertical is relative amplitude of the Electromagnetic Leakage information after signal processing, horizontal is the number of sampling point corresponding to the time 16.7 ns.



(a) The row synchronous character information of the Electromagnetic Leakage information.

(b) The filed synchronous character information of the Electromagnetic Leakage information.

(c) The row synchronous signal abstracted from Electromagnetic Leakage information.

(d) The filled synchronous signal abstracted from Electromagnetic Leakage information.

(e) The video signal leaked from character information and picture information.

(f) The video signal abstracted after signal processing.

Fig.1. Information distilled from Electromagnetic Leakage

3. THE TECHNOLOGY OF SIGNAL PROCESSING

3.1 The Principle of Phase Locks on

The recovery technique of getting field synchronous signals and row synchronous signals through signal processing of the collected Electromagnetic Leakage from Computer is one of the key techniques of Electromagnetic Leakage's Reconstruction. To make the field synchronous signals and row synchronous signals steady, we adopt the digital phase lock on technique--- making use of phase lock on loop circuit accurate to fix the output synchronous signals position. The basic structure of phase lock on is shown as Fig.2. It consists of 3 base component parts: comparer, LPF, and synchronous signals producer (synchronous circuit), the input signals are the field synchronous signals and row synchronous signals which are picked up by receiver. The comparer compares the input signals Si (t) with the output synchronous signals So (t) which pass the circuit to produce error voltage Se (t) corresponding to two signals. The function of LPF is to filter the high frequency component and the noise of the error voltage Se(t),the purpose is to ensure the function which is needed by the loop, and enhance the stability of the system. The synchronous circuit is controlled by the control voltage Sd(t) to make the synchronous signals approach to the input signals until the clearing of the error to lock on.

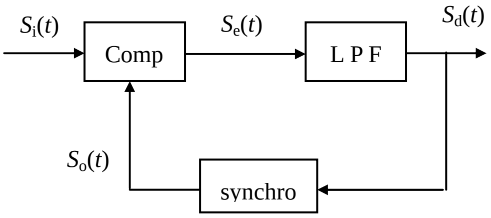


Fig.2. digital look on principle

3.2 Correlated Filter Wave

We could adopt the measure called shifting of function to enhance the SNR to reduce noise. Assume the included signal noise is

$$x(t) = s(t) + n(t) \tag{1}$$

Which s(t) is the signal whose circle is T, n(t) is the independent white noise whose average value is 0, square different is .The input signal SNR is :

$$SNR_i = W / \delta^2 \tag{2}$$

Assume the frame frequency of monitor is M, so the output of the system after following operation as following:

$$y(t) = \frac{1}{M}\sum_{k=1}^{M} x(t+kT) = s(t) + \sum_{k=1}^{M} n(t+kT) = s(t) + N(t) \tag{3}$$

In which N(t) is measure err. Its average value is obvious 0.its square err is

$$\sigma = D\left[N(t)\right] = \frac{1}{M}\sum_{k=1}^{M} E\left[n(t+T)^2\right] = \frac{\delta^2}{M} \tag{4}$$

Now, the processed SNR is

$$SNR_o = \frac{W}{\sigma} = M \cdot SNR_i \tag{5}$$

It is thus clear that SNR increase M times. It is easy to protect the pass function is

$$\left|H(\omega)\right| = \frac{1}{M}\left|\frac{1-e^{-jNM\omega\Delta t}}{1-e^{-jN\omega\Delta t}}\right| = \frac{1}{M}\left|\frac{\sin(\pi M\omega/\omega_0)}{\sin(\pi\omega/\omega_0)}\right| \tag{6}$$

It is thus clear that the gain of pass function of the filter is when w=kw0. The process equal to comb filter whose central frequency is w=kw0 called correlate filter.

3.3 Matching Filter

Matching filter could be considered as one correlate device which accurate input signals correlate function. The shape of wave becomes autocorrelation integral shape after passing match filter, and it is symmetrical about point t=t0.

t_0 is also the maximum point of the output signal. Consider the output signal is so(t) :

$$s_0(t) = \int_{-\infty}^{\infty} s(t-u)Ks(t_0-u)du \tag{7}$$

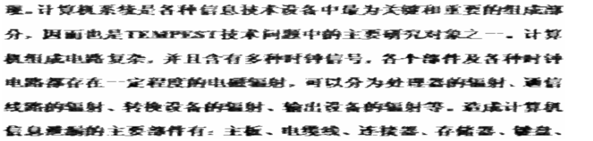
The impulse response of the match filter is the mirror image of the output signal wave. The Frequency spectrum related between input and output is:

$$S_o(\omega) = H(\omega)S(\omega) = S^*(\omega)e^{-j\omega t_0}S(\omega) = \left|S(\omega)\right|^2 e^{-j\omega t_0} \tag{8}$$

So, the frequency spectrum of output signal direct ratio to power spectrum input signal, only gap a delay factor that directs ratio to frequency. The stronger is the power spectrum of input signal, the stronger is the frequency spectrum of output signal. The frequency spectrum of output signal contains the conjugation of the power spectrum of input signal; the conjugation means that the phase can be dispel each other. Thereby, we can make phase modulation to the widen band signal, make all the phase to the identical phase, and algebra add the every frequency spectrum parts when the signal is finished, the output signal will have the most SNR. The characters of match filter make it have an advantage in the faint signal. Owing to this, we will use it in the filter disposal of the information divulge signal.

4. RESULTS AND DISCUSSIONS

The system we have designed and used received the electromagnetic information leakage when the monitor displays the word image, we have got the information reconstruction image through above disposal (Fig.3).It is the character of word document from up to down in it. The received place is 10 meters away from the computer. The result shows that: electromagnetic leakage information of computer is very weak, but we can intercept and reconstruction it.



TEMPEST是20世纪80年代提出的一种基于通信设备信息泄
漏的技术，它是在电磁兼容领域发展起来的一门全新的技术学
科。TEMPEST技术基于电磁辐射机理，更注意提取电磁辐射中
包含的有用信息，利用信息处理和信息识别的理论实现信息复
现。计算机系统是各种信息技术设备中最为关键和重要的组成部
分，因而也是TEMPEST技术问题中的主要研究对象之一。计算
机组成电路复杂，并且含有多种时钟信号，各个部件及各种时钟
电路都存在一定程度的电磁辐射，可以分为处理器的辐射、通信
线路的辐射、转换设备的辐射、输出设备的辐射等。造成计算机
信息泄漏的主要部件有：主板、电缆线、连接器、存储器、键盘、

Fig.3. Word re-appearance of electromagnetic information leakage

5. CONCLUSIONS

We have researched the problem of electromagnetic information leakage owing to the electromagnetic radiation by circuit inside computer monitor. It demonstrated that the electromagnetic information leakage of computer is objective through the experiment of the information received, extract and character image reconstruction from electromagnetic information leakage of computer. We can see from the reconstruction result: the resolution capability of details dispose must be improved. It is limited by the gather rate and signal technology we used. At present, the hotpot and emphasis of TEMPEST research are to come true the intercept ant reconstruction of Electromagnetic information leakage in far away. We have made an attempt works for the step further and have demonstrated the Electromagnetic information leakage related to the information secure and important latent factor to electromagnetic antagonism.

REFERENCES

[1] Berke Durak, "Hidden data transmission by controlling electromagnetic emanations of computers [EB/OL]," http://altern.org/berke/tempest/, 2000-08.

[2] Markus G. Kuhm, Ross J. Anderson, "Soft Tempest: Hidden Data Transmission Using Electromagnetic Emanations[EB/OL]," http://www.cl.cam.ac.uk

[3] Lu Ling, Nie Yan, Zhang Hongjin, "The electromagnetic leakage and protection for computer[A]," *Electromagnetic Compatibility Proceedings [C]*, 1997 International Symposium , 21-23 May 1997, 378 –382

[4] W. van Eck, "Electromagnetic Radiation form Video Display Units: An Eavesdropping Risk?," *Computers & Security* , 1985, vol.4, 269-286.

[5] P. Smulders, "The Threat of Information Theft by Reception of Electromagnetic Radiation form RS-232 Cables," *Computers & Security*, 1990, vol.9, 53-58.

[6] Raymod J. Lackey, Donald W. Upmal, "Speakeasy: The Military Software Radio," *IEEE Communications Magazine*, May, 1995, vol.33 (5): 56-61.

[7] Zhang Hongxin, Lu Yinghua, Qiu Yuchun, "The study on the electromagnetic leakage information arising from computer," *The Jjournal of China Universities of Posts and Telecommunications*, 2004, 11(2): 80-8.

Study on Interlace Video Coding Technique

Ruolin Ruan [1,2]
[1]**School of Information Engineering, Xianning College**
Xianning, Hubei 437100, China
[2]**National Engineering Research Center for Multimedia Software, Wuhan University**
Wuhan, Hubei 430072, China
Email: rlruan@163.com

ABSTRACT

Video coding technology mainly include two classes, that is, the progressive technology and the interlace technology. Interlace coding is a video coding technology that it divide a frame into two fields (top field and bottom field). The paper mainly analyses all kinds of interlace coding means, such as the fixed frame coding, fixed flied coding, picture-level adaptive frame/flied coding and MB-level adaptive frame/flied coding in the H.264/AVC, and so on. Through extensive simulation experiences to vary video sequences over the H.264/AVC, we find their advantages and disadvantages respectively, and analyses them by results of experience. At last, the paper concludes how these coding methods should be used according to the motion characteristic of vary video sequences.

Keywords: H.264/AVC, Interlace Sequence, Field Coding, Adaptive Frame/Field Coding

1. INTRODUCTION

Our definition of interlaced video comes from Annex W of H.263. An interlaced frame contains two fields, top and bottom, which are interleaved. The top field consists of the first (i.e., top), third, fifth, etc. lines of the complete picture. The bottom field consists of the second, fourth, sixth, etc. lines of the complete picture. A Top Field Picture consists of only the top field lines of a total frame. A Bottom Field Picture consists of only the bottom field lines of a picture. When sending interlaced field indications, an encoder shall use a picture size (custom picture size, if necessary) such that the picture dimensions correspond to those of a single field. In the work [1], there is very detailed description.

As in Annex W of H.263, the vertical sampling positions of the chrominance samples in interlaced field coding of a top field picture are specified as shifted up by 1/4 luminance sample height relative to the field-sampling grid in order for these samples to align vertically to the usual position relative to the full-picture sampling grid.

The vertical sampling positions of the chrominance samples in interlaced field coding of a bottom field picture are specified as shifted down by 1/4 luminance sample height relative to the field-sampling grid in order for these samples to align vertically to the usual position relative to the full-picture sampling grid. The horizontal sampling positions of the chrominance samples are specified as unaffected by the application of interlaced field coding. The vertical sampling positions are shown with their corresponding temporal sampling positions in Fig.1. And the detailed description in the W.2/H.263 below. [2]

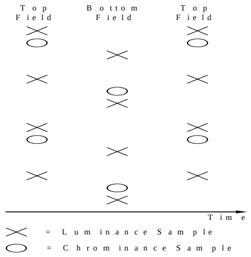


Fig.1. Vertical and Temporal Alignment of Chrominance Samples for Interlaced Field Coding

2. INTERLACED VIDEO CODING

The coding methods of interlaced video sequences can be the fixed frame coding, the fixed field coding, the adaptive frame/field coding at the picture level and the adaptive frame/field coding at the MB level. In the following, the paper will introduce these video coding methods, respectively.[2,3,4]

2.1 Fixed Frame Coding

In this approach, all the frames of a sequence are encoded in either frame mode. The code mode of frame will not be changed during entire sequence. And there are three ways of frame coding, that is,

(1) I I I I……
(2) I P P P……
(3) I B B P B B P ……

2.2 Fixed Field Coding

Note that picture types in bracket are for field coding. If field coded, a I frame is coded as one I field picture and one P field picture, a P frame as two P field pictures and a B frame as two B field pictures. It should be pointed that in general:

(1) A I frame can be coded as one I frame, or two I fields, or one I field and one P field,
(2) A P frame can be coded as one P frame, or two P fields, or one P field and one B field, and
(3) A B frame can be coded as one B frame, or two B fields.

However, in this contribution, the picture structures used in simulations follow the core experiments. In field coding, a frame is encoded as two field pictures. The rules for field coding are as follows.

(1) The sequence header indicates the field-coding mode.
(2) Two field of a frame are coded sequentially.
(3) The reference fields can be any coded I or P field.
(4) The code numbers assigned for field-based references stored in reference frame (field) buffer are slightly different from for frame-based references. Specifically, the code numbers of 0, 1, 2, 3, …, are grouped into pairs of (0,1), (2,3), (4,5), …. These code number pairs are assigned to the pairs of adjacent reference fields according to their distances to the current field (the field to be coded). Note that a pair of the adjacent fields

is not necessarily from the same frame. For each pair of reference fields, the field of the same parity as the current field is given the smaller number of the code number pair assigned for the pair of reference fields, as shown in Fig. 2.

(5) The skipped MB (copy mode) is reconstructed by copying the co-located MB in the most recently coded (past) I or P field of the same field parity.

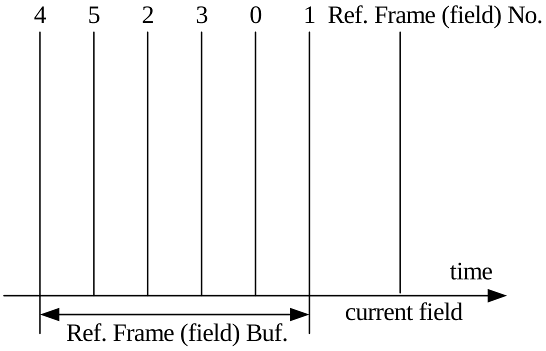


Fig.2. Reference Frame/Field Numbers

2.3 The Picture Level Adaptive Frame/Field Coding

For the picture level adaptive frame/field, an input frame of a sequence can be encoded as one frame or two fields. Specifically,

(1) An I frame can be coded as one I frame, or two I fields, or one I field and one P field, that is, I(or II)(or IP).
(2) P frame can be coded as one P frame, or two P fields, or one P field and one B field, that is, P(or PP)(or PB).
(3) A B frame can be coded as one B frame, or two B fields, that is, B(or BB).

In picture level adaptive frame/field coding, a frame can be encoded as one frame picture or two field pictures. The rules for picture level adaptive coding include:

(1) Picture header indicates whether the current frame is coded as one frame or two fields.
(2) For field coding, two fields of a frame are coded sequentially.
(3) For field coding, the reference fields can be any coded I field or P field.
(4) The code numbers are assigned to the reference fields stored in reference frame (field) buffer following the same rule as field coding (see Fig. 2).
(5) If in field coding, the skipped MB is reconstructed by copying the co-located MB in the most recently coded (past) I or P field of the same field parity.

The criterion for selecting either frame or field coding per frame is RD-based. The cost function is defined as

$$\cos t = Distortion + \lambda BitRate$$

where the Distortion is the distortion degree of the picture，and it is measured by sum of square difference of the original picture and the reconstructed picture.

At first, the coder code a picture by using the fixed frame coding method, and compute the cost of this coding method, then the coder again code the picture by using the fixed field coding, and also compute its cost, at last, comparing the two costs, if its cost is smaller than another, then the coder will choose accordant coding method with it. That it is said, the picture level adaptive frame/field coding method will choose either the fixed frame coding or the fixed field coding.

2.4 The MB Level Adaptive Frame/Field Coding

To improving efficient of coding further, H.264/AVC imports the concept of super MB. A super MB consists of 2 MBs of 16x16, as shown in Fig.3. A super MB of 32x16 can be coded as two frame MBs of 16x16, or one top-field MB of 16x16 and one bottom-field MB of 16x16.

The coding rules for super MB, such as intra prediction, reference frame/field and the assignment of code numbers for reference fields in the reference frame (field) buffer. For the skipped MB (copy mode), if in field, it is reconstructed by copying the co-located MB in the most recently coded (past) I or P field of the same field parity.

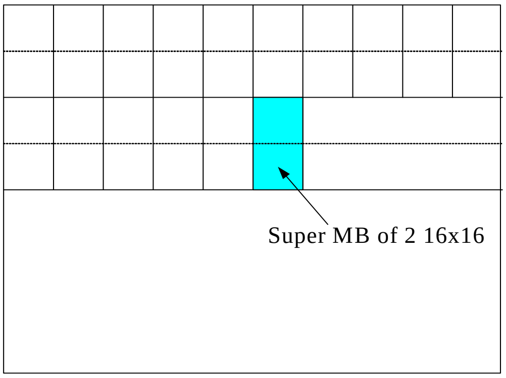


Fig.3. Sketch Map of the Concept of Super Macroblock

The AVS (Audio Video Standard) is the standard organization of audio and video coding in China, and she constitutes the correlation standard of audio and video coding. In the AVS, the vertical and correlation two macroblocks is called the macroblock pair, its size is 16× 32.

The macroblock pair has two classes, one is the non-sampling macroblock, and it is called the NS macroblock, the other is the vertical sampling macroblock, and it is called the VS macroblock, following as Fig.4. And in the Fig.5, the proposed [7][8] defined the adjacent macroblock, and between the macroblock have four relation of position, as following Fig, 5 (a), (b), (c), (d).

For MB level adaptation of frame/field coding, a MB can be coded in frame- or field-based. A frame/field flag may be required at MB level to indicate if the MB is coded in frame- or field- based, as shown in Fig.6 "0" indicates frame-based coding and "1" field-based coding.

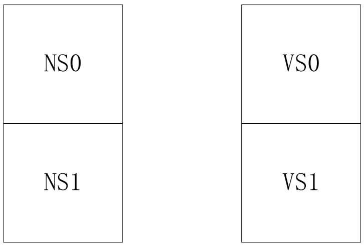
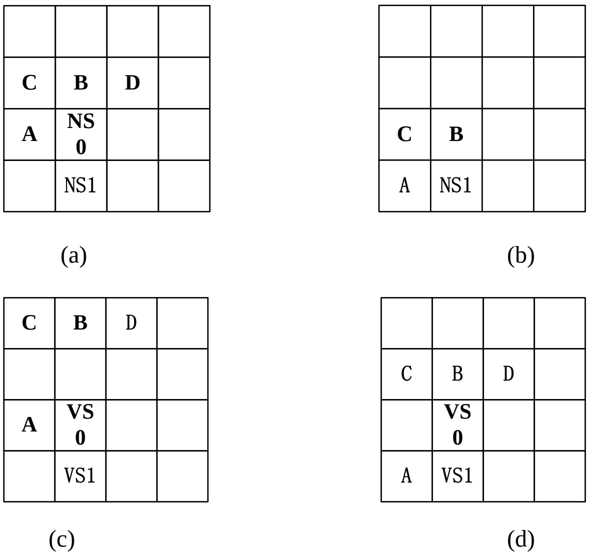


Fig.4. the Macroblock Pair



C	B	D	
A	NS 0		
	NS1		

(a)

C	B		
A	NS1		

(b)

C	B	D	
A	VS 0		
	VS1		

(c)

C	B	D	
	VS 0		
A	VS1		

(d)

Fig.5. the Define of the Adjacent Macroblock

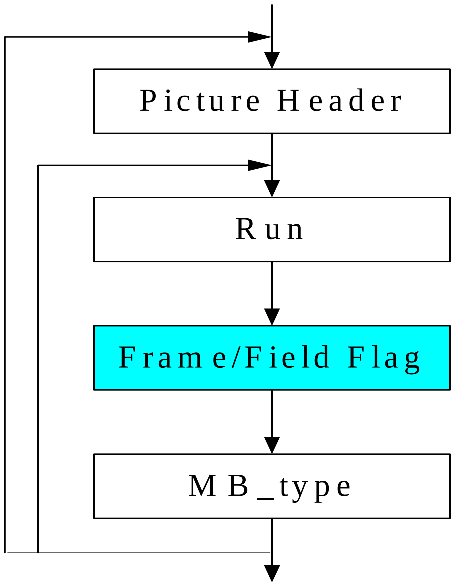


Fig.6. Sketch map of a frame/field flag

3. PERFORMANCE COMPARE OF VARIES VIDEO CODING MEHTODS IN H.264/AVC

In the following, the paper will compare the effect of the four coding method to varies video sequence through the simulation experiments. Simulations were carried out for two video sequences, as shown in Table 1. Mobile is a common test sequence. Tempete_Football is a hybrid sequence of Tempete of 352x480 on the right and Football of 352x480 on the left. Football contains high motions while Tempete is of slow motions. For comparison purpose, four coding methods, frame, field, picture-level adaptive, and MB adaptive coding, were performed for each of two sequences separately. A set of four quantization parameters (QP) was tested per coding method per sequence, where QP for I and P are 16, 20, 24, and 28. Other coding parameters are listed in Table 2. Picture structure used in tests is I and P only, that is, IPPPPP… [5, 6]

Table 1. Test Sequences

Sequence	Format 4:2:0	Length	Frame/Second
Mobile	704x480	150	30
Tempete_Football	704x480	150	30

Table 2. Test Conditions

Entropy Code.	MC	Hadamard	Ref. Frames	Search Res.	RD Opt.	Search Region
UVLC	1/4 pel	Yes	2	2	Yes	16

Fig. 7 and 8 show the PSNR with respect to bit rate for the two test sequences. The curves with diamond marks are for frame coding, the curves with square marks for field coding, the curves with triangle marks for picture-level adaptive coding and the curves with x marks for MB-level adaptive coding. As seen, picture-level adaptive coding performs better than both frame and field coding. MB-level adaptive coding provides an additional gain over picture-level adaptive coding. For sequence Mobile, the additional gain is around 0.25 dB and for the hybrid sequence Tempete_Football., the gain is up to 0.8 dB.

4. CONCLUSIONS

From the above analysis and simulation experiment, we can know that there are some picture that have very strong motion portions, that is, its content is vary quickly, and there will be a big vary in a very short time. So, if we use the fixed frame coding method then the efficient of coding will be very bad. Because the sequence is the interlace sequence, the scanning interval of the adjacent two rows (one row is in top field and another row is in bottom field) is very big. As change of picture is very acute, and the correlation of the adjacent row is very weak. But, if using the fixed field coding, as scanning interval of adjacent row is very short, even if the picture have very strongly move, and the temporal correlation of the adjacent row is very strong in the picture. And in this case, the fixed field coding method can remove the temporal redundant of picture. It is easy to be understood for the moveless picture, because the content of picture has little variety. The scanning interval is not important, and it is not affect to the correlation of the picture. And the adjacent row of a frame picture is real the adjacent row, its spatial correlation is better than the adjacent row of a field picture. And in the case, if we use the fixed frame coding method then the spatial redundant will be removed.

In a frame, the spatial correlation of the adjacent row is strong, and the temporal correlation is weak, so the fixed frame coding can be used to coding the picture with weak motility, and the picture with strong motility will be coded by using the fixed field coding, then their efficient will be improved. Furthermore, if there are varies motion speed in varies region of picture, then these region can be divided into the super MB, and they will be coded using the MB level adaptive frame/field coding method.

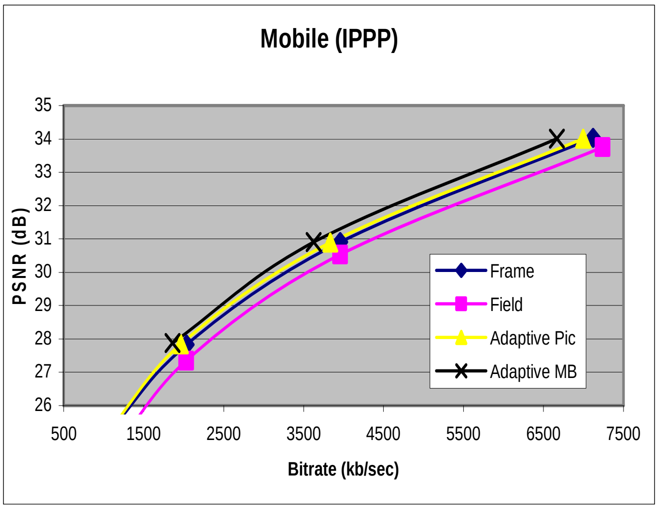


Fig.7. PSNR vs Bit Rate Curve for Sequence Mobile.

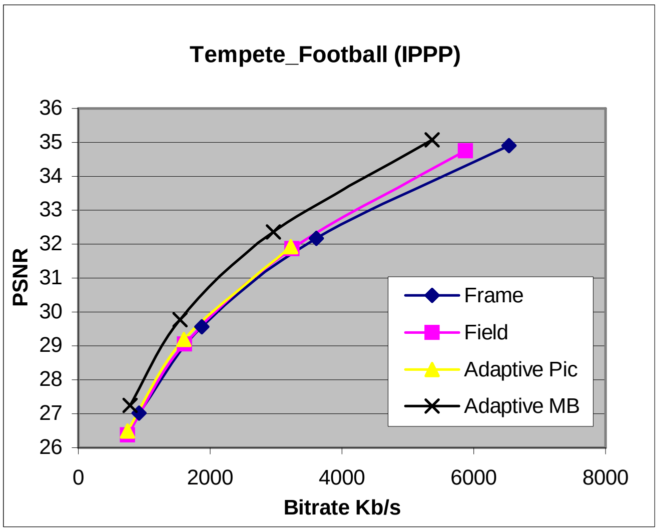


Fig.8. PSNR vs Bit Rate Curve for Sequence Tempete_Footbal.

REFERENCES

[1] ITU-T Recommendation H.263 (Video Coding for Low Bit Rate Communication), Annex W, specifying optional Additional Supplemental Enhancement Information, November 2000.

[2] Peter Borgwardt, “Handling Interlaced Video in H.26L,” ITU-T Q.15/16, Document VCEG-N57r2, September 2001.

[3] Michael Gallant and Guy Cote, “High Rate, High Resolution Video Using H26L”, ITU-T Q.15/16, Document VCEG-N84, September 2001.

[4] L. Wang, K. Panusopone, R. Gnadhi, Y. Yu, and A. Luthra, “Adaptive frame/field coding for JVT video coding”, JVT-B-071, Geneva, Jan. 2002.

[5] L. Wang, R. Gandhi, K. Panusopone, Y. Yu, and A. Luthra, “MB-level adaptive frame/field coding for JVT”, JVT-B-106, Geneva, Jan. 2002.

[6] L. Wang, K. Panusopone, R. Gandhi, Y. Yu, and A. Luthra, “Interlace coding tools for H.26L video coding”, VCEG-O37, Pattaya, Dec. 2001.

[7] He Yun, Chen Jianwen, “The Macroblock Level Frame/Field Adaptive Predict Coding Mean”, AVS-M1264, Beijing, June. 2004.

[8] AVS Group, “AVS-X CD 2.0: Information Technology – Advanced Coding of Audio and Video – Part 2: Video”, June, 2006.

Ruolin Ruan is currently pursuing his Ph.D at Wuhan Universirty. He is a lecturer of Xianning College, China. He received his M.S. degree in computer science and technology from Wuhan University of Technology in 2005. His current research interests include the modern long-distance education, multimedia technology, and video coding and video communication.

Digital Image Segmentation Based On Entropy

Xiaojun Tong[1], Qiuming Huang[2] ,Shan Zeng[3], Wenke Wang[4]
[1,2,3]Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan 430074, China
[1]Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[4]Department of Basic Courses, Command & Communications Academy of PLA, Wuhan 430010, China
Wuhan, Hubei, China
Email: tongxiaojun1998@yahoo.com.cn

ABSTRACT

We introduce entropy in fuzzy sets theory into digital image segmentation. According to the fuzzy character plane, firstly we get the segmentation threshold when the fuzzy bandwidth is given based on the maximal entropy principle. Then we develop the traditional binary segmentation to multi-value fuzzy segmentation, make modification of the threshold based on the multi-value fuzzy segmentation. When the difference between the object and the background is great, take the image ‘Cameraman’ for example, this method can effectively extract the object; When the difference between the object and the background is not great, take the image ‘Lena’ for example, this method can show plenty of edge information.

Keywords: Digital Image, Entropy, Fuzzy Bandwidth, Threshold Segmentation.

1. INTRODUCTION

Digital image processing[1] is a new crossed subject in the information technology which researches into all kinds of image theory, technology and application systematically. Its research object is the digital images produced by computer and other electronic equipments and the processing results is mainly used to show people, that is to say, the destination is human. So we should consider the image itself and the human vision in the course of image processing and recognition. The course of imaging is a multi-to-one mapping. So the image itself has much uncertainty or fuzziness. But the gray scale from black to white is fuzzy to human vision. People found that the fuzzy theory[2] has good description of such uncertainty. So we can introduce the fuzzy sets theory to be the model and method to effectively describe the image and human vision. Also we can analyse human behaviors such as estimation, awareness and distinguishing and get good effect[3]. In the field of signal processing, as a result of referring to a great deal of fuzzy factors, the fuzzy system theory has entered almost every field as a method of describing and dealing with the phenomena and thing which are of uncertainty (fuzziness). It has been used in the fields of automation,image processing , pattern recognition and machine vision etc[4].

In the course of studying the image, we are interested in only some parts of the image. We call these fields object or foreground (the rest is called background) which usually correspond to the fields with special quality. In order to recognize and analyse the object, we need to abstract these fields. This course which divides the image into object and background is the image segmentation course[4-5]. Traditionally we take the method of binary threshold segmentation. But the segmentation is not very perfect when the gray degree difference is not obvious or there is big superposition between the the gray degree scale. As the threshold segmentation method[6-7] is simple and can effectively finish image segmentation when the difference between gray scales of different objects is big, this method is of strong life. The disadvantage is that the correct result is hard to get and the segmentation result is not very good[7-10] when there is not obvious gray degree difference or there is great superposition between the gray degree scale. There are a lot of improving methods such as the maximal entropy criterion method and gray degree stretch method etc.[5] .We will study further the multi-value fuzzy segmentation and determine the threshold based on the maximal entropy principle. Then we make correction of the threshold by the fuzzy bandwidth and the determinate threshold.

2. THE FUZZY CHARACTER PLANE AND ENTROPY OF IMAGE

According to the fuzzy subset theory, suppose A is $I \times J$ dimension image with L grayscale, then the image can be seen a $I \times J$ rank fuzzy matrix . Image A is shown as :

$$\underset{\sim}{A} = \begin{bmatrix} \mu_{\underset{\sim}{A}}(x_{11}), \mu_{\underset{\sim}{A}}(x_{12}), \cdots\cdots, \mu_{\underset{\sim}{A}}(x_{1J}) \\ \mu_{\underset{\sim}{A}}(x_{21}), \mu_{\underset{\sim}{A}}(x_{22}), \cdots\cdots, \mu_{\underset{\sim}{A}}(x_{2J}) \\ \cdots\cdots\cdots\cdots, \cdots\cdots\cdots\cdots, \cdots\cdots\cdots\cdots \\ \mu_{\underset{\sim}{A}}(x_{I1}), \mu_{\underset{\sim}{A}}(x_{I2}), \cdots\cdots, \mu_{\underset{\sim}{A}}(x_{IJ}) \end{bmatrix} \tag{1}$$

$\mu_A(x_{ij})$ is the membership of pixel (i,j), that is to say, the extent to which pixel (i,j) belongs to the background fuzzy set A, and $0 \le \mu_A(x_{ij}) \le 1$.

If we set $x_{\max}$ as the biggest gray degree,then membership function $\mu_{\underset{\sim}{A}}(x_{ij})$ can be got by (x_{ij}) according to the standard fuzzy function S. The S function is as follows[4]:

$$S(x_{ij}, a, b, c) = \begin{cases} 0, x_{ij} < a \\ 2[(x_{ij} - a)/(c - a)]^2, a \le x_{ij} \le b \\ 1 - 2[(x_{ij} - c)/(c - a)]^2, b \le x_{ij} \le c \\ 1, x_{ij} > c \end{cases} \tag{2}$$

Here, $b = \dfrac{a + c}{2}$

$$\mu_{\underset{\sim}{A}}(x_{ij}) = S(x_{xj}, a, b, c) \tag{3}$$

In the function $S(x_{xj}, a, b, c)$, b is the threshold,when x_{ij} =b

$$S(x_{ij}, a, b, c) = S(b, a, b, c) = 0.5 \tag{4}$$

In the image segmentation, b is the segmentation threshold or edge threshold.

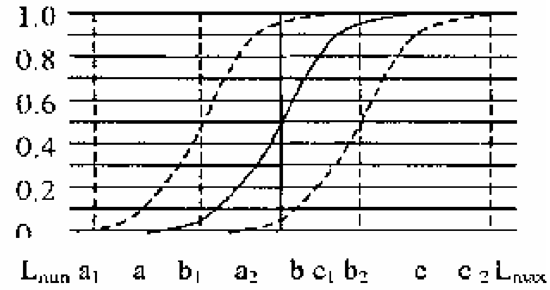


Fig.1. S Function

In Fig.1. the real line is S function curve, the two broken lines are left and right limitation . [a,c] is the fuzzy field, [$L_{\min}$,a] and [c, $L_{\max}$] are nonfuzzy field. Δb =c-a is the bandwidth, $\Delta b \in [\Delta b_{\min,} \Delta b_{\max}], [\Delta b_{\min,} \Delta b_{\max}]$ is the dynamic field of fuzzy bandwidth Δb . $a \in [L_{\min}, L_{\max} - \Delta b], c \in [L_{\min} + \Delta b, L_{\max}]$

The image is combained by pixels, visually speaking,, it is an arrangement of some pixels. So we should adopt the entropy with additivity. The entropy established basing on Shannon entropy in this paper has entropy in this paper has such property [2, 7].

If $\underset{\sim}{A}$ is a fuzzy set, in the universe of discourse U, x∈U. So when U is a discrete universe, the calculating formula[2] is

$$e\left(\underset{\sim}{A}\right) = \frac{1}{n \ln 2} \sum_{i=1}^{n} \left[-\mu_{\underset{\sim}{A}}(x_i) \ln \mu_{\underset{\sim}{A}}(x_i) - \left(\mu_{\underset{\sim}{A}^c}(x_i) \right) \ln \left(\mu_{\underset{\sim}{A}^c}(x_i) \right) \right] \quad (5)$$

we call $e\left(\underset{\sim}{A}\right)$ the fuzzy entropy of $\underset{\sim}{A}$.

3. THE THRESHOLD SEGMENTATION METHOD AND ITS IMPROVEMENT

The image segmentation is to divide the image into a few areas which don't intersect according to gray degree、colour、dimensional texture、geometrical shape, etc. These characters show coherence or similarity in the same area but obvious difference in different areas. Simply speaking, we separate the object from the background in an image to make further processing. Image segmentation is one of the most basic and important fields in image processing and computer vision. It is the basic precondition of visual analysis and pattern recognition of the image[4-5].

The threshold segmentation is the most common and simplest method. It is especially fit for the image whose object and background are in different gray degree scales. It can not only compress the data hugely but also simplify the process of analysis and processing. So it is the necessary pre-processing before image analysis、character abstraction and pattern recognition. The purpose of the threshold processing is to get a division of the pixel set according to the gray degree scale. The gained subset becomes an area which corresponds with the real scenery. The character is the same in the same area and between the neighboring areas. Such division can be got by choosing one or more threshold from the gray degree scale[5-7].

The theory of threshold segmentation is as follows:
Suppose the original gray image is f(x,y), we find a gray value t as the threshold by a certain regulation and divide the image into two parts, the binary image g(x,y) after segmentation is

$$g(x,y) = \begin{cases} 0, f(x,y) < t \\ 1, f(x,y) \geq t \end{cases} \quad (6)$$

3.1 The Maximal Entropy and The Binary Segmentation Method

In the membership function abstracted from the S function, when x_{ij} takes the value of threshold b,

$$\mu_A(x_{ij}) = \mu_A(b) = S(b,a,b,c) = 0.5 \quad (7)$$

S function is the continuous and derivable function，so when x_{ij} is closer to threshold b, $\mu_A(x_{ij})$ is closer to 0.5, the fuzzy matrix A is more fuzzy. We can confirm b by the maximal entropy principle:

$$b = \arg\{\max[e(b, \Delta b)]\} \quad (8)$$

so $c = b + \frac{\Delta b}{2}, a = b - \frac{\Delta b}{2}$, S function is the function of $b, \Delta b$,and $\mu_A(x_{ij})$ and $e\left(\underset{\sim}{A}\right)$ is the function of $b, \Delta b$.

Suppose

$$e\left(\underset{\sim}{A}\right) = e(b, \Delta b) \quad (9)$$

then $b = \arg\{\max[e(b, \Delta b)]\}$, arg represents inverse function, that is to say, to get the coordinate b of the maximal extremum of function $e(b, \Delta b)$.

We choose the classical image Lena and Cameraman and get Δb , b and e with the maximal entropy principle as follows:

Fig.2. Lena

Fig.3. Cameraman

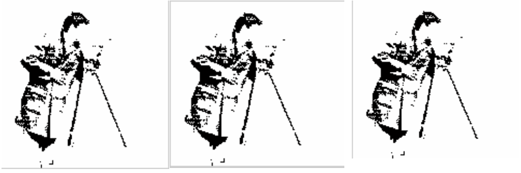
Table 1 Lena

Δb	b	e
10	115.54	0.0364
20	124.42	0.0674
30	113.77	0.0937
120	120.76	0.3066
140	122.06	0.3555
160	125.47	0.4029
240	120.43	0.5795

Table 2 Cameraman

Δb	b	e
10	12.47	0.1270
20	12.69	0.1815
25	12.82	0.1929
120	154.38	0.4955
140	152.94	0.5342
160	152.02	0.5647
240	135.00	0.6295

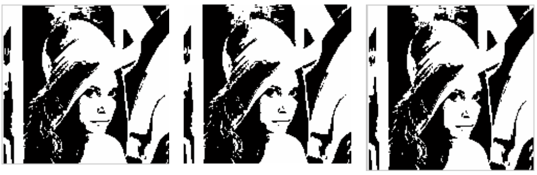
The binary segmentation results are as follows:



Δb =10 Δb =20 Δb =25

Fig.4. The Binary Segmentation (Cameraman)

We can see from the figure above that when $\Delta b < 30$,the object figure can be abstracted effectively and separated from the background well. When $\Delta b =$25, the segmentation result is good correspondingly.



Δb =10 Δb =20 Δb =30

Fig.5. The Binary Segmentation (Lena)

We can see from the figure above that when $\Delta b \le 30$, the figure after segmentation has abundant details and reserves good edge, loses little information.When $\Delta b =$ 30, the segmentation result is good correspondingly.

3.2 Multi-Value Segmentation

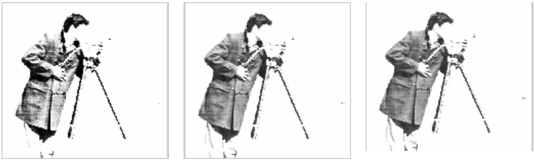
We can get corresponding a and c according to Δb and b. We carry out the multi-value segmentation according to the standard fuzzy S function. The method is as follows:

$$f(x,y)=\begin{cases}0, x_{ij}<a \\ S(x_{ij}), a\le x_{ij}\le c \\ 1, x_{ij}>c\end{cases} \qquad (10)$$

Because the binary segmentation strengthen the gray degree of the object (when gray degree$\ge b$, we order it to be 255). The fuzzy multi-value sementation consider the bandwidth besides we reserve the object(gray degree$\ge b$) strengthening of the binary segmentation. If there is certain difference between the object and the background ,under proper bandwidth,the fuzzy segmentation separates the object well and reserves part of the background. We combine the two methods and create the improved fuzzy multi-value segmentation, that is to say, improve the b in the binary segmentation. The process is as follows:

Order $b^{'} = b + \lambda\Delta b$ (λ is $0.5 \sim 1$),and make binary segmentation of the original image.

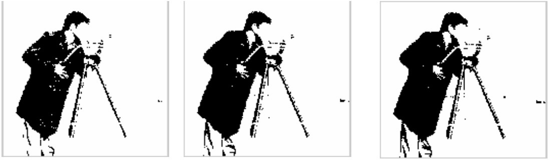
(1) The difference between object and background is big---- Cameraman



Δb =10 Δb =20 Δb =25

Fig.6. Multi-value Segmentation (Cameraman)

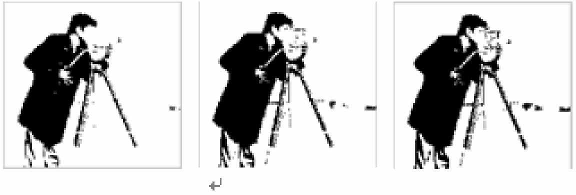
When $\lambda = \frac{1}{2}$, according to $f(x,y)=\begin{cases}0, x_{ij}<b^{'} \\ 1, x_{ij}\ge b^{'}\end{cases}$,we make new binary segmentation and the results are as follows:



Δb =10 Δb =20 Δb =25

Fig.7. Modification Images , $\lambda = 0.5$ (Cameraman)

When λ =1, the result of the binary segmentation is as follows:



Δb =10 Δb =20 Δb =25

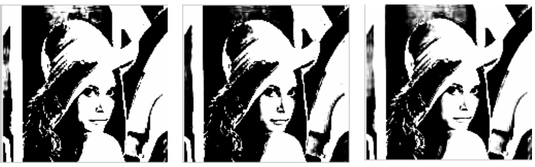
Fig.8. Modification Images, λ =1 (Cameraman)

The third image in Figure 3.7 has part of reserved background,but the image has better strengthening effect of the object.We can wipe off the reserved background by the smoothness filtering noise wiping off.

When $\Delta b < 30$, for the images whose object and background have big difference, we can effectively abstract the object and separate the object from the background very well by developing the traditional binary segmentation to multi-value fuzzy segmentation.

We can make modification and complementarity of the fuzzy multi-value segmentation by binary segmentation with $b^{'}$.

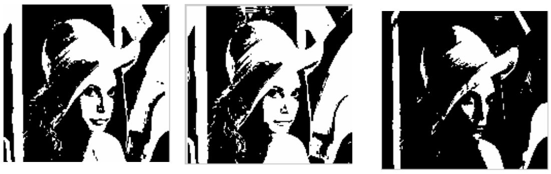
(2) The Difference between Object and Background Is Not Big---- Lena



Δb =10 Δb =20 Δb =30

Fig.9. Multi-value Segmentation (Lena)

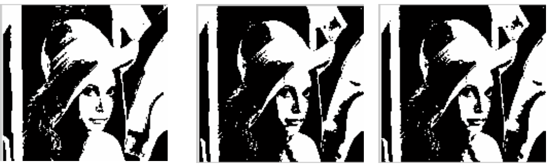
When $\lambda = \frac{1}{2}$, the result of the binary segmentation is as follows:



Δb =10 Δb =20 Δb =30

Fig.10. Modification Images, $\lambda = \frac{1}{2}$ (Lena)

When λ =1, the result of the binary segmentation is as follows:



Δb =10 Δb =20 Δb =30

Fig.11. Modification Images, λ =1 (Lena)

When $\Delta b < 30$, for the images whose object and background have small difference, the object after segmentation shows abundant edge information and reserves good details and clear edge by multi-value fuzzy segmentation or modification threshold segmentation. When $\Delta b = 30$,the segmentation effect is good correspondingly.

4. CONCLUSIONS

We introduce Shannon entropy in fuzzy sets theory into digital image segmentation. According to the fuzzy character plane, we get the segmentation threshold when the fuzzy bandwidth is given based on the maximal entropy principle. We develop the traditional binary segmentation to multi-value fuzzy segmentation. The segmentation results are both good when the difference between the object and background is big or not very big.

5. ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation under Grant 79970025, 60403002 and 30370356 of China, and the plan of Science and Technological Innovation Team of the Outstanding Young and Middle-aged Scholars of Hubei Provincial Department of Education, and Hubei provincial Natural Science Foundation under Grant 2004ABA031 and 2005ABA233, and National Postdoctoral Science Foundation of china (Grant 2004036016), and Foundation of Hubei Provincial Department of Education Grant 2003X130 and Scientific Research Team of Wuhan Polytechnic University Grant 03-T-06.

REFERENCES

[1] Rafael C.Gonzalez, Richard E.Woods. *Digital Image Processing*, Bejing:Publishing House of Electronics Industry,2005

[2] Wang Peizhuang, *The theory and application of fuzzy sets* , Shanghai Scientifical and Technological Press, 1983

[3] Xu Yan,Wang Chao, “The entropy of digital image basing on fuzzy math,” *The Computer Aided Design and Graphics*, 2000,12(5):747-749

[4] Sun Guanglin, “The research on the image threshold segmentationbasedfuzzythreshold,”[Dissertation],Hefei, Hefei University of Technology,2005.6

[5] Gao Xiujuan, “The theory, method and application of image segmentation,” [Dissertation],Changchun, Jilin University,2006.4

[6] Pei Jihong, Xie Weixin, “Self-adaption and multi-threshold image segmentation using FCM clustering with diagram as fuzzy restriction,” Electronic Transaction, 1999, Vol.27, No.10:38-42.

[7] Li Yan, Fan Xiaoping, Li Gang, “a new image threshold segmentation,” *The Computer Simulation*, 2006, 6.

[8] Pham T.D, “Image Segmentation Using Probabilistic Fuzzy C-Means Clustering. *Proceedings of International Conference on Image Processing*,” 2001, 21(3):722-725.

[9] S.K.Pal, R.A.King “Image Enhancement using Fuzzy Set,” *IEEE Electronics Letters*.1980.16:376-378.

[10] Sun Zaoling “Matlab 6.x image processing,” Beijing, Tsinghua University Press, 2002.

Performance Analysis of Embedded Runge-Kutta Methods in Cloth Simulation*

Xinrong Hu , Lan Wei
Dept. of Computer Science & Technology, WuHan University of Science & Engineering
Wuhan, Hubei, 430074, China
Email: hxr@wuse.edu.cn

ABSTRACT

A satisfied cloth simulation should be general, accurate, efficient and stable. Explicit, implicit and semi-implicit integration methods have contributed to large performance enhancements in the field of cloth simulation. However, these methods are not ideal in improving the cloth simulation result because of their efficiency, accuracy or stability. In this paper, we consider embedded Runge-Kutta methods used to solve the ordinary differential equations and study their efficient implementation on cloth simulation. With the classical spring-mass model, we investigate and solve the ordinary differential equation for cloth simulation. First, we determine the extent to which the overall quality must be compromised in order for the stable conditions to be satisfied. Two embedded Runge-Kutta methods are presented and their simulation properties are compared with general fourth order Runge-Kutta method. The result shows that the embedded Runge-Kutta methods have some advantages for cloth simulation: adaptive time step, controllable errors, good stability and satisfied precision. Experiments demonstrate that this approach results in simulation efficiency improvements.

Keywords: Cloth Simulation, Embedded Runge-Kutta, Ordinary Differential Equation, Numerical Integration, Spring-Mass Model, Time Step

1. INTRODUCTION

Garment designers, textile engineers and computer graphics researchers are all interested in predicting the motion and static drape of cloth. Cloth simulation can be used to solve this complex problem. It is an important integral component of virtual character animation and can be used in future film, games and virtual reality applications. A general, accurate, efficient and stable technique is a challenge for solving 3D cloth simulation formulation. In order to improve the reality of 3D cloth simulation, maintaining a reasonable computation time and error, a deeper study of the cloth model and the identification of cloth behavior at different levels are necessary.

The study is not intended to integrate another more precise physical model of cloth behavior, but rather focus on the real-time constraints for the simulation and the visual cloth motion features to which an observer is sensitive. Most of the existing approaches use a general-purpose simulation method using discrete simulation model of the cloth [3-4]. Unfortunately, simulations that simply calculate all potentially colliding vertices may generate a realistic result, but do not provide a guaranteed frame time and error. While an effective method should be implemented which avoids heavy computation of the spring-mass model wherever possible.

This problem has already received more and more attentions from mechanical and textile engineers, as well as computer scientists. In fact, spring-mass model [6] is the most frequently considered discrete cloth representations, and tends to prefer discrete ordinary differential equation (ODE) model, typically formulated through Lagrangian or Newtonian dynamics, formally simpler and computationally less expensive.

In this paper, we aim to develop an implicit integration method--embedded Runge-Kutta method solving the cloth motion equation during the real-time simulation. While simulating cloth, the method can reduce the computational costs, adjust the size of time step automatically and control errors availably, without losing accuracy or generality.

2. RELATED WORKS

2.1 Spring-Mass Model

The simplest and intuitive way of designing a cloth simulation system is to consider the cloth as a discrete masses structure that interact with each other through springs [10]. This discrete cloth structure is called spring-mass model. The technique was first introduced in 1988 by Hamann and Parent, and was developed further in 1995 by Provot [7,11]. Since then, spring-mass model is the most common modeling tool for cloth simulation. The spring-mass model considers a triangular mesh where the vertices are masses and the edges are springs with constant rigidity and optional viscosity, as shown in fig 1 and fig 2.

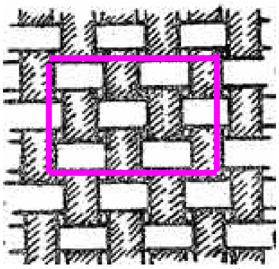


Fig.1. Close View of Cloth

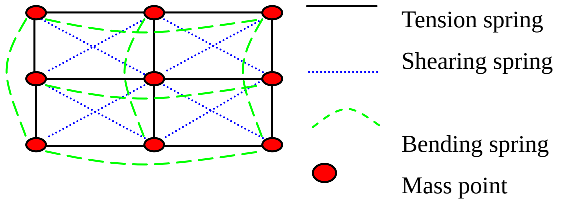


Fig.2. Corresponding Simulation Model

This representation simulates cloth by modeling the low level structure of the fabrics and the anisotropy behavior due to

* The Hubei Educational Council, China (2004D005) and Hubei Digital Textile Equipment Key laboratory (DTL200702) provide the funding for this work.

different warp and weft properties. This model yields visually plausible deformable cloth surface, which can be solved through adding cross-springs shown in figure 1. Bending springs are connected between two mass points across each edge. Tension springs are connected between two nearest neighbor mass points along horizontal and vertical direction. Shearing springs are connected between two diagonal mass points. The three types of springs have different physical and mathematical formulations to describe the forces applied on the cloth. The combination of the mass points and spring forces is integrated with respect to time to provide a new acceleration for each point. Through the spring-mass model, the motion and static drape of cloth can be computed with several numeric integration methods.

2.2 Numeric Integration Methods for Cloth Simulation

Numerical integration is the essence of spring-mass based cloth simulation and is the most important factor in the design of cloth simulation systems [11]. With numerical integration methods, an ordinary differential equation is solved to compute the time-varying state of the masses.

Semi-implicit integration scheme was used to solve the motion differential equation of the particle system [8], which including explicit space domain and implicit time domain. It offers first-order accuracy and a better stability than a pure explicit integration scheme. It does not require the resolution of a large sparse linear system for each of iterations. Several semi-implicit integration methods have been discussed in terms of stability and accuracy [5]. However, large time step may be used, most of the time without any instability issue. When time steps are infinite, it is equivalent to iterations of the Newton resolution method.

Since the pioneering work of Baraff and Wikin [15], implicit integration scheme is currently widely used in applications including cloth simulation, from real time animation in Virtual Reality applications to accurate garment simulation for design and prototyping applications. It has good simulation stability. The computation may be performed more quickly than semi-implicit scheme because of adopting large simulation time steps. Many implicit integration methods are now available in cloth simulation. Hairer [14] described a detail review of them.

Simulation errors, which usually break the stability of explicit integration methods, only appear in the implicit scheme as a form of “numerical damping” that does not convergence to equilibrium completely. This may alter the motion of cloth, which results in decreasing the reality of simulation result and increasing the size of time step. This may be a problem when expecting an accurate cloth motion and simulation.

In order to alleviate the disadvantage of implicit integration, a various different methods have proposed for cloth simulation, such as Euler method, Midpoint method, higher-order Runge-Kutta method, and so on. However, the accuracy or simulation error of these methods is difficult to control.

Explicit integration schemes or implicit ones all have a badly limitation that a fixed time steps must be given in advance. If the time step is too large, the cloth simulation will exceed the stability area and cannot achieve equilibrium. Else if the time step is too small, the cloth simulation will increase computational time. An improved method to provide more accuracy, efficiency and stability is expected.

The scheme proposed in this paper is an adaptive Runge-Kutta method. Through embedded optimized parameters, the method achieves an adaptive time steps, higher efficiency, controllable errors and accuracy. We compare the performance of the embedded Runge-Kutta with the normal Runge-Kutta when they are tested on cloth simulation.

3. EMBEDDED RUNGE_KUTTA METHOD

3.1 Motion Differential Equation of Cloth Simulation

As mentioned above, the surface of motion cloth is represented with spring-mass model. That is, the cloth structure is discrete into a mesh of point masses. The dynamics of the system is then simulated in time according to Newton’s second law:

$$F_{\text{int}\,ernal} + F_{external} = m\frac{d^2x}{dt^2} + C_{damping}\frac{dx}{dt}$$

Here, $F_{\text{int}\,ernal}$ and $F_{external}$ are the internal forces and external forces applied on the mass points respectively. $x(t)$ is the location of mass point at time t. $C_{damping}$ is the damping coefficient. For cloth simulation, the motion of masses is always described with second order or higher order differential equation of time t. When simulating the dynamic cloth, the initial state is known. Solving equations of the cloth motion is a solving initial conditions problem of the ordinary differential equations.

In these schemes of solving equations, fourth-order and fifth-order Runge-Kutta scheme are applied widely in deformation simulations because of their excellent performances. However, a better solution to solving ordinary differential equations of the cloth behavior should be that time steps are adaptive and errors are predictable.

3.2 Problems with Runge-Kutta Scheme

To animate such a spring-mass system, the following fourth-order Runge-Kutta scheme can be used:

$$v_{i+1} = v_i + h(f_1 + 2f_2 + 2f_3 + f_4)/6$$
$$x_{i+1} = x_i + v_{i+1}dt \qquad (1)$$

Here, $v_{i+1} = v_i + h$

$f_1 = f(t_i, v_i)$;

$f_2 = f(t_i + \frac{h}{2}, v_i + \frac{h}{2} f_1)$;

$f_3 = f(t_i + \frac{h}{2}, v_i + \frac{h}{2} f_2)$;

$f_4 = f(t_i + h, v_i + hf_2)$;

f is the acceleration of the mass point at time t;

v_i is velocity of the mass point at time t_i;

h is a fixed time step of the simulation system;

The final overall errors of Runge-Kutta scheme achieves to $O(h^4)$ and partial errors achieves to $O(h^5)$. That is, the higher the orders of Runge-Kutta scheme is, the more accuracy the simulation system is, and the more complicated the computation is. At the same time, since we often have to handle collisions (which give rise to discontinuities in the motion during simulation), this scheme is not appropriate. The scheme is a recurrent process. The computation may not a crucial problem, whereas the difficult problem is the choice of time step h. The simulation system can diverge rapidly since assuming the force as constant over too large a time step may induce a wild change in position. In practice, we effectively notice a stable behavior of the system only for a small time step. Therefore, Runge-Kutta scheme in cloth simulation is often unsatisfied in practice. We can adjust the time steps to reduce the computation times: small time steps are adopted in some parts of the simulated cloth, while large time steps are adopted in other parts. That is, according to the cloth animation requirements, time steps are adaptive to the simulation and the system is stable. In the following section, we perform a study of Runge-Kutta method relatively to embedding parameters that control the time steps and errors.

3.3 Runge-Kutta Method of Embedding Parameters

An alternative stepsize adjustment method is based on the embedded Runge-Kutta formulas, originally invented by Fehlberg. An interesting fact about Runge-Kutta formulas is that for orders M higher than four, (M+1) function evaluations are required.

Fehlberg [11,13] proposed a fifth order method with six function evaluations where another combination of the six functions gives a fourth-order method. The difference between the two estimates can then be used as an estimate of the truncation error to adjust the time step size. Since Fehlberg's original formula, several other embedded Runge-Kutta formulas have been found.

The general form of a fifth-order Runge- Kutta formula is:

$$f_1 = hf(t_i, v_i)$$
$$f_2 = hf(t_i + a_2 h, v_i + b_{21} h f_1)$$
$$f_3 = hf(t_i + a_3 h, v_i + b_{31} f_1 + b_{32} f_2)$$
$$f_4 = hf(t_i + a_4 h, v_i + b_{41} f_1 + b_{42} f_2 + b_{43} f_3)$$
$$f_5 = hf(t_i + a_5 h, v_i + b_{51} f_1 + b_{52} f_2 + b_{53} f_3 + b_{54} f_4)$$
$$f_6 = hf(t_i + a_6 h, v_i + b_{61} f_1 + b_{62} f_2 + b_{63} f_3 + b_{64} f_4 + b_{65} f_5)$$
$$v_{i+1} = v_i + c_1 f_1 + c_2 f_2 + c_3 f_3 + c_4 f_4 + c_5 f_5 + O(h^5) \quad . (2)$$

The embedded fourth-order formula is:

$$v'_{i+1} = v_i + c'_1 f_1 + c'_2 f_2 + c'_3 f_3 + c_4' f_4 + c_5' f_5 + c_6' f_6 + O(h^6) \quad (3)$$

From the equation (2) and (3), the error estimate:

$$error = v_{i+1}' - v_{i+1} \quad (4)$$

Here, $error$ is the controllable tolerance of specified error. Compared the error with the allowable accuracy, the system can decide if time step is adjusted automatically.

Runge-Kutta-Cash-Karp[11,13] is an another method of solving ordinary differential equations. This method is essentially the same as the Runge-Kutta-Fehlberg method. The motion equations of solving initial conditions with Cash-Karp method are similar with Fehlberg method. The difference between these two methods is the coefficients. The former is said to give a more efficient method. In these two methods, we can determine the new time step h_{new} with the help of the accuracy, the old time step h, and the error $|v_{i+1}' - v_{i+1}|$, the expression is:

$$h_{new} = h \left| \frac{accuracy}{error} \right|^{1/5} \quad (5)$$

Runge-Kutta-Cash-Karp method is the most accurate integration technique available on the simulator, having a sixth order error term.

4. EXPERIMENTS AND CONCLUSIONS

We animate the cloth motion with general fourth order Runge-Kutta (RK4) method, Runge-Kutta -Cash-Karp (RKCK) method and Runge-Kutta-Fehlberg (RKF) method respectively. The following simple procedure provides the final part of our experiments.

Fig 3-5 shows cloth simulation graphics with these three methods:

Table 1 shows the data comparisons of these three methods in cloth simulation on HP graphics workstation XW6000.

In summary, our experiments show that a good locality behavior and a scalable error strategy are crucial for an efficient solution of ODEs of cloth simulation using embedded Runge-Kutta methods. However, due to the large range of different types and characters of cloth simulation systems, hardware and memory architectures, compilers and optimization techniques, we cannot give a general recommendation on which implementation should be used.

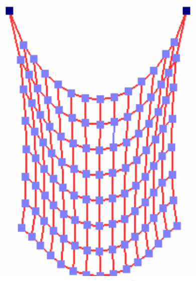


Fig.3.RK4 Method(t=0.25ms)

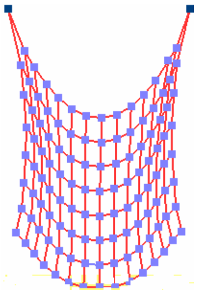


Fig.4.RKF Method (t=0.25ms)

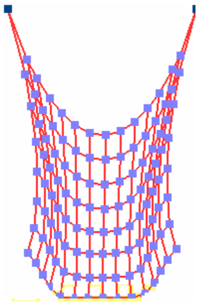


Fig.5.RKCK Method (t=0.25ms)

Table 1 Experiment Data Comparison

Method	RK4	RKF	RKCK
Simulation points	120	120	120
Time step (ms)	0.005	Adaptive	Adaptive
Time to stability	130ms	145ms	146ms
Error	bigger	small	sammler
Precision	high	higher	highest
Frames to stability	260	266	271
Stability	good	better	better
Implement	easy	easier	easier

For a simple simulation system that emphasized particularly on the computation time, RK4 maybe a better choice to solve the cloth motion equations. But it results the precision decrease and a bad stability. For a cloth simulation system that emphasized particularly on precision of system, embedded RK methods are the best choice to achieve satisfied simulation results.

Considering the controllable error and stability, embedded RK methods are currently successful in ODEs solutions of cloth simulation system. However, the careful selection of the most efficient implementation for the solution of a specific cloth simulation system is usually worth the effort as it can save a large percentage of computation time. Further improvements of the performance of embedded RK solvers might be obtained by a specialization in a particular RK method.

REFERENCES

[1] D.Parks, D.A. Forsyth, Improved Integration for Cloth Simulation, Eurographics short presentations, 2002.

[2] P. Volino, N. Magnenat-Thalmann, Accurate Garment Prototyping and Simulation,Computer-Aided Design and Applications, CAD Solutions, 2(5), pp 645-654, 2005.

[3] Kwang-Jin Choi, Hyeong-Seok Ko. Stable but responsive cloth. In Proceedings of the 29th annual conference on Computer graphics and interactive techniques, pp 604–611, ACM Press, 2002.

[4] M. Oshita, A. Makinouchi, Real-time Cloth Simulation with Sparse Particles and Curved Faces. Proceedings of Computer Animation 2001, Seoul, Korea, November 2001.

[5] P. Volino, N. Magnenat-Thalmann, Comparing efficiency of integration methods for cloth simulation, in Computer Graphics International 2001, pp 265–272, 2001.

[6] B. Eberhardt, A.Weber, W.Straßer, A fast, flexible particle-system model for cloth draping, IEEE Computer Graphics and Applications, vol. 16, no. 5, pp 52–59, Sept. 1996.

[7] J. Eischen, T. May-Plumlee, N. Kenkare, P. Pandurangan. Accurate 3d Virtual Drape Simulations: A Developmental Method, Proceedings of International Textile and Apparel Association 2003 annual conference.

[8] John H.Mathews, Kurtis D.Fink. Numerical Methods Using MATLAB (Fourth Edition). Pearson Education, Inc., 2004.

[9] M. Prieto, R. Santiago, D. Espadas, I. M.Llorente, F.Tirado, Parallel Multigrid for Anisotropic Elliptic Equations. Journal of Parallel and Distributed Computing, vol. 61 (1) pp.96-114. Academic Press, 2001.

[10] Etzmuss, O., Gross, J., Straßer W. Deriving a Particle System from Continuum Mechanics for the Animation of Deformable Objects, IEEE Transaction on Visualization and Computer Graphics, Vol. 9 (4), pp. 538-550, 2003.

[11] Hauth M., Etzmuss O., Straßer W. Analysis of numerical methods for the simulation of deformable models, The Visual Computer, 2002.

[12] Parks D., Forsyth D.A..Improved integration for cloth simulation, EUROGRAPHICS Short Presentation, 2002.

[13] Jacob W. Foshee. “Resolution Independent Curved Seams In Clothing Animation Using a regular particle grid,” master thesis, 2004.

[14] E. Hairer, S. P. Nørsett, and G. Wanner. Solving Ordinary Differential Equations I: Nonstiff Problems. Springer-Verlag, Berlin, 1993.

[15] D. Baraff, A. Witkin. “Large steps in cloth simulation,” In *ACM Computer Graphics (Proceedings of SIGGRAPH'98)*, pp. 43-54, 1998.

Xinrong Hu is an associate professor of Wuhan University of Science & Engineering. She graduated from Wuhan University of Technology.in 1997, and a Ph.D candidate in Pattern Recognition and Intelligent Control of Huzhong University of Science & Technology. Her research interests include numerical analysis of fabric drape and simulation, dynamics and control of flexible mechanisms and physics properties analysis in cloth simulation, virtual reality.

Investigation of Shape Retrieval Based on HAAR’S Function and Hierarchical Evolution Algorithm

Zhou Ge, Shihong Qin
Wuhan Polytechnic University, Wuhan, 430023 P.R.China
Email: Qinsh@whpu.edu.cn, Shinegogo@ sohu.com

ABSTRACT

The outline of a shape image is described using Haar’s function. A multi-scale shape matching approach is presented based on discrete curve evolution. Some algorithm about the hierarchical polygon evolution & the comparability measurement degree for two outlines are put forward. The experimental studies are carried using Haar’s transformation under different maximum transformation error. The results shows that the approach studied not only decomposes 2D object into polygonal curve but also induces a hierarchical structure of shape, and that combining “Coarse level” with “fine level” in hierarchical similar matching, the retrieval of shape image can be effective for both retrieval accuracy and efficiency.

Keywords: Haar’s Transformation, Hierarchical Haracterization, Shape Retrieval

1. INTRODUCTION

The recognition of the object is a question extensively concerned in the computer vision research fields[1], There has been a few approaches of resolving these problems put forward, such as the template matching, the string match, the match of the form feature point, the dynamic programming, the chart matching, slack and elasticity matching etc. But these approaches mostly according to model[2], they limit at the particular image type. As the sharp of a portrait data processed increases, the ways of the real-time management and retrieval to the shape are necessary to be carried out for application fields [3].

A kind of shape matching method of outline evolution is put forward. Using the Haar’s transformation and given different maximum permissible error of transformation, the object can be decomposed into polygonal curve with hierarchical structure. The study results show that the retrieval of shape image can be effective for both retrieval accuracy and efficiency.

2. PRINCIPLE OF HIERARCHICAL EVOLUTION ALGORITHM

2.1 The evolution of hierarchical polygon

How to approach image outline feature is discussed using a series of line segments connected with each other. The purpose is to realize shape retrieval with compact polygon model using a kind of controllable evolution method to decompose curve outline to polygon. In order to check the line segment on the outline, we describe outline P with coordinates:

$$P=\{(\Delta X_n,\Delta Y_n)=(X_n-X_{n-1}),(Y_n-Y_{n-1})\} \qquad (1)$$

Supposing the value of the coordinates is a integrity and in proper order, namely on the outline, each point has corresponding coordinates. The value of coordinates of the next point equals to that of n decreasing 1, △xn and △yn take the value from ｛-1,0,1｝. Regarding equation (1) as a picture related n, the outline line segment becomes a horizontal line in the picture because of invariable inclined rate. The sharp shift of the line segment direction in its link place causes the discontinuity describing for the whole picture. Thus, the outline is decomposed to line segments in the form of the square of wave. By selecting the fundamental function to keep the discontinuity, and naturally the changing characteristic between the line segments is sustained.

Haar’s function was put forward in 1910[4]. It constitutes an integrity regular function set. For the points with equal interval, the discrete transformation of Haar’s function is:

$$h_{o,n}=1 \qquad \text{for} \quad 0 \le n < 2^p$$

$$h_{i,j,n}=\begin{cases} 2^{i/2} & \text{for} \quad 2^{p-i}(j-1/2) \le n < 2^{p-i}(j+1/2) \\ -2^{i/2} & \text{for} \quad 2^{p-i}(j+1/2) \le n < 2^{p-i}(j+1) \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Where N=2^p, i, j are integers,0<i<p,0<j<2i, These functions represent the unsymmetrical square wave, where the pulse width is related with the square of number of the points. Fig.1 shows the Haar’s functions at the beginning 8 points.

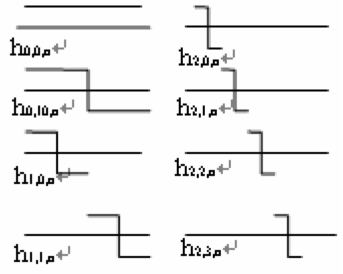


Fig.1. Harr’s functions

Harr’s function includes i natural groupings, in which each function is marked with j. Each grouping is divided into 2i regions, and includes one Haar’s function only..Haar’s function satisfies the following condition:

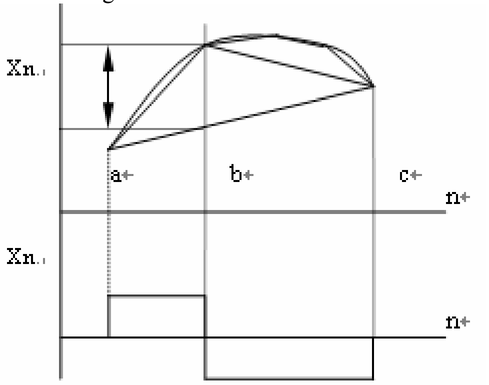


Fig.2. The transform of shape and its corresponding Haar’s function

$$\Sigma h_{i,j,n}h_{k,l,n}=\begin{cases} 1 & i=k \quad \text{and} \quad j=1 \\ 0 & i\ne k \quad \text{and} \quad j\ne 1 \end{cases} \qquad (3)$$

Because the functions constitute integrity set, either of these function fn sequences can be described as the sum of Haar's functions.

Where the amplification coefficient was given by the following equation:

$$f_n = \Sigma\, a_{i,j} h_{i,j,n} \qquad (4)$$

$$a_{i,j} = 1/N \Sigma f_n h_{i,j,n} \qquad (5)$$

The amplification coefficients of Haar sequence of the difference between outline coordinates is worked out according to equation (5), some coefficients chosen are set to zero, in order to form an approximate polygon. Each ΔXn and ΔYn can be regarded as function fn, and the amplification coefficients of ΔXn and ΔYn can be calculated according to equation (5). For the three points of equal interval space on the outline, coordinates x was marked as xa, xb, xc individually and the subscripts satisfy c-b=b-a. Selecting these points make the beginning side corresponding to n=a, the midpoint corresponding to n=b, and the ending side corresponding to n=c, which belongs to one of Haar's functions hi,j,n. As shown in Fig.2, the difference of coordinates x corresponding to Haar's function is △xn=ai,jhi,j,n, Thus:

$$Xn = \begin{cases} a_{i,j} 2^{i/2}(n-a) & a \le n < b \\ a_{i,j} 2^{i/2}(2b-n-a) & b \le n < c \\ 0 & n < a \quad \text{or} \quad n \ge c \end{cases} \qquad (6)$$

It can be shown that coordinates x constituted by Haar's function is located between two lines, one of which connects the beginning point n=a and midpoint n=b, the other connects the midpoint and ending point n=c. In this way Haar's function decomposes an outline segment to two lines which are located at these points.

The amplification coefficient ai,j corresponding to Haar's function is taken into account below, the sum of equation (5) limits to a≤n≤c, hi,j,n=0 beyond this limit.

$$\begin{aligned} a_{i,j} &= 1/N \Sigma_{n=a}^{c} \Delta X_n h_{i,j,n} = 2^{-p+i/2}(\Sigma_{n=a}^{b} \Delta X_n - \Sigma_{n=b}^{c} \Delta X_n) \\ &= 2^{-p+i/2}(X_b - (X_a + X_c)/2) \end{aligned} \qquad (7)$$

The last term of the equation (7) is the distance which begins at the outline midpoint to the line $\overline{x_a\, x_c}$ as shown in Fig. 2. So the coefficient of equation (4) is related to the error caused by using line segment to take place of the outline along the midpoint. Equation (4) is explained as a approximate result which comes from using more and more thick line segments to take place of curve outline where each segment is the half of the former one.

Equation (7) can be use as a criterion for filtration. Given the maximum permissible error ξ which the segment midpoint strayed off outline is, every amplification coefficient is tested from the highest to the lowest (the maximum substripts are i and j) in order.

If the coefficient related to ith，jth satisfies:

$$\left|a_{i,j}\right| \ge 2^{1-p=i/2}\zeta \qquad (8)$$

The coefficient is preserved, otherwise it is thought as nothing important, in this case it is set 0, then the corresponding Haar's function can be efficiently eliminated from the sequence.

By filtering the amplification coefficient, the coordinates difference ΔX_n^F and ΔY_n^F of the outline can be calculated. The beginning point and ending point of the line segment which approximate to the outline can be gotten from here.

Algorithm 1: the hierarchical polygon evolution of outline shape

(1) Step 1: Divide the outline curve into N(N=2p) equal parts randomly. Select one point as the reference point among them. Give the coordinates a different description of the outline equinoctial points.

(2) Step 2: Outline coordinates difference can be further described as the expanded form △xn=ai,jhi,j,n of Harr's function, where 0<i<p,0<j<2i.

(3) Step 3: Give the maximum permissible error ξ which the segment midpoint strayed off outline, then every amplification coefficient is tested from the highest to the lowest in order, if $\left|a_{i,j}\right| \ge 2^{1-p=i/2}\zeta$, the coefficient is preserved, otherwise, hi,j,n=0.

(4) Step 4: By filtering the amplification coefficient, the approximate polygon description of the outline can be obtained.

2.2 The comparability measurement of the outline shape

In the way of evolution transform, any outline curve C of digital image can be described using as many approximate polygons as possible in order to approach it. Supposing digital curve C is a polygon whose apexes are $v_0v_1v_2$……v_m, then curve C can be divided into digital segment D(C)=S_0……S_m, where S_i is the segment connected Vi with V_{i+1}. Again supposing the arc length $S(V_k)$ denotes the distance from V_0 to V_k on curve C along counter clockwise direction, when the length of curve C is normalized, the total length of the curve is equal to 1, and $S(V_m)=1$ is obtained. Taken the beginning line segment S0 on the polygon as the reference direction, the relative angle δ(s) of line segment S in counter clockwise direction is acquired.

Definition 1: The corner function Θ(s) is the function of the relative angle along counter clockwise direction and the arc length.

Now the question about object evolution polygon and comparability measurement of the model can be further discussed. For two polygons A and B and their corresponding corner functions $\Theta_{A(s)}$ and $\Theta_{B(s)}$, the degree of comparability between A and B can be obtained from the distance between corner function $\Theta_{A(s)}$ and $\Theta_{B(s)}$. Letting the reference point A move a distance t along the outline, $t \in [0,1]$, then the new corner function $\Theta_A(s+t)$ is obtained. For all value of t, calculate the minimum value, namely:

$$d(a,b) = \text{main}_{t\in[0,1]}\left(\int_0^1 \left| \Theta_A(s+t) - \Theta_b(s) \right|^2 ds\right)^{1/2} / 2\pi \qquad (9)$$

When the displacement operation is carried on along outline a, that is $\Theta_A(s) \to \Theta_A(s+t)$, $t \in [0,1]$, the breakpoint of arc a will meet with that of arc b. Value t on the condition that breakpoint of outline a and outline b meet each other is defined as a key event. If outline a has m breakpoints, outline b has n breakpoints, there are mn key events at the most. If S_0S_1…S_N are the n key events of $\Theta_a(s)$ respectively, when t= S_0, S_1,…S_N, the distance measure between segment a and segment b can be calculated respectively, the minimum value among them is the final comparability measurement required.

Algorithm 2: the comparability measurement degree for two random curves A and B

```
(1) Step 1: Call algorithm 1, the evolution polygon description form of two curves is obtained.
(2) Step 2: Normalize the length of evolution polygon, make the total length equal to 1.
(3) Step 3: Covert evolution polygon to corner function $\Theta_{A(s)}$ and $\Theta_{B(s)}$.
(4) Step 4: Suppose S[0],S[1]…S[n] are the key events of $\Theta_{A(s)}$, S[0]=0
        For j=1 to m do
            For i=1 to n do
            t=S[i]
                compute
$D(i)=\int_0^1 |\Theta_A(s+t)-\Theta_B(s)|^{1/2}/2\pi$ ……..
        end do
        d(A,B)…..
    end do
(5) Step 5: the ultimate matching comparability degree $S_C(A,B)=1-d(A,B)$ ……
```

3. HIERARCHICAL RETRIEVAL OF THE OUTLINE SHAPE

3.1 Hierarchical description of shape

As far as two-dimensional shape is concerned, its description method can be got out based on the occupied outline region. As demand of practical application, the shape description of single layer is usually not enough. Using Haar’s transformation, it can be generated that a set of hierarchical polygon with the sides gradually decreased, and two-dimensional hierarchical shape description of outline can be constituted, which deviates from the original. Where the description of key layer should correspond to those obvious deviations, these layers are the coarse hierarchy, not the key hierarchy layers. Those are the fine hierarchy layers, corresponding to the general deviations. Fig.3 shows evolution processing of a hierarchy shape.

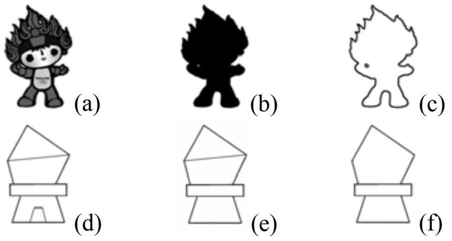


Fig.3. hierarchical evolution processing

(a) the color image of a rabbit (b)two values image (c) evolution shape of outline $\xi=1$ (d) evolution shape of outline $\xi=4$ (e) evolution shape of outline $\xi=8$ (f)evolution shape of outline $\xi=16$

3.2 Hierarchical retrieval of outline shape

In the process of shape retrieval, the similar matching retrieval way combining “coarse level” with “fine level” in hierarchical is usually adopted. At first use the retrieval method of “coarse level” to compute the distance of key hierarchical shape description, and to filter the absolutely dissimilar images in order to decrease the data size of retrieving image. Then use the retrieval method of “fine level” to compute the distance of fine hierarchical shape description to improve the retrieval accuracy. In this way, the retrieval needs less computer source compared to completely adopting “fine level”.

4. ASSESSMENT OF EXPERIMENT RESULTS

A robust shape matching approach should be independent of direction, displacement, and scale. shown in Fig.4., 9 kinds of patterns of different shape are chosen for assessment respectively. Selecting 8 kinds of samples of different directions, and reducing the scale of shape image to 90%,80%,70% along direction X and Y, thus the sample data base of outline shape of are obtained.

The different retrieval effects obtained by different matching approach are given in table 1. From table 1 the effects are not so good using string matching method[5] and histogram method[6] for shape recognition. The effect of two stage matching methods[7] is very good, and can reach to 100% repeatability. In the same way, hierarchical evolution matching method has the same retrieval effect but quicker than two stage matching method. The result shows that hierarchical evolution matching method is effective for both retrieve accuracy and efficiency.

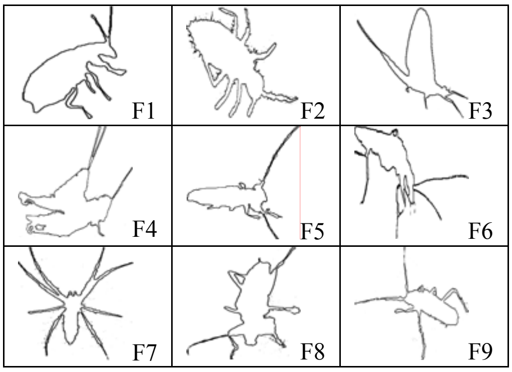


Fig.4. the sample data base of 9 kinds of outline Shape

Table 1 the result comparison of shape retrieve for Fig.5

Method	F1	F2	F3	F4	F5	F6	F7	F8	F9	Time(ms)
String matching	96.1%	100%	100%	98.6%	100%	100%	92.4%	73.4%	87.8%	0.10
Histogram	83.5%	92%	73.6%	65%	100%	92%	87.2%	43.1%	82%	0.10
Two-stage matching	100%	100%	100%	100%	100%	100%	100%	100%	100%	0.20
Evolution matching	100%	100%	100%	100%	100%	100%	100%	100%	100%	0.12

REFERENCES

[1] E.M.Arkin,L.P.chew etc, “an efficiently computer metric for comparing polygonal shape.” *IEEE Trans on Pattern Analysis and Machine Intelligent* Vol13, No3, pp209-216, 1991.

[2] F.stein, “Structural Indexing: Efficient 2-D object Recognition.” *IEEE Trans on Pattern Analysis and Machine Intelligent* Vol14, No12, pp1855-1870, 1992

[3] R.Horaud and H.Sossa, “Polyhedral Object Recognition by indexing.” *Pattern Recognition*, Vol 28, No12, pp1855-1870, 1995

[4] K.G.Beauchamp, *Walsh function and their Application*. New York Academic Press, 1975

[5] S.W.Chen, S.T.Tung and C.Fang, “Extended attributed string matching for shape recognition.” *Computer Vision and Image Understanding* Vol70, No1, pp36-50, 1998

[6] Benoit Huet and Edwin R.Hancock, “Line pattern retrieval using Relational histogram.” *IEEE Trans on Pattern Analysis and Machine Intelligent* Vol21, No12, pp1363-1370, 1999

[7] Wen-Yen Wu and Mao-Jian J, “Two-dimensional object recognition trough two-stage string matching.” *IEEE Trans on Image Processing.* Vol8, No7, pp978-981, 1999

Shihong Qin is a full professor and a vice dean of electrical and information engineering department, graduated from Hua Zhong university of science and technology in January 1999, and awarded PH.D degree. His most interest fields are information processing, electrical intelligentized instrument on-line lest, etc.

Medical Image Segmentation by Geodesic active contour methods*

Guiyun Ye[1] Changzheng Liu[2]
[1]College of Electrical and Information Engineering, Heilongjiang Institute of Science and Technology
Harbin, Hei Longjiang, 150027, P.R.China
Email: yeguiyun@yahoo.com.cn
[2]College of Computer Science and Technology, Harbin University of Science and Technology
Harbin, Hei Longjiang, 150080, P.R.China
Email: fox@hrbust.edu.cn

ABSTRACT

Geodesic active contour methods are powerful numerical techniques for image segmentation and analysis. However the edge detector in the model ensures that the data on both sides of the contour is as dissimilar as possible, it cannot deal with the requirement that also the interior of a region should be as homogeneous as possible. In this paper, we present a new region based geodesic active contour method. The new method gives a global view of the boundary information within the image. The method here proposed is particularly well adapted to situations where edges are weak and overlap, and the curve initialization is a very bad guess. A number of experiments on CT medical images were performed to evaluate the new method. The experimental results demonstrate the reliability and efficiency of this new method.

Keywords: Image Segmentation, Initial Contour,Geodesic Active Contour, and Region Statistics Information, Gradient Information.

1. INTRODUCTION

Active contour models, also known as snakes were introduced by Kass[1] et al.(1988). The classical approach is based on deforming an initial contour C0 towards the boundary of the object to be detected. The deformation is obtained by trying to minimize a functional designed so that its (local) minimum is obtained at the boundary of the object. These active contours are examples of the general technique of matching deformable models to image data by means of energy minimization. The energy functional is basically composed of two components, one controls the smoothness of the curve and another attracts the curve towards the boundary. Geodesic active contour [2] is a particular case of active contour models. In this case the classical energy snakes model is proved to be equivalent to finding a geodesic curve in a Riemannian space with a metric derived from the image content. This means that in a certain framework, boundary detection can be considered equivalent to finding a curve of minimal weighted length. Then, assuming that this geodesic active contour is represented as the zero level-set of a 3D function, the geodesic curve computation is reduced to a geometric flow that is similar to the one obtained in the curve.

The level set approach were introduced by S. Osher and J. A. Sethian[3] in 1988, which actually originate from computational fluid dynamics. Level sets are designed to handle problems in which the evolving interfaces can develop sharp corners and cusps, change topology and become very complex. In the level set approach, the convergence to the final result may be relatively independent of the initial shape, and branches, splits and merges can develop without problems as the front moves. Generally, the method may be applied even where no a priori assumptions about the object's topology are made. Meanwhile, the level set framework has become a very popular tool for curve representation in image analysis.

2. GEODESIC ACTIVE CONTOUR

In this section we discuss the geodesic active contour and it's advantages, shortcomings.

The geodesic active contour model was proposed by Caselles in 1997[4]. Based on the curve length minimization with mean curvature motion, geodesic active contours additionally introduce an edge detector g(|▽I|), which performs the task to draw the curve towards edges in the image I, a possible choice is, for instance Eq. (1):

$$g(|\nabla I|) = \frac{1}{\sqrt{|\nabla I|^2 + \epsilon^2}} \qquad (1)$$

This edge indicator can be added to the evolution Eq. (2) in form of a weighting factor:

$$\partial_t \Phi = |\nabla \Phi| \operatorname{div} \left(g(|\nabla I|) \frac{\nabla \Phi}{|\nabla \Phi|} \right) \qquad (2)$$

Applying the chain rule, Eq. (2) can be rewritten as Eq. (3):

$$\partial_t \Phi = g\, |\nabla \Phi| \operatorname{div} \left(\frac{\nabla \Phi}{|\nabla \Phi|} \right) + \nabla g^\top \nabla \Phi \qquad (3)$$

It becomes obvious that the evolution equation actually consists of two terms. The first one performs mean curvature motion weighted by the edge indicator function g. With solely this term, the curve would still shrink to a circular point for t→ ∞, since g > 0. The shrinkage would only be slowed down in the presence of edges, yet it could not be stopped. Only the second term in the evolution equation can prevent further shrinkage, as it evolves the curve in direction of smaller g. With this term, the curve can even evolve in outer normal direction, provided the attracting image edge is sufficiently steep and close enough to the evolving contour.

This already indicates the dependence of the outcome from the initialization of the contour. The evolution equation is only a local optimization method. Consequently, it can yield only the next local optimum. This means that the curve is attracted by

* This work is supported by the Natural Sciences Foundation of Heilongjiang Province under Grant QC04C44, Doctor Foundation mgb05006, Natural Sciences Foundation of China mgz06011 and the Foundation of office of Education of Heilongjiang Province under Grant 11511070.

the next relevant edge in the image, although an edge further away from the initial curve could yield a smaller weighted curve length.

In conclusion, the geodesic active contour model has several advantages:

① Its main advantage is the sound minimization of the energy functional by means of the discredited steepest descent equation. There are no algorithmic supplements like regrinding necessary.

② Additionally, the implicit curve representation allows for topological changes. Hence, partially occluded objects that are split into several parts can be handled without additional efforts.

③ The energy functional is very simple and contains only a few parameters, this is the choice of the edge detector and the edge weighting function g.

Apart from this, however, the model has also some shortcomings:

① First of all, the interior of the extracted regions is completely neglected. The only cues that drive the evolution are the edges in the image. However, edges are not very reliable for extracting objects. Especially in the presence of textured objects, edges do hardly contain any useful information for the segmentation.

② Since the evolution is drawn to the next significant edge in the image, any edge in between the initialization and the sought object contour can attract the solution. While in supervised medical segmentation applications this strong dependence on the initialization might be considered as beneficial, it is certainly a severe disadvantage in unsupervised segmentation.

The goal of the technique described in the next section is to keep the advantages of geodesic active contours while addressing its shortcomings.

3. GEODESIC ACTIVE CONTOURS

Both problems that appear with the geodesic active contour model originate from the negligence of region information. While the edge detector in the model ensures that the data on both sides of the contour is as dissimilar as possible, it cannot deal with the requirement that also the interior of a region should be as homogeneous as possible. An immediate consequence of these considerations is to include region information in the active contour model. An important question that arises with region based active contours is how to model the interior of regions. The Chan-Vese[5] model based on the Mumford-Shah functional employs actually the simplest possible model, namely the model of piecewise constant regions. One can certainly think of cases where this model is not appropriate to distinguish the particular regions where the two regions contain Gaussian noise with the same mean but different standard deviation. In such a case, where the mean value is not sufficient to distinguish the regions, segmentation with the piecewise constant model must fail.

A reasonable way to find a remedy for this shortcoming is to extend the complexity of the region statistics. According to the maximum a-posteriori criterion, the segmentation has to maximize the a-posteriori probability P(M|D) of the model M given the data D. Reinterpreting this criterion in the notation of the active contour model, the model M is one of the regions, i.e. either Ω 1 or Ω 2 and the data D is the image gray value I. The Bayes rule allows to express P(x∈ Ω i|I(x),i={1,2},as

$$P(\mathrm{x}\in\Omega_i|I(\mathrm{x})=s)=\frac{P(I(\mathrm{x})=s|\mathrm{x}\in\Omega_i)P(\mathrm{x}\in\Omega_i)}{P(I(\mathrm{x})=s)}. \tag{4}$$

While P(x∈ Ω i) is the a-priori probability of region Ω i, which allows to integrate further assumptions about the region besides its region statistics, pi(s) := P(I(x) = s|x∈ Ω i) is the probability density in region Ω i, i.e, the gray value distribution within this region. The a-priori probability of the image gray values P(I(x) = s) is independent of the choice of the region and can therefore be neglected for optimization.

The only a-priori knowledge about the regions that is available at this point is the length constraint on the object contour. Supplementing the assumption that the gray values of particular image pixels are independent, which is reasonable as long as we do not know the kind of interdependence between the pixels, this yields the task to maximize [6]

$$\prod_{\mathrm{x}\in\Omega_1} p_1(\mathrm{x}) \prod_{\mathrm{x}\in\Omega_2} p_2(\mathrm{x}) \;\; e^{-\nu\int_\Gamma ds} \tag{5}$$

or equivalently to minimize

$$E(\Gamma)=-\int_{\Omega_1}\log p_1(\mathrm{x})\,\mathrm{dx}-\int_{\Omega_2}\log p_2(\mathrm{x})\,\mathrm{dx}+\nu\int_\Gamma ds. \tag{6}$$

It can further be expressed by means of the level set framework[7]

$$E(\Phi)=\int_\Omega\Big(-H(\Phi)\log p_1-(1-H(\Phi))\log p_2+\nu|\nabla H(\Phi)|\Big)\,\mathrm{dx}. \tag{7}$$

In fact, the Chan-Vese model is only a special case of this energy functional. A typical choice to model the probability density functions is a Gaussian function with mean μ and standard deviation δ :

$$p(s)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(s-\mu)^2}{2\sigma^2}}. \tag{8}$$

Since the energy in and Eq. (6) is formulated by means of a level set function, minimization can be performed by a gradient descent according to the Euler-Lagrange equations of Eq. (6):

$$\partial_t\Phi=H'(\Phi)\left(\log p_1-\log p_2+\nu\,\mathrm{div}\left(\frac{\nabla\Phi}{|\nabla\Phi|}\right)\right) \tag{9}$$

The probability densities p1 and p2 have to be updated after each iteration. For the Gaussian model in Eq. (8) this comes down to updating

$$\mu_1=\frac{\int_\Omega I\,H(\Phi)\,\mathrm{dx}}{\int_\Omega H(\Phi)\,\mathrm{dx}} \tag{10}$$

$$\mu_2=\frac{\int_\Omega I\,(1-H(\Phi))\,\mathrm{dx}}{\int_\Omega(1-H(\Phi))\,\mathrm{dx}} \tag{11}$$

$$\sigma_1=\sqrt{\frac{\int_\Omega(I-\mu_1)^2\,H(\Phi)\,\mathrm{dx}}{\int_\Omega H(\Phi)\,\mathrm{dx}}} \tag{12}$$

$$\sigma_2=\sqrt{\frac{\int_\Omega(I-\mu_2)^2\,(1-H(\Phi))\,\mathrm{dx}}{\int_\Omega(1-H(\Phi))\,\mathrm{dx}}}. \tag{13}$$

In contrast to the piecewise constant model, the Gaussian probability density function in Eq. (8) is capable to capture the boundary between the two regions.

In the case where the curve initialization is a very bad guess, it yields slightly different probability densities in the two regions. Thus for each specific gray value in the image, the assignment of the pixel to one region is more probable than to the other. This difference in the probability drives the evolution of the curve in Eq. (9). The curve evolution, on the other hand, causes the difference in the probability density functions to grow until the optimum assignment of all pixels is reached. As a result, one region captures the area with small standard deviation, while the other region takes care of the rest.

4. EXPERIMENTAL RESULTS

To demonstrate the performance of our new method, we carried out a series of experiments on medical images.

The key challenge for the neurosurgeon during brain surgery is to remove as much as possible of a tumor without destroying healthy brain tissue. This can be difficult because the visual appearance of healthy and diseased brain tissue can be very similar. It is also complicated by the inability of the surgeon to see critical structures underneath the brain surface as it is being cut. It was our goal to be able to rapidly and faithfully capture the deformation of the brain during neurosurgery, so as to improve intraoperative navigation by allows preoperative data to be aligned to volumetric scans of the brain acquired intraoperatively. Image guided surgery techniques are used in operating rooms equipped with special purpose imaging equipment. The development of image guided surgical methods over the past decade has provided a major advance in minimally invasive therapy delivery. Image guided therapy has largely been a visualization driven task. Quantitative assessment of intraoperative imaging data has not been possible in the past, and instead qualitative judgements by experts in the clinical domains have been relied upon. In order to provide the surgeon or interventional radiologist with as rich a visualization environment as possible from which to derive such judgements, previous work has primarily been concerned with image acquisition, visualization and registration of intraoperative and preoperative data. Biomechanically accurate registration of brain scans acquired during surgery, as proposed here for fluoroscope, has the potential to be a significant aid to the automatic interpretation of intraoperative images and to enable prediction of surgical changes.

The fluoroscope navigation system (FNS) is one of optical tracking system (OTS), which can be utilized in a variety of trajectory-based procedures such as tumor biopsy, catheter placement and endoscopy, thus obviating the need for framed stereotaxis. When comparing the FNS to a framed stereotactic system, the FNS has advantages in the areas of patient acceptance (no ring placement), less cumbersome hardware, and easier trajectory planning. Disadvantages include the need for general anesthesia in most patients for placement of the Mayfield head holder and slightly less accuracy compared to the rigid system. The following case study illustrates these points. The patient is a 40-year-old woman with a one-month history of progressive headaches. Neurological examination was significant only for papilledema. MR scan demonstrated a 3 x 3 x 5cm irregularly enhancing mass located within the anterior corpus callosum with significant surrounding edema. The tumor was deemed unresectable and FNS stereotactic biopsy was recommended prior to the initiation of adjuvant therapy. The patient underwent placement of six scalp fiducials and then 3mm contiguous MR slices were obtained after contrast infusion. At the time of surgery, the patient's head was fixed in the Mayfield threepin headholder and the FNS calibrated using the scalp fiducials. The FNS was used to determine the target, entry point, and trajectory for the tumor biopsy. The proposed entry site was prepped and draped in the usual sterile fashion. A skin incision and burr hole were performed. The 3-D positioner was then attached to the side rail and the FNS depth probe secured at the distal end. The FNS probe was positioned at the burr hole such that the screen cursor was superimposed on the proposed tumor target. The depth probe pointer was then replaced by a standard Radionics reducing tube. A Nashold biopsy needle was set to a depth of 23 cm as measured from the bottom of the plastic depth stop to the center of the side cutting window and passed through the reducing tube to the tumor target. Quadrant samples were taken and frozen section confirmed glioblastoma.

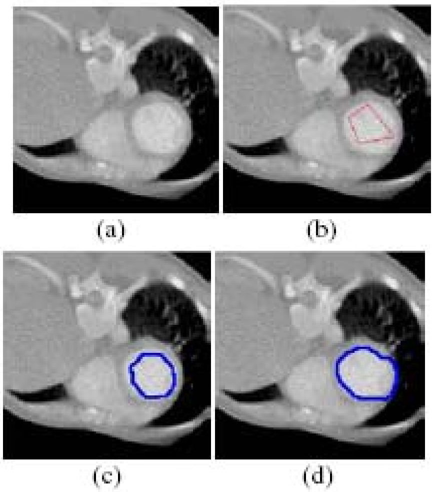


Fig.1. CT liver image segmentation results; (a)original image; (b) initial curve; (c) intermediate iterations; (d) our proposed method results.

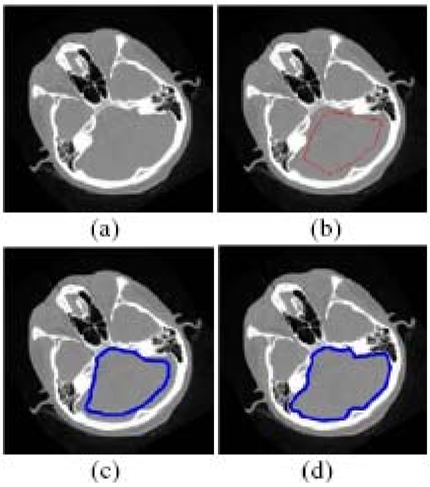


Fig.2. CT brain image segmentation results; (a) original image; (b) initial curve; (c) intermediate iterations; (d) our proposed method results.

As mentioned above, the proposed method seems ideal for use on a wide variety of medical imagery. The power of this method in extracting feature from even fuzzy boundary and overlap boundary medical images has been demonstrated, even the curve initialization is a very bad guess.

5. CONCLUSIONS

In this paper, we have proposed a new method for image segmentation. The new models modify the geodesic active contour utilizing region statistics information and gradient information. The scheme here proposed is particularly well adapted to situations where edges are weak and overlap, and the curve initialization is a very bad guess. The method has been tested with numerical real CT medical images. The experimental results show the reliability of the approach.

In each neurosurgery case several volumetric FNS scans were carried out during surgery. The first scan was acquired at the beginning of the procedure before any changes in the shape of the brain took place, and then over the course of surgery other scans were acquired as the surgeon checked the progress of tumor resection. The final scan in each sequence exhibits significant nonrigid deformation and loss of tissue due to tumor resection. In order to test our biomechanical simulation approach each subsequent scan was aligned to the first by rigid registration using maximization of mutual information. This method computes a global alignment accounting for positioning differences in the scan coordinates but does not attempt to correct for nonrigid deformation. The first scan was manually segmented to act as an individualized anatomical model. The last scan in each sequence was then segmented with our intraoperative segmentation approach. Our algorithm for biomechanical simulation of the brain deformation was then executed to compute the volumetric deformation between between the twodata scans.

The results here focus upon the biomechanical simulation of brain deformation because in the past obtaining sufficiently accurate results in a clinically compatible small amount of time has been seen as extremely difficult. Often less accurate but fast models of deformation have been used. In order to provide a context for the biomechanical simulation amongst the other image analysis and acquisition tasks that must occur intraoperatively.

Ultimately the ability to use these intraoperative image analysis methods relies upon them being sufficiently robust to provide accurate results for typical clinical cases, and critically, to be sufficiently fast to provide feedback to the surgeon at a rate that can be practical to use during neurosurgery. We collected performance results on three different architectures in order to assess the absolute performance and the scaling behavior of our parallel implementation.

REFERENCES

[1] M. Kass, A. Witkin, and D. Terzopoulos. “Snakes: Active contour models”. *International Journal of Computer Vision*, pp321~331, 1988.

[2] V.Caselles, R.Kimmel, and G.Sapiro. "Geodesic active contours". *International Journal of Computer Vision*, 21:61~79, 1997.

[3] S.Osher and J.A.Sethian. "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations". *Journal of Computational Physics*, 79:12~49, 1988.

[4] M.Rousson and R.Deriche. "Dynamic segmentation of vector valued images". *Geometric Level Set Methods in Imaging, Vision and Graphics*. Springer, August 2003.

[5] L.D.Cohen, E.Bardinet, and N.Ayache. "Surface reconstruction using active contour models".In *SPIE Conference on Geometric Methods in Computer*

[6] T. Chan and L.Vese. "An active contour model without edges". In M. Nielsen, P. Johansen,O. F. Olsen, and J. Weickert, editors, *Scale-Space Theories in Computer Vision*, volume 1682 of Lecture Notes in *Computer Science*, pages 141~151. Springer, 1999.

[7] S.C. Zhu and A.Yuille. "Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884~900, September 1996.

[8] Reddi AH, Roodman D, Freeman C, Mohla S. Review. “Mechanisms of tumour metastasis to the bone: challenges and opportunities,” J *Bone Mineral Res 2003*; 18: 19~24

[9] Ratanatharathorn V, Powers WE, Moss WT, Perez CA. “Bone metastasis. Review and critical analysis of random allocation trials of local field treatment,” Int J *Radiat Oncol Biol Phys* 1999; 44 : 1~18

[10] Porter AT, McEwan AJ, Powe JE et al. Results of a randomised phase-III trial to evaluate the efficacy of strontium-89 adjuvant to local field external beam irradiation in the management of endocrine resistant metastatic prostate cancer. Int J *Radiat Oncol Biol Phys* 1993; 25 : 805~13

[11] Quilty PM, Kirk D, Bolger JJ *et al*. “A comparison of the palliative effects of strontium-89 and external beam radiotherapy in metastatic prostate cancer,” *Radiother Oncol* 1994; 31: 33~40

Guiyun Ye is a vice Professor of College of Electrical and Information Engineering, Heilongjiang Institute of Science and Technology, She graduated from Harbin Engineering University in 1986. She has published over 30 Journal papers. Her research interests are in distributed parallel processing, Visualization in Scientific Computing.

Changzheng Liu is a vice Professor of Computer Science and Technology College, Harbin University of Science and Technology. He graduated from Harbin Engineering University in 1993; was a postdoctor of Harbin Medical University (2004~2006). He is secretary-general of Hei Longjiang Biomedical Engineering Society. He has published over 20 Journal papers. His research interests are in distributed parallel processing, Visualization in Scientific Computing.

A New Rate Control Scheme in Low Bit-Rate for H.264/AVC

Yangchun Li, Ruolin Ruan
School of Information Engineering, Xianning College
Xianning, Hubei 437100, China
Email: rlruan@163.com

ABSTRACT

Rate control is not a necessary part of the current video coding standards, as the network bandwidth and delay constraint, Rate control has become a key technique for video coding in order to obtain consecutive and high quality video picture, especially the low bit rate communication need the better efficiency rate control scheme. Rate control Scheme of basic unit in H.264 mainly adopts the linear MAD predict model and quadratic rate distortion model, in the process of implementing, after coding a macroblock, the parameters of model will be updated, and computes the quantization parameter of the current macroblock. So, its computation cost is very high, and its complexity is also very high, so, it is not fit the low bit rate communication, especially the real time communication. Through analyzing the rate control scheme of the H.264/AVC, the paper proposed an improved low complexity MAD weighted predict model, and make the accurate rate control in the macroblock layer to reduce the complexity of the rate control scheme to adapt the real time communication, and lastly carried out it in the JM98 platform of JVT. Extensive experiment results show the complexity of this scheme is lower than the JVT-G012rl of H.264, and the average PSNR of the standard test sequences increased 0.146dB, at the same time, its accuracy of the rate control averagely improved 0.506kpbs, it will fit the low bit rate communication if it is further modified.

Keywords: H.264/AVC, Rate Control, MAD, Macrobock

1. INTRODUTION

Because the bit rate of video bit stream is not constant, in order to obtain high quality the decoded video picture in the constant bit rate (CBR), the encoder must adopt rate control. So the rate control has become an indispensability part of the every encoder. Rate control mainly include two parts: one is rational allocation bit; the other is realization the allocation of the bit and it is general realized through adjusting the quantization parameter.

The literature [1,2] proposed the scheme of bit allocation that is adopted by MPEG-2 TM5, it allocates the constant bit amount for the each group of picture(GOP), the bit amount in GOP will reallocate to each frame and each macroblock, the latter many schemes of bit allocation also continue this idea. In the realization of bit allocation side, the literature [3,4] proposed the rate control scheme based on the using Lagrange rate distortion optimization(RDO), and the scheme is adopted by the TMN8 of H.263, but this scheme use the square difference as the control parameter, and its complexity is very high; the literature [5]discovers that they are correlation between the number of non-zero coefficient through transforming and quantizing and bitrate of video sequence, and proposed a simple linear model, but the scheme must handle the transform coefficients; the literature [6,7] proposed the quadratic quantization model, and this model is very simple, and it is very near the real R-Q relation and it is no necessary to transform the coefficient, so the present many video rate control scheme use it. Up to now, many scholars and research institutions proposed many vary rate control schemes. The typical schemes mainly include MPEG-2 TM5, H.263 TMN8, MPEG-4 VM8 and VM18, ρ-region model and rate control scheme in H.264/AVC JM. TM5 rate control scheme is early research achievement, and rate control scheme of TMN8, VM8 and H.264 JM adopted RD model, and ρ-region rate control scheme is experience model.

Z.G.Li proposed rate control scheme that is adopted by the present standard of H.264/AVC in the literature [8, 9], this scheme uses the layered rate control way, there are GOP layer rate control, frame layer rate control and basic unit (BU) layer rate control. The target bit allocation way of GOP is the same as TM5, and use the constant bit allocation; the frame layer target bit allocation is common decided by network bandwidth, buffer occupancy, buffer size and remain bits; the BU layer target bit is allocated based on mean absolute difference (MAD). JVT-G012rl uses the linear MAD predict model, the model parameters will be updated after handling a BU. These schemes are very successful, at the same time, they also have some problems. For example, because they adopt the linear MAD and quadratic RD models, they increase the complexity of scheme in the process of computing the MAD and parameter of the models.

In this paper, the paper focus on solving the above issues by using the weighted MAD model and an adaptive macroblock layer rate control strategy according to the complexity of vary macrobloack in the same BU. The organization of the rest paper is as follows. The section 2 reviews the rate control algorithm for H.264/AVC. The section 3 proposed the weighted MAD model. The section 4 presents our detailed proposed algorithm. Our extensive experiment results are provided in Section 5. This paper concludes with Section 6.

2. REVIEWS THE RATE CONTROL SCHEME FOR H.264/AVC JM98

2.1 Gop Layer Rate Control

The main purpose of GOP layer rate control is: using the formula (1) to computing the useableness bits of the no coding frame in the current GOP; using the formula (2) to computing the initialize quantization parameter(QP) of the IDR frame and the first stored frame. The detailed description is in the literature [8].

$$B_i(j)=\begin{cases}\dfrac{R_i(j)}{f}\times N_i - V_i(j) & j=1\\ B_i(j-1)+\dfrac{R_i(j)-R_i(j-1)}{f}\times(N_i-j+1)-b_i(j-1) & j=2,3,\ldots,N_i\end{cases} \tag{1}$$

$$QP_1(1) = \begin{cases} 35 & bpp \le l1 \\ 25 & l1 < bpp \le l2 \\ 20 & l2 < bpp \le l3 \\ 10 & bpp > l3 \end{cases} \tag{2}$$

2.2 Frame Layer Rate Control

Considering the particularity of the I frame and B frame, the frame layer bit allocation is no same to TM5, and its bit allocation of I frame and B frame is no relation with the coding complexity of picture. The center of the whole rate control scheme is in P frame. It will decide the target bits for the current coding P frame according to the current buffer occupancy, frame rate, remain bits in GOP and the structure of GOP (IPPP or IBBP), and then compute the QP, and carry out RDO, after coding a frame, it still will update the parameter of the linear predict model and R-D model.

2.3 Basic Unit Layer Rate Control

Supposing a frame video picture composes of N_{mbpic} MB, a BU is a group consecutive N_{mbpic} MB, and the numbers of BU in a frame N_{unit}, computing as following:

$$N_{unit} = \frac{N_{mbpic}}{N_{mbunit}} \tag{3}$$

A BU may be a macroblock, a group consecutive macroblocks, a field or a frame. In general, the more the BU is chosen, the bigger the PSNR, but fluctuation of the coding bits is also very big. On the contrary, the coding bits will be smooth, and the PSNR is small. The BU layer rate control has three mainly processes, the linear MAD predict, allocating target bit and using the quadratic R-Q model to compute QP. JM98 uses the linear model to predict the BU MAD through the BU MAD of the previous frame the same position predicts the current frame BU MAD of the same position. It uses the linear model in the formula (4):

$$MAD_i = a \times MAD_{i-1} + b \tag{4}$$

where a and b are the coefficient of the model, and their initial value is 1.0 and 0, respectively. They will be updated after coding a BU.

According to the predict MAD, using formula (5) to compute the current BU target bit $R_{i.}$

$$R_i = R_r \times \frac{MAD_i^2}{\sum_{j=1}^{N_{unit}-1} MAD_j^2} \tag{5}$$

where R_r is the remain bits of the current frame.

The QP of current BU may be compute through the quadratic R-Q model in the formula (6):

$$R(QP) = MAD_i \times (c / QP + d / QP^2) \tag{6}$$

3. WEIGHTED MAD PREDICTION IN THE BU LAYER [10]

The predict of MAD is very key in the BU rate control, it will affect the target bit allocation, and it is also an input parameter in the process of computing QP. The linear predict model requests real-time updated the parameter a and b. The accuracy of the predict model will be ensured through some input reference points. In general, a beeline is drew up through no less than ten points in the coordinate, some point that has very big error with beeline must be wipe off after completing a linear draw up, and then it will make the second draw up, so it is very complex. The researcher discovers that the actual MAD is correlation in the temporal and spatial through extensive experiments and statistical analysis, and may propose the weighted MAD model to replace the linear MAD predict model by using the correlation of temporal and spatial of macroblock, and it will reduce the computation cost of the predict coefficient updated.

The current BU MAD may be computed through the previous frame actual MAD of the same position and its fore-and-aft two BU actual MAD, the detailed denoted as following:

If the current BU is the first BU of the current frame, then

$$MADofCurre\ ntBU[0] = MADof\ \Pr eviousBU[0] \tag{7}$$

If the current BU is the last BU of the current frame, then

$$\begin{aligned} &MADofCurre\,ntBU[i] = (2 \times MADof\ \Pr eviousBU[i] \\ &+ MADof\ \Pr eviousBU[i-1])/3 \end{aligned} \tag{8}$$

or else

$$\begin{aligned} &MADofCurrentBU[i] = (MADof\ \Pr eviousBU[i] + \\ &4 \times MADof\ \Pr eviousBU[i] + MADof\ \Pr eviousBU[i-1])/6 \end{aligned} \tag{9}$$

Where $MADofCurre\,ntBU[i]$ denotes the ith BU MAD of the current frame, $MADof\ \Pr eviousBU[i]$ denotes the actual MAD of the previous frame the same position. This model is very simple, and it reduces the dynamic update of the predict parameter and the complexity of computation, the effect of predict is very good, it may replace the linear MAD predict model.

4. THE ACCURACY MACROBLOCK LAYER RATE CONTROL STRATEGY [11]

The least unit of rate control is the BU in H.264, and it only makes the RDO in the macroblock layer. In general, to obtain a good trade-off between average PSNR and bit fluctuation, number of macroblock is recommended to be the number macroblock of a row. For example, the number of macroblock in a BU is eleven for the QCIF test sequence, that is to say, the eleven macroblock of a BU use the same QP. As using the constant QP to coding the macroblock of a BU, it doesn’t consider the difference of macroblock in the same BU. The difference of vary macroblock in the same BU is very big. To further accurately allocate bit in the same BU, according to the difference of the macroblock, we may adaptive adjust the QP to obtain the more bit for the some macroblock with bigger activity, and the macroblock with the smaller activity obtain the least bits. We use the MAD of macroblock to denote their activity, the distribution of the macroblock MAD of the 7th frame in Football as Table.1.

Table 1. Macroblock MAD of the 7th Frame in Football

4.19	2.61	3.06	2.93	2.89	2.81	2.97	2.76	2.90	2.97	3.05
2.48	2.91	9.83	5.35	2.80	4.22	7.03	4.70	3.15	6.23	8.82
3.61	7.69	8.55	11.6	5.57	15.4	13.9	10.7	14.8	**18.7**	8.72
6.92	18.1	17.0	13.5	14.9	8.89	16.7	12.9	14.3	21.7	17.5
8.66	15.2	18.7	14.8	12.3	21.2	22.0	18.4	19.0	15.2	13.7
10.4	16.6	14.6	12.2	17.5	26.4	12.6	10.6	13.9	18.8	10.3
14.4	15.0	9.39	7.19	4.66	14.4	15.0	8.62	11.6	16.0	17.0
10.7	8.73	**18.7**	17.7	8.75	3.23	17.2	12.8	12.7	8.34	**3.99**
12.5	10.3	8.25	16.6	4.68	9.72	13.0	10.6	12.3	8.94	3.68

The Table.1 shows difference of macroblock MAD is very big in a frame, and difference of macroblock for the same BU is also very big, the paper may be further control bit allocate in macroblock layer for a BU in order to obtain accurate bit allocation, and the paper imports a parameter Complex$_{MB}$ to scale the complexity of the macroblock.

$$Complex_{MAD} = (3 \times MADofCurrentMB + 2 \times MADofCurrentBU) \\ /(2 \times MADofCurrentMB + 3 \times MADofCurrentBU) \tag{10}$$

The formula (10) describes the relation of the MAD predict value of the current macroblock MADofCurrentMB and the MAD predict value of the current BU MADofCurrentBU, the more the MADofCurrentMB, the more the Complex$_{MAD}$. If the Complex$_{MAD}$ is big, then the complexity of macroblock is very big, and the texture of picture need more bit to describe, so its QP should be reduced, on the contrary, the picture is more smooth, it need very least bits to describe the picture, and its QP may be increased.

Therefore, after making rate control for BU layer, and the QP of macroblock may be adaptive adjusted according to the Complex$_{MAD}$ in the macroblock layer. Through extensive experiments and statistical analysis, the adjusted value DeltaQP for macroblock is defined as following:

```
if (Complex_MAD<=0.75)
  DeltaQP=2;
else if (Complex_MAD<=0.85)
  DeltaQP=1;
else if (Complex_MAD<=1.15)
  DeltaQP=0;
else if (Complex_MAD<=1.35)
  DeltaQP=-1;
else
  DeltaQP=-2;
CurrMB->qp = CurrMB->qp+DeltaQP;
```

5. EXPERIMENT RESULTS AND ANALYSIS

The paper uses the reference software of JVT JM98 to implement the experiment, and uses the 4:2:0 standard test sequences Bus, Football, Foreman, News and Paris, the bitrate is 32, 48, 64kpbs, CABAC is on, the BU is eleven, the results follow as the Table.2.

Where TB denotes the Target Bitrate, AB denotes the Actual Bitrate, P denotes the PSNR(dB), the △B denotes the difference between the AB of JM 98 and our proposed and the TB, △P denotes the difference of PSNR between JM 98 and our proposed.

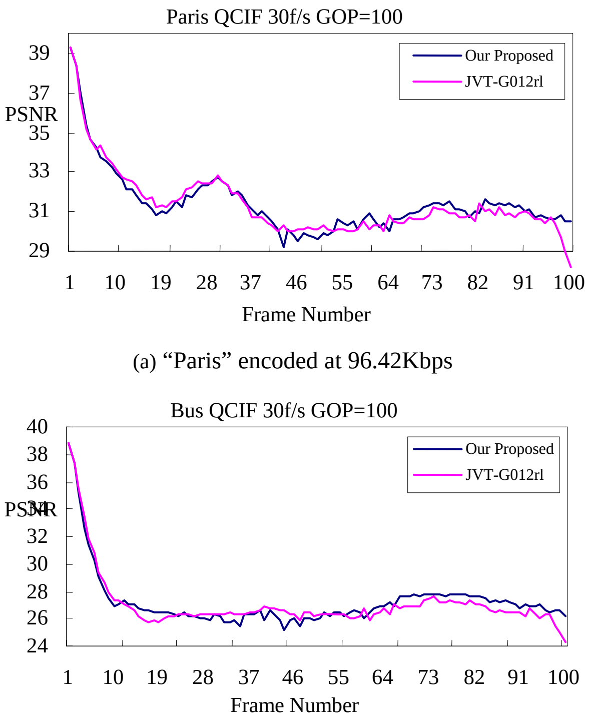
The Table.2. shows the performance of our proposed scheme is more than the JM98 in average PSNR and bitrate, the average PSNR of all test sequences improved 0.11dB, the rate control is more accurate and improved 0.498kpbs. The Fig.1 shows the compare plot of the average PSNR of our proposed and JVT-G012rl for the test sequence Paris and Bus, and the results show that the performance of our proposed is more than JVT-G012rl of H.264/AVC JM98.

6. CONCLUSIONS

The paper firstly introduces the rate control scheme of H.264/AVC, and then analyses its complexity and accuracy of the linear MAD predict model, and further proposed the weighted MAD predict model to replace the linear MAD predict model, and then proposed the QP adaptive adjusted scheme in the macroblock layer based on the BU layer according the complexity of vary macroblock in the same BU to further control bitrate. The extensive experience results show that our proposed rate control scheme reduces effectively the computation complexity of BU MAD, and the average PSNR of standard test sequences increases 0.146dB, at the same time, the accuracy of the rate control also improves 0.506kpbs.

Table 2. Bitrate and PSNR of the Standard Test Sequence

Sequence	TB	JM98		Our Proposed		△P	△B
		AB	P	AB	P		
Bus	32	33.6	20.9	32.4	21.2	+1.2	+0.3
	48	49.9	22.7	48.3	22.8	+1.6	+0.1
	64	65.7	24.6	64.4	24.8	+1.3	+0.2
Football	32	32.1	29.6	32.1	29.9	0	+0.3
	48	49.5	22.8	49.2	22.6	+0.3	-0.2
	64	64.3	23.9	64.3	24.4	0	+0.5
Foreman	32	64.5	22.9	63.4	22.8	-0.1	-0.1
	48	48.2	32.3	48.2	32.5	0	+0.2
	64	64.9	33.9	64.4	33.9	+0.5	0
News	32	32.3	33.6	32.2	33.5	+0.1	-0.1
	48	48.4	35.8	48.4	35.7	0	-0.1
	64	64.7	37.4	64.0	37.8	+0.7	+0.4
Paris	32	33.7	23.3	32.4	23.9	+1.3	+0.6
	48	48.7	27.1	48.3	27.1	+0.4	0
	64	64.5	29.0	64.2	29.0	+0.3	0



(a) “Paris” encoded at 96.42Kbps

(b) “Bus” encoded at 95.91Kbps

Fig.1. PSNR results for sequence

REFERENCES

[1] J. W. Lee, and Y.S. Ho, “Target Bit Matching for MPEG-2 Video Rate Control”, *Proceedings of Tencon 1998, New Delhi, India, 1998,*vol.1, pp: 66-99.

[2] “MPEG-2 Video Test Model 5”, *ISO/IEC/JTC1/SC29/ WGll*, MPEG93/457, 1993.

[3] J.Ribas-Corbera, and S.Lei, “Rate Control in DCT Video Coding for Low-Delay Communications”, *IEEE Trans. on Circuits and Systems for Video Technology*, 1999, vo1.9 (1), pp: 172-185.

[4] “Video Codec Test Model Near-Term”, TMN8. *ITU-T/SG16/VCEG/Q15 A59, Portland*, USA, 1997.

[5] Zhihai. He, Yong Kwan Kim, and Sanjit K. Mitra, “Low-Delay Rate Control for DCT Video Coding via p-Domain Source Modeling”, *IEEE Transactions on Circuits and Systems for Video Technology*, 2001, vol.11(8), pp:928-940.

[6] T.Chiang, and Y.Q.Zhang,“A New Rate Control Scheme Using Quadratic Rate Distortion Model”, *IEEE Transactions on Circuits and Systems for Video Technology*,1997, vol7(1), pp: 246-250.

[7] Vetro, H.Sun, and Y.Wang, “MPEG-4 rate control for multiple video objects”, *IEEE Transactions on Circuits and Systems for Video Technology*, 1999, vol.9, pp.186-199.

[8] Z.G. Li, Pan F, and Lim K.P, “Adaptive Basic Unit Layer Rate Control for JVT”, *JVT-G012r1, 7th meeting,* Pattaya, Thailand, Mar. 2003.

[9] S.W.Ma, W.Gao, and Y.Lu, “Rate Control on JVT Standard Document”, *JVT_D030*, Klagenfurt, Austria, Jul. 2002.

[10] M.Q. Jiang, X.Q. Yi, and Nam Ling, “Improved Frame-Layer Rate Control for H.264 Using MAD Ratio”, *ISCAS 2004, III*, pp: 813-816.

[11] S.L. Shang, Q.X. Du, and H.Q. Lu, “A Low-Complexity Macroblock Level Rate Control for H.264”, *Chinese Journal of Computers*, 2006, vol.29 (6), pp: 914-919.

Yangchun Li is a lecturer of Xianning College, China. He is currently pursuing his M.S. at Wuhan University of Technology. He received his B.S. degree in computer science and technology from China University of Geosciences in 1996. His current research interests include the computer graphics and image process and communication.

Ruolin Ruan is currently pursuing his Ph.D at Wuhan Universirty. He is a lecturer of Xianning College, China. He received his M.S. degree in computer science and technology from Wuhan University of Technology in 2005. His current research interests include the modern long-distance education, multimedia technology, and video coding and video communication.

Study on Meshing Force of Involute Gear Based on Simulation

Qingbin Cui, Xinmin Huo, Linchun Xing, Renbin Zhou
Wuhan Mechanical Technology College
Wuhan, Hubei Province, China 430075
Email: cqb0430@126.com

ABSTRACT

Considering the variation of meshing point of involute gears in the transmission process, in order to obtain the data of dynamic meshing force of involute gears, computer simulation of involute gears meshing is proposed based on the MSC-ADAMS software. Then, combined with the tooth profile equation modeling of involute gear, the gear meshing force is obtained in the transmission process based on the MATLAB software, while one pair of teeth is form entering to quitting meshing. The results show that the impact forces in the meshing process of involute gears are obviously. Which provide scientific warranty for strength, stiffness, lifetime prediction and optimum design of involute gear.

Keywords: Involute Gear, Meshing Force, Simulation

1. INTRODUCTION

The periodical impact forces in the meshing process of involute gears in the armored vehicle transmission system influence not only the performance of the vehicle transmission system but also failure and design of the gear. So, the study of meshing force is significant to the performance of the vehicle transmission system, strength, stiffness, lifetime prediction and optimum design of involute gear. According to the well-known software ADAMS and MATLAB, the virtual prototyping model of gear meshing is built in this paper. Then, the impact forces of involute gear are achieved by virtual driving, which is foundation of strength, stiffness, lifetime prediction and optimum design of involute gear.

2. CONTACT MODEL OF ADAMS

Meshing force of involute gear is a constant force in the theory of mechanical design, but which is bigger different in the real situation.

The software ADAMS can simulate the real load of the mechanical system by virtual prototyping. The software ADAMS calculates normal contact force by using of penalty function method. Contact model as follow:

$$F_n = k \cdot g^e + step\ (g, 0, 0, D_{\max}, C_{\max})\frac{dg}{dt}$$

Where, F_n —normal contact force

k —penalty parameter

g —the penetration of one geometry into another

e —a positive real value denoting the force exponent

$step$ —step function

$D_{\max}$ —maximum penetration

$C_{\max}$ —maximum damping coefficient

3. SIMULATION MODEL OF INVOLUTE GEAR

The MSC-ADAMS software offers various contact models, such as solid-to-solid, circle-to-circle, curve-to-curve, and so on. In this paper, the contact model is curve-to-curve contact. Figure 1 shows the simulation model of involute gear.

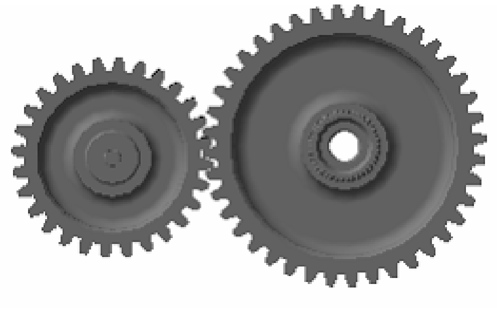


Fig.1. The Simulation Model of Gear Mesh

3.1 Definition of involute gear profile

According to the processing of involute, polar equation of involute gear is described as follow:

$$\theta_K = tg\,\alpha_k - \alpha_K$$

$$r_K = \frac{r_b}{\cos\ \alpha_K}$$

Where, θ_k —expansion angle

α_k —pressure angle

r_k —radius vector

r_b —base radius

By using of polar equation of involute gear, the matrix data of involute gear are obtained based on the MATLAB software. This data are imported to the simulation model which is built by the ADAMS software. Figure 2 shows the profile of driven gear.

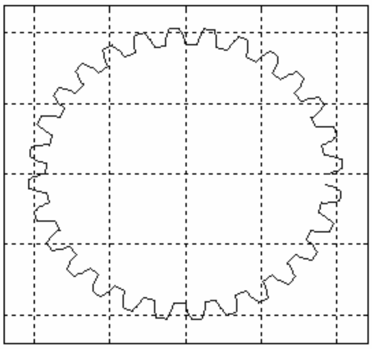


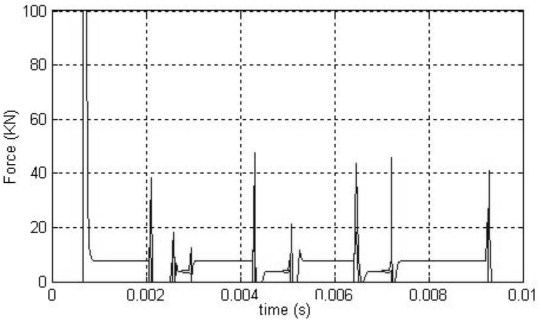
Fig.2. the Profile of Driven Gear

3.2 Definition of load and constraints

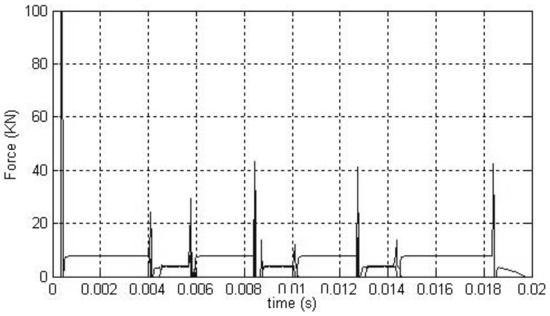
According to the request of contact model, the curve-to-curve contact is defined. When the drive gear wheel is applied load, the driven gear wheel is applied constrain, the simulation model of involute gear is built.

4. ANALYSIS OF SIMULATION AND RESULTS

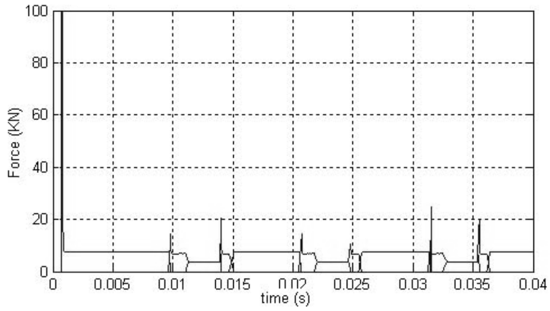
When the drive torque T=1000N/m, the driven constraint ω =104.7rad/s, ω =52.36rad/s, ω =20.94rad/s, the meshing force curve of involute gear are achieved by simulation as followed Figure 3. Where, K=1e+6N/mm, $D_{\max}$ =0.01mm, e=2.2, $C_{\max}$ =1000.



a) ω=104.7rad/s



b) ω =52.36rad/s



c) ω =20.94rad/s

Fig.3. Meshing force curve of gear

The change of meshing force of involute gear is achieved by the simulation model of gear meshing. The results show that the value data of simulation equals the value data of theoretical arithmetic. When the angular velocity of gear is bigger, the impact force is bigger. So, the meshing force which is achieved by simulation is reasonable.

5. DISCUSS BETWEEN MESHING FORCE AND CONTACT POINT

When one pair of teeth is form entering to quitting meshing, the relationship (Figure 4) between meshing force and time history can be known by means of gear mesh's virtual prototyping simulation.

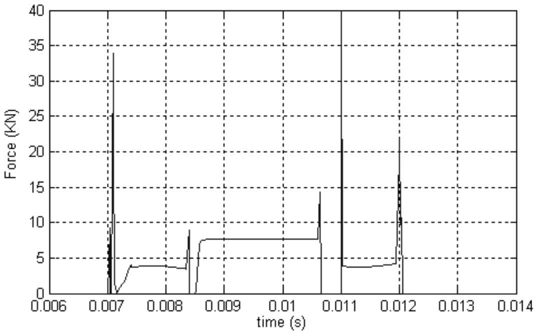


Fig.4. Meshing force variable with time

The relationship (Figure 5) between expansion angle and time history is known in the gear mesh theory.

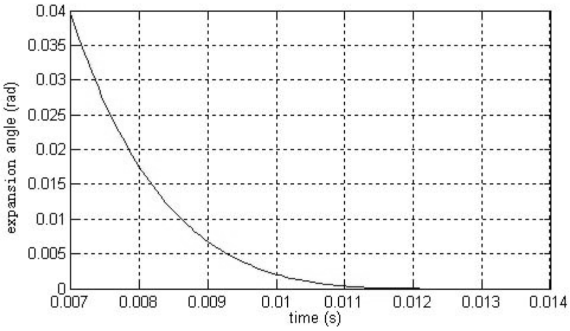


Fig.5. Expansion angle variable with time

Combining Figure 4 and Figure 5, the relationship (Figure 6) between meshing force and contact point can be found by expansion angle.

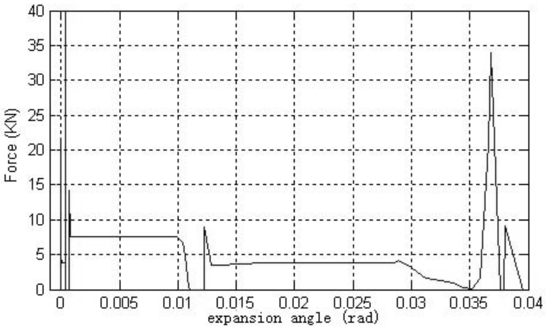


Fig.6. Meshing force variable with expansion angle

6. CONCLUSIONS

(1) Meshing force of involute gear is achieved by virtual simulation.The results show that the angular velocity of gear is bigger, the impact force of meshing force is bigger.

(2) The scientific data are provided for the strength,stiffness, lifetime prediction and optimum design of involute gear by the simulation method.

REFERENCES

[1] Qingbin Cui,Jingzhu Zhang,Guanhai Xue,Shichun Chen, Lei Lei,"Fatigue Life Prediction of Gear Based on Simulation Technology,"*Key Engineering Materials* Vols.324-325(2006),pp.431-434.

[2] Ian Howard,Shengxiang Jia,Jiande Wang,"The Dynamic modeling of a Spur Gear in Mesh Including Friction and a Crack," *Mechanical Systems and Signal Processing*, 2001,15(5),pp.831~851.

[3] B.Ozdalyan, M.V.Blundell. "Anti-lock Braking System Simulation and Modeling in ADAMS," *IEEE* 1998, pp.140~144.

[4] W.R.Kruger, W.Kortum,"Multibody Simulation in the Integrated Design of Semi-active landing Gears," American Institute of Aeronautics, Inc,1998,pp.246~256.

[5] W.B.Ferry,P.R.Frise,G.T.Andrews,M.A.Malik, "Combining Virtual Simulation and Physical Vehicle Test Data to Optimize Durability Testing," Fatigue & Fracture of Engineering Materials & Structures, 2002,25,pp.1127~1134.

Qingbin Cui is a lecturer of Self-propelled Gun department of Wuhan Mechanical Technology College. He graduated from Ordnance Engineering College of PLA in 2001. He has published two books, over 10 Journal papers, and one of them has been indexed by EI. He was born in Qiqihaer city in Heilongjiang province in 1978. His research interests are in dynamic simulation, structure fatigue lifetime prediction.

A New Kind of SVM with Spline Wavelet Kernel

Yafan Yue [1], Dayou Zeng [1], Xufang Li [2]
[1]North China Institute of Aerospace Engineering, Langfang Hebei, 065000, China
[2]School of Information, Jiangnan University, Wuxi Jiangsu, 214122, China
Email: zhangql1972@yahoo.com.cn

ABSTRACT

We propose a kind of admissible support vector kernel called spline wavelet kernel. This is based on the theory of multi-resolution analysis and support vector kernel function. In fact, spline wavelet kernels are the multi-dimensional spline wavelet function with translation vectors and dilation factors and they are a set of complete orthonormal bases in the square integral space. Hence, the support vector machine (SVM) with spline wavelet kernel can approximate almost any objective function in the square integral space theoretically. Thus, the generalization ability of the SVM is improved greatly. The results obtained by our simulations show that the regression's accuracy of SVM with spline wavelet kernel is significantly higher than that of SVM with Gaussian kernel under the same conditions.

Keywords: Support Vector Machine, Spline Wavelet Kernel, Gaussian Kernel

1. INTRODUCTION

SVM is a state-of-the-art learning machine which has been extensively used as a tool for data classification , regression analysis, etc.. SVM has found a great deal of success in many applications (see [1-5]) due to its generalization ability. Unlike traditional methods which minimizing the empirical training error, a noteworthy feature of SVM is that it minimize an upper bound of the generalization error through maximizing the margin between the separating hyperplane and a data set. This can be regarded as an approximate implementation of the Structure Risk Minimization principle which provides a guaranteed bounded risk value even when the number of the training samples are small. What makes SVM attractive is the property of condensing information in the training data and providing a sparse representation by using a very small number of data points (SVs, Support Vectors, see [6]). To facilitate the discussion, we give a very brief review of SVM in this section. More details can be referred to [1-3].
Consider N pairs of training samples:
$\{X(1),Y(1)\}$, $\{X(2),Y(2)\}$,---, $\{X(N),Y(N)\}$, where , $X(i)=[x_1(i),x_2(i),\cdots,x_k(i)]^T$ is a k -dimensional feature vector representing the i th training sample, and $Y(i)\in\{-1,1\}$ is the class label of $X(i)$.

A hyperplane in the feature space can be described as the equation $W\cdot X+b=0$, where $W=[w_1,w_2,\cdots,w_k]^T$ is a weight vector , and b is a scalar. The signed distance $d(i)$ from a point $X(i)$ to the hyperplane in the feature space is $d(i)=\frac{W\cdot X(i)+b}{\|w\|}$. When training samples are linearly separable, SVM yields the optimal hyperplane that separates two classes with no training error, and maximizes the minimum value of $|d(i)|$. It is easy to find that the parameter pair (W,b) corresponding to the optimal hyperplane is the solution to the following optimization problem:

Minimize $L(W)=\frac{1}{2}\|w\|^2$

Subject to

$Y(i)(W\cdot X(i)+b)\geq 1,\ i=1,\cdots,N.$ (1)

For linearly nonseparable cases, there is no such a hyperplane that is able to classify every training point correctly. However, the optimization idea can be generalized by introducing the concept of soft margin. The new optimization problem thus becomes:

Minimize $L(W)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi(i)$

Subject to

$Y(i)(W\cdot X(i)+b)\geq 1-\xi(i),\ \ i=1,\cdots,N.$ (2)

where ξ_i are called slack variables which are related to the soft margin, and C is the tuning parameter used to balance the margin and the training error. Both optimization problems (1) and (2) can be solved by introducing the Lagrange multipliers $\alpha(i)$ in Lagrange function $L(W,b,\alpha)$ that transform them to quadratic programming dual problem.

In the classification phase, a point $\overline{X}$ in the feature space is assigned a label $\overline{Y}$ according to the following equation:

$\overline{Y}=\mathrm{sgn}(W\cdot\overline{X}+b)=\mathrm{sgn}[\sum_{i=1}^{N}\alpha(i)Y(i)(X(i)\cdot\overline{X})+b]$.

SVM is a linear classifier in the parameter space, but it can be easily extended to a nonlinear classifier of the Ψ -machine type (see [7]). For the applications where linear SVM does not produce satisfactory performance, nonlinear SVM is suggested. The basic idea of nonlinear SVM is to map X by a nonlinear mapping $\Psi(X)$ to a much higher dimensional space, in which the optimal hyperplane is found. By choosing an adequate mapping Ψ , the data points become linearly separable or mostly linearly separable in the high-dimensional space, so that one can easily apply the structure risk minimization. Based on the observation that only the dot products of two mapped patterns are needed to solve the corresponding quadratic programming problem, we need not compute the mapped patterns $\Psi(X)$ explicitly and can define the nonlinear mapping implicitly by introducing the so-called kernel function which computes the inner product of vectors $\Psi(X(i))$ and $\Psi(X(j))$. Actually, in the SVM study, people always work in the inverse way by starting from a kernel. Among a variety of kernel functions available, the radial basis function (RBF) and the polynomial function are often chosen for many applications. SVM can realize RBF , Polynomial and Multi-layer Perceptron classifiers.

(1) RBF: $K(X_1,X_2)=\exp(-\frac{\|X_1-X_2\|}{\sigma^2})$ where σ is the parameter controlling the width of the kernel.

(2) Polynomial Function

$K(X_1,X_2)=(X_1\cdot X_2+1)^d$, where d is the degree of the polynomial.

Accordingly, the class label of an unseen point $\overline{X}$ is given by

$\overline{Y} = \text{sgn}(\overline{W} \cdot \overline{X} + b) = \text{sgn}[\sum_{i=1}^{N} \alpha(i)Y(i)K(X(i), \overline{X}) + b]$. For pattern recognition and regression analysis, the non-linear ability of SVM can use kernel mapping to achieve. For the kernel mapping, the kernel function must satisfy Mercer Condition (see [7]). The Gaussian kernel is extensively used with good generalization performance. However, SVM with Gaussian kernel, Polynomial kernel and Sigmoid kernel can not approach any curve in $L^2(R)$ (square integral space), because it can not form a set of complete orthonormal bases in $L^2(R)$. Therefore, we need find a new kernel function which can build a set of complete bases in $L^2(R)$ by translation and dilation. We knew scaling and wavelet function could build a set of complete bases by translation and dilation[8]. In this paper, we prove spline wavelet function is a new support vector kernel function which is named as spline wavelet kernel. The spline wavelet kernel can construct a set of orthonormal bases in $L^2(R)$ by translation and dilation. Thus, SVM with spline wavelet kernel can approximate any function in $L^2(R)$. The experimental results support our ideas.

2. SPLINE WAVELET KERNELS

2.1 Support Vector Kernels

If a function satisfied Mercer condition, it is the allowable support vector kernel.

Theorem 1 [7] (Mercer condition) The symmetry function $K(X, X')$ is the dot products in a feature space if and only if

$$\iint_{R^d \otimes R^d} K(X, X')g(X)g(X')dXdX' \geq 0$$

always holds for all $g \in L^2(R^d)$

We can construct kernel function using this theorem. Reference [8] gives another theorem for the translation-invariant kernel function.

Theorem 2 $K(X - X')$ is a allowable support vector's kernel function if and only if the Fourier transform of $K(X)$ satisfies

$$F[K(\omega)] = (2\pi)^{-\frac{d}{2}} \int_{R^d} \exp(-j\omega X)K(X)dX \geq 0 \cdot$$

2.2 Spline Wavelet Kernel

In this paper, we select two order spline wavelet with many good properties and denote by $\psi_2(x)$ it with one dimension, where

$$\psi_2(x) = \begin{cases} -\frac{1}{6}|x| + \frac{1}{4} & 1 < |x| \leq 1.5 \\ \frac{7}{6}|x| - \frac{13}{12} & 0.5 < |x| \leq 1 \\ -\frac{8}{3}|x| + \frac{5}{6} & |x| \leq 0.5 \\ 0 & elsewhere \end{cases}$$

More details can be referred to [9].

Using tensor theory (see [l0]), we can obtain multi-dimensional spline wavelet function as follows:

$$\Psi_d(X) = \prod_{i=1}^{d} \psi_2(x_i)$$

We define the finite element multi-scale function kernel function as follows:

$K(X, X') = K(X - X') = \prod_{i=1}^{d} \psi_2(\frac{x_i - x_i'}{a_i})$ where a_i is dilation coefficient and $a_i > 0$.

Theorem 3 $K(X, X')$ is a allowable support vector kernel function.

Proof According to the theorem 2, we only need to prove

$$F[K(\omega)] = (2\pi)^{-\frac{d}{2}} \int_{R^d} \exp(-j\omega X)K(X)dX \geq 0 \cdot$$

$$F[K(\omega)] = (2\pi)^{-\frac{d}{2}} \int_{R^d} \exp(-j\omega X)K(X)dX$$

$$= (2\pi)^{-\frac{d}{2}} \prod_{i=1}^{d} a_i \int_{-\infty}^{+\infty} \psi_2(\frac{x_i}{a_i}) \exp(-j\omega a_i \frac{x_i}{a_i}) d(\frac{x_i}{a_i})$$

$$= (2\pi)^{-\frac{d}{2}} \prod_{i=1}^{d} \frac{4}{3a_i\omega^2}[\cos(\frac{a_i\omega}{2}) - 1]^2[2 - \cos(\frac{a_i\omega}{2})] \geq 0$$

3. EXPERIMENTAL STUDIES

To evaluate the performance of our method, we did simulations on two regression problems. one function of one-variable and the other function of two variables. The parameters a_i are fixed to be one.

For these two function's regression, we use the approaching error of [10].

3.1 One-Variable Function Identification

We select the following function:

$y = \cos(\exp(x))$

In this example, we sampled 1000 points distributed uniformly over $[0,1]$,we used 500 for training and 500 for testing.

Simulation results for Gaussian kernel and multi-scale kernel are compared in the table1.

Table.1 Comparison about the results for multi-scale kernel and Gaussian kernel

Kernel function	Regression error	Number of support vectors
Spline wavelet kernel	0.0397	107
Gaussian kernel	0.0637	133

Obviously, the performance of SVM with multi-scale kernel is superior to that of SVM with Gaussian kernel with low support vectors and good accurate.

3.2 Two Variables Function Identification

The two variables function is as follows:

$z = \cos(x)\exp(x^2 + y^2)$

We sampled 400 points distributed uniformly over $[0,1] \times [0,1]$ and used 200 for training, 200 for testing.

Simulation results for Gaussian kernel and multi-scale kernel are compared in the table 2.

Table 2 Comparison about the results for multi-scale kernel and Gaussian kernel

Kernel function	Regression error	Number of support vectors
Spline wavelet kernel	0.0174	53
Gaussian kernel	0.0256	76

4. CONCLUSIONS

In this paper, we construct a new SVM based on spline wavelet for function approximation. The presented SVM with spline wavelet kernel integrates wavelet theory with SVM concepts such that spline wavelet kernel SVM possesses good sparse property. Compared with Gaussian kernel SVM, the proposed spline wavelet SVM not only reserves the multi-resolution capability of wavelet analysis, but also has the advantages of the approximation accuracy and good generalization performance. However, it need to deeply research for high dimension system identification.

REFERENCES

[1] C.Cortes, V.N. Vapnik, “Support vector networks,” *Machine Learning*, Vol. 20, No. 3 ,1995,pp.73-297.

[2] V.N.Vapnik,“An overview of statistical learning theory,” *IEEE Trans. Neural Networks*, Vol.10,No.5,1999,pp.988-989.

[3] O.Chapelle,V.N.Vapnik,Y. Bengio,“Model selection for small sample regression,”*Machine Learning*,Vol.48,No.1,2002,pp.9-23.

[4] Y.Liu,Y.F. Zheng,FS_SFS, “A novel feature selection method for support vector machines,” *Pattern Recognition*,Vol.39,2006,pp.1333-1345.

[5] N.Acir,“A support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recongnition systems,”*Expert Systems with Application*,Vol.31,2006,pp.150-158.

[6] Girosi,F,“An equivalence between sparse approximation and support vector machine,” *Neural Computation*,Vol.20,1998,pp.1455-1480.

[7] John Shawe_Taylor,Nello Cristianini, *Kernel Methods for Pattern Analysis*,Cambridge:Cambridge University press,2004.

[8] J.A.K.Suykens,J.Vandewalle,“Least squares support vector machine classifiers,”*Neural Processing Letter*,Vol. 9,No.3,1999,pp.293-300.

[9] Chui K,*An Introduction to Wavelet,* San Diego: Academic Press,1992.

[10] Q.Zhang,A.Benveniste,“Wavelet networks,” I*EEE Trans on Neural Networks*,Vol.3,No.6,1992,pp.889-898.

Yafan Yue is a lecture and a teacher of North China Institute of Aerospace Engineering. He received the B.S. degree from Northeast Normal University in 1994, the M.S. degree from Hebei Normal University in 2007. His research interests are in Mathematics and artificial intelligence.

Dayou Zeng is a associate professor and a teacher of North China Institute of Aerospace Engineering. He graduated from Sichuan University in 1986 with specialty of Math. His research interests are in Mathematics and artificial intelligence.

Xufang Li is currently working toward M.S. degree from School of Information, Jiangnan University. He graduated from Hebei Institute of Architecture Civil Engineering in 2005 with specialty of Computer. Her research interests are in machine learning and artificial intelligence.

A Fast algorithm of GLCM Computation Based on Programmable Graphics Hardware

Zhipeng Xu, Hongyan Liu
School of Physics Science and Information Engineering, Liaocheng University
Liaocheng, Shandong, 252059, China
Email: xu0510cn@yahoo.com.cn

ABSTRACT

It is well-known that GLCM is the effective method for segmenting texture images. GLCM is widely used in image analysis, image segmentation, etc. But because the essence of GLCM is a problem of second-order statistical computation, so the calculation of GLCM is very expensive. In this paper, a new method is presented to calculate GLCM. It is based on Programmable Graphics Hardware. The results show that the time cost of computation is greatly reduced.

Keywords: Texture, Grey Level Co-occurrence Matrix(GLCM), GPU computing.

1. INTRODUCTION

Texture almost presents everywhere in natural and real world images. Texture, therefore, has long been an important research topic in image processing. Successful applications of texture analysis methods have been widely found in industrial, medical and remote sensing areas. Grey Level Co-occurrence Matrix (GLCM), one of the best known tools for texture analysis, estimates image properties related to second-order statistics.

However, their computations are highly intensive especially for very large images such as medical ones. For an image of size 5000 ＊ 5000 pixels with 16 bands, the time required is approximately 260 seconds using Pentium 4 machine running at 2400 MHz. There are different images for each patient and even the number of patients can vary. For example, if there are 100 patients with 10 images for each patient, the overall time required is approximately 100 hours which is quite computational intensive [1].

The traditional Von Neumann style of fetch-operate-writeback computation can not realize the inherent parallelism in computing GLCM and Haralick features. Therefore, acceleration of computation of GPU has been paid much attention. Methods based on hardware accelerated are proposed.

The aim of this paper is to investigate the use of graphics processing unit (GPU) to accelerate the computation of GLCM. GPU was originally developed for 3D games; however, GPU has become so powerful and fast that they are expected to be powerful tools in the fields of floating point calculations. Employing graphics hardware for purposes it was not designed for has a long tradition. Today, it has been utilized in many fields, such as artificial neural networks [2], fast matrix multiplies [3], 3d convolution [4], morphological analysis [5], ray tracing [6], robot motion planning [7], sparse matrix solvers [8].

The paper is organized as follows. Section 2 reviews grey level co-occurrence matrix. Section 3 describes the proposed system model. Section 4 describes graphics hardware implementation. Experiments and discussion are presented in Section 5.

2. GREY LEVEL CO-OCCURRENCE MATRIX (GLCM)

GLCM, first introduced by Haralick[9], is a powerful technique for measuring texture features. Suppose an image to be analyzed is rectangular and has Nx columns and Ny rows. Suppose that the grey level appearing at each pixel is quantized to L levels. Let Lx = 1, 2, . . . , Nx be the columns, Ly = 1, 2, . . . , Ny be the rows, and G = 0, 1, 2, . . . , Ng be the set of Ng quantized grey levels. The texture-context information is specified by the matrix of relative frequencies Pi, j with two neighboring pixels separated by displacement d and angle θ.

The GLCM is calculated with the following equation[9]:

$$P(i, j, d, \theta) = \# \{(x_1, y_1)(x_2, y_2) | f(x_1, y_1) = i, \\ f(x_2, y_2) = j, |(x_1, y_1) - (x_2, y_2)| = d, \\ \angle((x_1, y_1)(x_2, y_2)) = \theta \}$$

where # is the number of occurrences inside the window sizes where the intensity level of a pixel pair changes from i to j , the location of the first pixel is (x1, y1) and that of the second pixel is (x2, y2), d is the distance between the pixel pair, θ is the angle between the two pixels. The co-occurrence matrix defined is not symmetric.

3. PROPOSED SYSTEM ARCHITECTURE

The block diagram for calculating GLCM on GPU is shown in Fig. 1. The image must be grayed and quantized to L grey levels. Since the floating point buffer has been adopted, the grey values are expressed in integer value instead of floating point value between [0,1]. The image is divided into regions R of size N ＊ N.

Each element in GLCM concerns two pixels, which is not convenient for calculation. So we propose a new method to represent the pixel pair. An index of position in GLCM replaces the pixel pair in image. For the pixel pair (i,j), the corresponding index is i * L + j. Then the Region R is converted to texture T1. The texel value in T1 varies form 0 to L*L-1. When we count the number of certain value in T1, we can get the corresponding element in GLCM. The algorithm can be realized by recursively summing up the result of the previous pass in multiple rendering passes[10]: starting with the initial texture T1 and the quadrilateral lined up with the texture in screen space; in each step a quadrilateral scaled by a factor of 0.5 is rendered, in the shader program, each fragment that is covered by the shrunken quadrilateral sums up the texel that is mapped to it and the three adjacent texel in positive (u,v) texture space direction. The entire process is illustrated in figure 2[10]. Several methods are implemented to improve the parallelism of computation. Firstly, each pixel has RGBA channels, while the pixel in region R only has one gray value.

Four neighboring pixels can be combined to 1 pixel. The process can be realized by drawing half size of original region, in fragment shader, the current pixel and the neighboring 3 pixels were sampled and written back into the RGBA of current pixel. So we get the texture T2. With reduction process, finally 4 GLCM values can be obtained in one single pixel. Secondly, a big texture T3 is constructed. In the prostate cancer research, experiments have shown that the size of region equals to 128X128 exhibits the best trade off in terms of good localization and accurate measurements of texture features for the application. The size of corresponding texture T2 of such region will be 64X64.

Obviously the size can not exert all GPU’s power. So we propose a new method to compute the GLCM with texture T3. Many textures T2 are combined to construct a big texture T3. For each texture T2 in texture T3, 4 elements of GLCM will be computed. Different elements of GLCM will be calculated according to the position of texture T2. If the gray level is 32, then the size of GLCM will be 32X32. The total number of texture T2 is (32X32)/4=16X16=256. The size of texture will be 16X64=1024.

Usually the maximum texture dimensions are 4096X4096. If such texture T3 is utilized to calculate GLCM, the maximum region that can be calculated with texture T3 will be 512X512. The configuration of texture T3 for region R 128X128 is shown in fig.3. The size of every texture block T2 is 64X64 after size reduction, and 4 elements of GLCM are calculated from each texture T2.

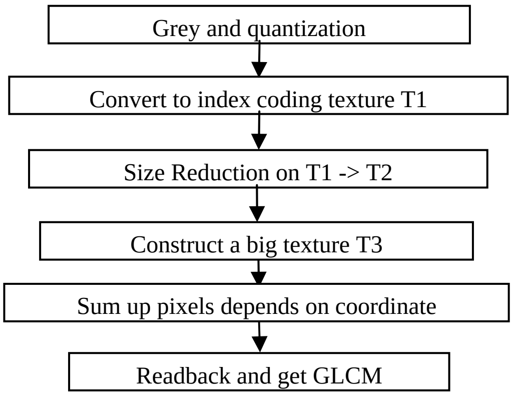

Grey and quantization
Convert to index coding texture T1
Size Reduction on T1 -> T2
Construct a big texture T3
Sum up pixels depends on coordinate
Readback and get GLCM

Fig.1. Block diagram for calculating GLCM on GPU.

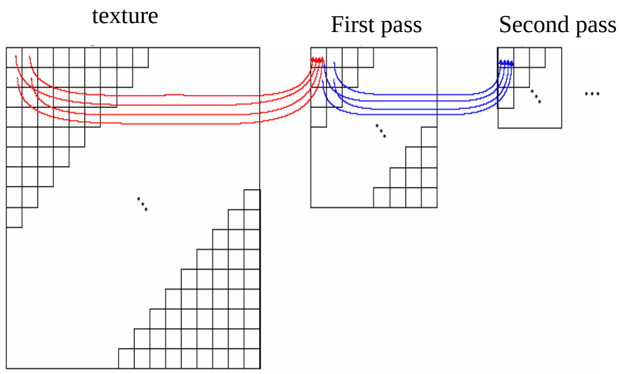


Fig.2 count number of certain value in T1

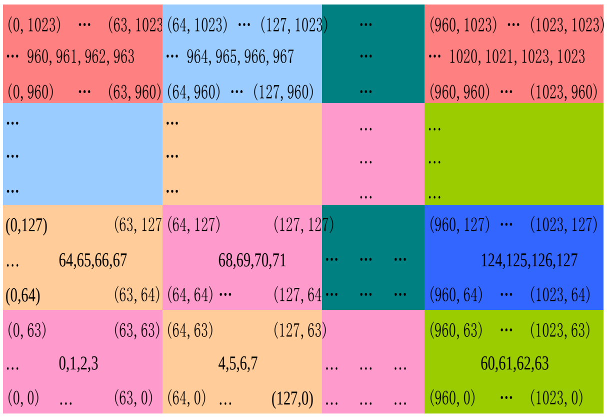


Fig.3 The configuration of texture T3，R=128X128，L=32

4. GRAPHICS HARDWARE IMPLEMENTATION

The basic idea of realizing above algorithm with graphics hardware is: firstly, the image is expressed as textures or vertices, then the main part of the algorithm is implemented in a fragment shader, thirdly a quadrilateral of same dimensions with the texture is drawn, finally we get the result by reading back the pixel buffer. In order to assuring the exact corresponding map of the texels in the texture and pixels on the screen, we use an orthographic view (no perspective adjustments). We also use the a few floating point buffers to avoid the bias and scale of middle result.

The image must be grayed and quantized to 32 grey levels. The grey values are expressed in integer value instead of floating point value between [0,1]. For each pixel i, another pixel j is determined according to the displacement vector d. Then the position index of pixel pair in GLCM is calculated: index = i * L + j, the index is written back into pixel i. The displacement vector d and gray level L can be transferred to shader with parameters. The texture T3 is constructed by glCopyTexSubImage2D.

The complex part of algorithm is summing up according to coordinate position. Two steps are included: firstly, the 4 elements s,s+1,s+2,s+3 of GLCM for each texture T2 is calculated. Suppose the size of texture T2 is PXP, for one pixel whose coordinate is (x,y) in texture T3, let m1=floor(x/P), m2=floor(y/P), then the first element s=m2*64+m1*4, the number 64 is the total results of horizontal 16 textures blocks T2. The number 4 is results of each texture T2. Secondly, the sum operation is implemented through size reduction of texture T3, in fragment shader, the current pixel and adjacent 3 pixels are sampled and compared with s,s+1,s+2,s+3, the number 1 or 0 is calculated and count of certain value is summed up, finally the sum is written back into current pixel. When a square quadrilateral of 16X16 is drawn, the complete GLCM is on the pixel buffer. Some code is shown in fig.4.

5. EXPERIMENTS AND DISCUSS

The algorithm is tested with a common PC which has a P4 920 2.8Ghz CPU, 1G RAM and a nVidia Geforce 7900GT. The original image size is 512X512. Different size of region R is tested including 256x256,128X128,64x64,32x32. The displacement vectors are (2,0) The results are shown in table 1.

We now compare our GLCM run on a GPU with a GLCM run on a traditional CPU. We performed a GLCM of region R=128X128 in image 512X512. We found this operation to take about 13.9 milliseconds. We feel that our GPU implementation, which took about 8.5 seconds, is comparable to the GLCM on a traditional CPU. When GLCM of region 256X256 Table 1 is performed, the results on CPU are better than on GPU. We believe that cache of CPU influences the result. The L2 cache of P4 920 is 2MB, while the image data is about 256KB. Definitely there are many reading and writing instructions taking place in cache. For very large medical image, the image data can not be fully loaded into cache, the cost time will increased. At the same time, some experiments have shown that the size of region equals to 128X128 exhibits the best trade off in terms of good localization and accurate measurements of texture features for the application[10] GPU.

```
#calculation of 4 elements according to texel coordinate
MOV R6,{0.5,0.5,0,0};
SUB R5,f[TEX0],R6;          # remove offset of texel coordinate
MUL R4,R5,{4.0,4.0,1.0,1.0};
RCP R1.x,p[0].x;
MUL R1.x,R1.x,2.0;          # 1/P＝2/L
MUL R2,R4,R1.xxxy;          # m1,m2,put into R2.x,R2.y
FLR R2,R2;

MUL R2.y,R2.y,64.0;         # m2*64

MAD R2.x,R2.x,4.0,R2.y;     #s=m2*64+m1*4

ADD R2,R2.xxxx,{0.0,1.0,2.0,3.0};
                            #get 4 elements s,s+1,s+2,s+3
#calculation of adjacent pixel coordinates,
MOV R1,{1.0,2.0,3.0,0.0};
MAD R17,R5.xxxy,{2.0,2.0,2.0,2.0},R1;
ADD R17,R17,R6;               #add offset of texel coordinate

……                           #other adjacent pixel coordinates
#sample the texture and count the numbers of 4 elements, the
result is put into
#RGBA of current pixel
TEX R1,R4,TEX0,RECT;        #sample texture of current pixel
SEQ R3,R2.xxxx,R1;
DP4 R3,R3,{1.0,1.0,1.0,1.0};  #how many R2.x are there in
                                R1
ADD R0.x,R0.x,R3.x ;
SEQ R3,R2.yyyy,R1;
DP4 R3,R3,{1.0,1.0,1.0,1.0};  #how many R2.y are there in
                                R1
ADD R0.y,R0.y,R3.x ;
SEQ R3,R2.zzzz,R1;
DP4 R3,R3,{1.0,1.0,1.0,1.0};  #how many R2.z are there in
                                R1
ADD R0.z,R0.z,R3.x ;
SEQ R3,R2.wwww,R1;
DP4 R3,R3,{1.0,1.0,1.0,1.0};  #how many R2.z are there in
                                R1
ADD R0.w,R0.w,R3.x ;
……                           #comparison   of adjacent pixels
MOV o[COLR],R0;
```

Fig.4 calculation of 4 elements for each texture block T2

Table 1 Cost of GLCM computation

	Cost of GLCM computation （ms）	
Size of R	Geforce 7900	CPU
256X256	21.5	14.9
128X128	8.5	13.9
64X64	4.5	13.7

We feel that GPU has superior speed performance when compared with a general-purpose processor for the computation of GLCM.

REFERENCES

[1] M. A. Tahir, A. Bouridane and F. Kurugollu, “An FPGA Based Coprocessor for the Classification of Tissue Patterns in Prostatic Cancer,” Lecture Notes in *Computer Science*, pp771-780, published by Springer Verlag, 2004, ISBN 3-540-22989-2, New York (USA).

[2] C.-A. Bohn. “Kohonen feature mapping through graphics hardware,” In Proceedings of *Int. Conf. on Compu. Intelligence and Neurosciences* 1998.

[3] E. S. Larsen and D. McAllister. “Fast matrix multiplies using graphics hardware,” *The International Conference for High Performance Computing and Communications*, 2001.

[4] M. Hopf and T. Ertl. “Accelerating 3d convolution using graphics hardware,” In *IEEE Visualization* ‘99, pages 471﹣474, October 1999.

[5] M. Hopf and T. Ertl. “Accelerating morphological analysis with graphics hardware,” In Workshop on *Vision, Modelling ,and Visualization VMV’* 00, pages 337﹣345, 2000.

[6] T. Purcell, I. Buck, W. Mark, and P.Hanrahan. “Ray tracing on programmable graphics hardware,” In *Proceeding of SIGGRAPH 2002*. ACM, 2002.

[7] J. Lengyel, M. Reichert, B. Donald, and D. Greenberg. “Real-time robot motion planning using rasterizing computer graphics hardware,” In *Proceedings of SIGGRAPH 1990*, pages 327–335, 1990.

[8] Jeffrey Bolz, Ian Farmer,Eitan Grinspun, and Peter Schr¨oder. “Sparse matrix solvers on the gpu: Conjugate gradients and multigrid,” In Jessica Hodgins, editor, Siggraph 2003, *Computer Graphics Proceedings*, 2003.

[9] R.M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification.” IEEE Trans. on Systems,Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610–621,1973.

[10] Jens Kruger, Rudiger Westermann, “Linear Algebra Operators for GPU Implementation of Numerical Algorithms,” *SIGGRAPH 2003*.

A Color Image Quantization Algorithm Based on Quantum-behaved Particle Swarm Optimization

Xiaohong Wang[1], Wenbo Xu[2]
[1]Dept. of Computer Science and Engineering, Wuxi Radio And TV University,wuxi 214021
[2]School of Information Technology, Jiangnan University,wuxi 214122
Email: wangxh@wxtvu.cn, xwb@sytu.edu.cn

ABSTRACT

A color image quantization algorithm based on Quantum-Behaved Particle Swarm Optimization (QPSO) is developed in this paper. The proposed algorithm randomly initializes each particle in the swarm to contain K centroids (i.e. color triplets). The K-means clustering algorithm is then applied to each particle at a user-specified probability to refine the chosen centroids. Each pixel is then assigned to the cluster with the closest centroid. The QPSO is then applied to refine the centroids obtained from the K-means algorithm. The proposed algorithm is then applied to commonly used images. It is shown from the conducted experiments that the proposed algorithm generally results in a significant improvement of image quality compared to Particle Swarm Optimization (PSO) approaches.

Keywords: Color image quantization, K-means clustering algorithm, Quantum-behaved particle swarm optimization

1. INTRODUCTION

Color image quantization is the process of reducing the number of colors presented in a digital color image [1]. It is an important problem in the fields of image processing and computer graphics. Color image quantization consists of two major steps:

- Creating a colormap (or palette) where a small set of colors (typically 8-256) is chosen from the (224) possible combinations of red, green and blue (RGB).
- Mapping each color pixel in the color image to one of the colors in the colormap.

Therefore, the main objective of color image quantization is to map the set of colors in the original color image to a much smaller set of colors in the quantized image. Furthermore, this mapping, as already mentioned, should minimize the difference between the original and the quantized images. The color quantization problem is known to be NP-complete. This means that it is not feasible to find the global optimal solution because this will require a prohibitive amount of time. To address this problem, several approximation techniques have been used. One popular approximation method is the use of a standard local search strategy such as K-means. K-means has already been applied to the color image quantization problem [2]. However, K-means is a greedy algorithm which depends on the initial conditions, which may cause the algorithm to converge to suboptimal solutions. This drawback is magnified by the fact that the distribution of local optima is expected to be broad in the color image quantization problem due to the three dimensional color space. In addition, this local optimality is expected to affect the visual image quality. The local optimality issue can be addressed by using stochastic optimization schemes.

In this paper, a new color image quantization algorithm based on Quantum-Behaved Particle Swarm Optimization (QPSO) is proposed. As we all know, Particle Swarm Optimization (PSO) is a population-based stochastic optimization algorithm modeled after the simulation of the social behavior of bird flocks and follows similar steps as evolutionary algorithms to find near-optimal solutions. QPSO is the newest improvement version of PSO. QPSO and other evolutionary algorithms that depend on heuristics to find 'soft' solutions are considered to be soft computing algorithms. This population-based search approach reduces the effect of the initial conditions, compared to K-means (especially if the size of the population is relatively large). The feasibility of the approach is demonstrated by applying it to commonly used images. The results show that, in general, the proposed approach performs better than state-of-the-art color image quantization approaches.

2. QUANTUM-BEHAVED PARTICLE SWARM OPTIMIZATION

In the Standard PSO model, each individual is treated as a volume-less particle in the D-dimensional space, with the position vector and velocity vector of the ith particle represented as $X_i(t) = (x_{i1}(t), x_{i2}(t), \dots, x_{iD}(t))$ and $V_i(t) = (v_{i1}(t), v_{i2}(t), \dots, v_{iD}(t))$ .The particles move according to the following equation:

$$v_{id}(t+1) = w * v_{id}(t) + \varphi_1(P_{id} - x_{id}(t)) + \varphi_2(P_{gd} - x_{id}(t)) \qquad (2.1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \qquad (2.2)$$

where φ_1 and φ_2 are random numbers whose upper limits are parameters of the algorithm that have to be selected carefully. Parameter w is the inertia weight introduced to accelerate the convergence speed of the PSO. Vector $P_i = (P_{i1}, P_{i2}, \dots, P_{iD})$ is the best previous position (the position giving the best fitness value) of particle i called pbest, and vector $P_g = (P_{g1}, P_{g2}, \dots, P_{gD})$ is the position of the best particle among all the particles in the population and called gbest.

In essence, the traditional model of PSO system is of linear system, if pbest and gbest are fixed as well as all random numbers are considered constant. Trajectory analysis [3] shows that, whatever model is employed in the PSO algorithm, each particle in the PSO system converges to its Local Point (LP) $p = (p_1, p_2, \dots, p_D)$, one and only local attractor of each particle, of which the coordinates are

$$p_d = (\varphi_1 P_{id} + \varphi_2 P_{gd}) / (\varphi_1 + \varphi_2) \qquad (2.3)$$

so that the pbests of all particles will converges to an exclusive gbest with t → ∞ .

In the quantum model of a PSO, the state of a particle is depicted by wavefunction $\Psi(\vec{x},t)$, instead of position and velocity. The dynamic behavior of the particle is widely divergent from that of the particle in traditional PSO systems in that the exact values of X and V cannot be determined simultaneously. We can only learn the probability of the particle's appearing in position X from probability density function $\left|\Psi(X,t)\right|^2$, the form of which depends on the potential field the particle lies in.

In literature [4], Delta potential well with the canter on point $p=(p_1,p_2,\ldots,p_D)$ is employed to constrain the quantum particles in PSO in order that the particle can converge to their local p without explosion.

In Quantum-behaved Particle Swarm Optimization (QPSO), the particle moves according to the following equation:

$$mbest=\frac{1}{M}\sum_{i=1}^{M}P_i=\left(\frac{1}{M}\sum_{i=1}^{M}P_{i1},\frac{1}{M}\sum_{i=1}^{M}P_{i2},\ldots,\frac{1}{M}\sum_{i=1}^{M}P_{iD}\right) \tag{2.4}$$

$$p_{id}=\varphi * P_{id}+(1-\varphi)*P_{gd} \qquad \varphi=rand() \tag{2.5}$$

$$X_{id}=p_{id}\pm\alpha*\left|mbest_d-X_{id}\right|*\ln(1/u) \qquad \varphi=rand() \tag{2.6}$$

where mbest is the mean best position among the particles. p_{id}, a stochastic point between P_{id} and P_{gd}, is the local attractor on the dth dimension of the ith particle, φ is a random number distributed uniformly on [0, 1], u is another uniformly-distributed random number on [0, 1] and α is a parameter of QPSO that is called Contraction-Expansion Coefficient.

3. THE QPSO-BASED COLOR IMAGE QUANTIZATION (QPSO-CIQ) ALGORITHM

Define the following symbols:

- Np denotes the number of image pixels
- K denotes the number of clusters (i.e. colors in the colormap)
- Zp denotes the coordinates of pixel p
- Mk denotes the centroid of cluster k (representing one color triple in the colormap)

3.1 Measure of Quality

The most general measure of performance is the mean square error (MSE) of the quantized image using a specific colormap. The MSE was defined in Eq. (3.1):

$$MSE=\frac{\sum_{k=1}^{K}\sum_{\forall z_p\in C_k}(Z_p-M_k)^2}{N_p} \tag{3.1}$$

where Ck is the kth cluster.

3.2 The QPSO-CIQ Algorithm

In the context of color image quantization, a single particle represents a colormap (i.e. a particle consists of K cluster centroids representing RGB color triplets). The RGB coordinates in each color triple are floating-point numbers. Each particle X_i is constructed as $X_i=(M_{i1},M_{i2},\ldots,M_{iK})$ where M_{ik} refers to the kth cluster centroid vector of the ith particle. Therefore, a swarm represents a number of candidate colormaps. The quality of each particle is measured using the MSE (defined in Eq. 3.1) as follows:

$$f(X_i)=MSE(X_i) \tag{3.2}$$

The algorithm initializes each particle randomly from the color image to contain K centroids (i.e. color triplets). The set of K color triplets represents the colormap. The K-means clustering algorithm is then applied to each particle at a user-specified probability, p_{kmeans}. The K-means algorithm is used in order to refine the chosen colors and to reduce the search space. Each pixel is then assigned to the cluster with the closest centroid. The fitness function of each particle is calculated using Eq. 3.2. The QPSO update Eq.'s 2.4-2.6 are then applied. The procedure is repeated until a stopping criterion is satisfied. The colormap of the global best particle after tmax iterations is chosen as the optimal result.

The QPSO-CIQ algorithm is summarized below:

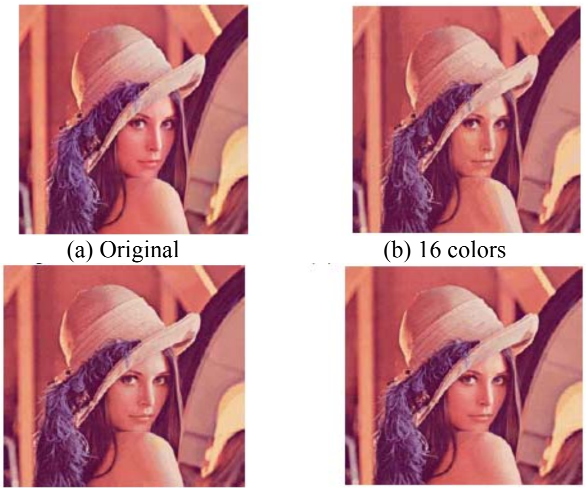
```
1. Initialize each particle by randomly choosing K color triplets from the image.
2. For t = 1 to tmax
   (a) For each particle i
      i. Apply K-means for a few iterations with a probability $p_{kmeans}$
      ii. For each pixel Zp
         Calculate $d^2(Z_p-m_{ik})$ for all clusters $C_{ik}$ .
         Assign Zp to $C_{ih}$ where $d^2(Z_p-m_{ih})=\min_{\forall k=1,\ldots K}\left\{d^2(Z_p-m_{ik})\right\}$
      iii. Calculate the fitness, $f(X_i)$
   (b) Find the global best solution $\hat{Y}(t)$
   (c) Update the centroids using Eq.'s 2.4-2.6
```

4. EXPERIMENTAL RESULTS

The QPSO-CIQ algorithm was applied to a commonly used color images namely: Lenna (shown in Figure 1(a)). The size of each image is 512 × 512 pixels. All images are quantized to 16, 32 and 64 colors. The QPSO-CIQ parameters were initially set as follows: pkmeans = 0.1, s = 20, tmax = 50, number of K-means iterations is 10.The parameter α of QPSO decreases linearly from 1 to 0.5. Table 1 summarizes the results for the four images. Figure 1 shows the visual quality of the quantized image generated by QPSO-CIQ when applied to Lenna. The results showed that the color image quantization algorithm based on Quantum-Behaved Particle Swarm Optimization (QPSO) is effective and feasible.

Table 1. Results for the Lenna image

K	MSE
16	208.147
32	115.422
64	71.846



(a) Original (b) 16 colors

(c) 32 colors (d) 64 colors

Fig.1. Quantization results for the Lenna image using QPSO-CIQ

5. CONCLUSIONS

This paper presented a QPSO-based color image quantization algorithm (QPSO-CIQ). The QPSO-CIQ uses the K-means clustering algorithm to refine the color triplets. Future research can investigate the use of other more efficient clustering algorithms such as FCM and KHM [5]. Finally, the QPSO-CIQ uses the RGB color space. Although the RGB model is the most widely used model, it has some weaknesses. One of these weaknesses is that equal distances in the RGB color space may not correspond to equal distance in color perception. Hence, future research may try to apply the QPSO-CIQ to other color spaces (e.g. the L*u*v* color space [6]).

REFERENCES

[1] Braquelaire J, Brun L (1997) “Comparison and optimization of methods of color image quantization,” *IEEE Transactions on Image Processing* 6(7): 1048-1052.

[2] Celenk M (1990) “A color clustering technique for image segmentation,” *computer vision, Graphics and Image Processing* 52: 145-170.

[3] Clerc, M., Kennedy, J.: “The Particle Swarm: Explosion, Stability and Convergence in a Multi-Dimensional Complex Space,” *IEEE Transaction on Evolutionary Computation*. 6 (2002) 58 – 73.

[4] Sun, J., Feng, B., Xu, W. “Particle Swarm Optimization with Particles Having Quantum Behavior,” *IEEE Proc. of Congress on Evolutionary Computation*. (2004)

[5] Zhang B (2000) “Generalized K-Harmonic means -boosting in unsupervised learning,” *Technical Report HPL*-2000-137, Hewlett-Packard Labs.

[6] Watt A (1989) *Three-Dimensional Computer Graphics*, Addison-Wesley.

Reconstruction of the Objects with Fractal Features in the Virtual Geographic Environment*

Dan Liu[1], Yun Han[2], Daguo Chan[3]
[1]Department of Forensic Science and Technology, China Criminal Police University
Shenyang, Liaoning 110035, P.R.China
[2]Department of Foreign Language, Dalian University of Technology
Dalian, Liaoning 116024, P.R.China
[3]Department of Computer Science and Technology, China Criminal Police University
Shenyang, Liaoning 110035, P.R.China
Email: dliudlmu@gmail.com

ABSTRACT

On the basis of computer aided geometry design (CAGD) and fractional Brownian motion (FBM) theory, two-dimensional and even higher-dimensional models and algorithms based on data distribution have been proposed. The disadvantages in dynamic scene integration of the navigating simulator have been overcome. The reliable results are available.

Keywords: Virtual Geographic Environment, CAGD, FBM.

1. INTRODUCTION

The Virtual Geographic Environment can be used as a multi-function geographic environment simulation system. The dynamic scene integration is one of the important parts of the system, which provides 3-D virtual geographic environments for supporting regional planning, conservation of water and soil, ecological environment construction. From the point of technique, there are some disadvantages of dynamic scene integration in the virtual geographic environment, such as the mountains can't be displayed in multi-resolution and the relative movements of the objects can't be described well [1-5]. In this paper, firstly, we try to give the mathematical models of the mountains in multi-resolution. In order to do this, we provide the stochastic interpolation model based on the smooth curves, and then reconstruct multi-node FBM curves using the provided stochastic model. Secondly, we extend the models of FBM curves to higher-dimensional FBM surfaces. Finally, we use the landform of the Loess Plateau as an example to test the mathematical models, the results are nice.

2. THE STOCHASTIC INTERPOLATION MODEL BASED ON THE SMOOTH CURVES

The k-th basic Spline functions based on the nodes set $\left\{-\frac{k+1}{2},-\frac{k-1}{2},\cdots,\frac{k+1}{2}\right\}$ can be expressed as[6-8]:

$$\Omega_k(x)=\frac{\delta^{k+1}x_+^k}{k!} \tag{1}$$

Where, $i=0,1,\cdots,k+1$, $u_+^k=\max\{u^k,0\}$, δ is the central difference operator，so the Eq. (1) can be denoted as：

$$\begin{gathered}\Omega_k(x)=\frac{1}{k!}\sum_{i=0}^{j}(-1)^i C_{k+1}^i\left(x+\frac{k+1}{2}-i\right)^k \\ j-\frac{k+1}{2}\le x\le j-\frac{k+1}{2}+1,\ j=0,1,\cdots,k\end{gathered} \tag{2}$$

Suppose $P_0,P_1,\cdots,P_{m+n}$ is the given points sequence. If the edge restricted condition is given, the value points can be extended to (*m*+*n*+*5*) points, otherwise, the extended points P_{-1},P_{-2} , P_{m+n+1},P_{m+n+2} can be obtained by simple translation, overlapped points and linear or nonlinear outer interpolation, so, the interpolated curves are:

$$P^{(k)}(t)=\sum_{j=-\left[\frac{k}{2}\right]}^{m+n+\left[\frac{k}{2}\right]}P_j\Omega_k(t)$$

Define the stochastic function as follows：

$$B(t)=\sum_{i=1}^{n}A_i\alpha(t-t_i)$$

Where, $\alpha(t)=\begin{cases}0 & t<0\\ 1 & t\ge 0\end{cases}$, $A_i(i=1,2,\cdots,n)$ is a stochastic variable in (-1,1), which satisfies Gauss distribution，in fact，$B(t)$ is the sum of a series of independent leaps. Hence, the following stochastic interpolation model based on the smooth curves is obtained：

$$f_0(x)=P^{(0)}(t)+B(t)=\sum_{i=1}^{n}[P_j\Omega_0(x-j)+A_j\alpha(x-j)] \tag{3}$$

This model appends the high-frequency ingredients gradually on the basis of the veracity of the low-frequency ingredients. Nowadays, one of the virtual systems — the navigating simulator adopts the line intersection algorithm to seek all the interceptive points of the scanning lines with all the objects, then connects the interceptive points with segments, which can not get multi-resolution. One main problem of the radar image display in the navigating simulator is how to simulate the coastline echo in the radar image. If the model of Eq. (3) is used, the effects will be better [9].

3. THE RECONSTRUCTION OF MULTI-NODE FBM CURVES

The main reason of putting forward FBM algorithm [10-12] is to provide a kind of interpolation method between the given data points:

Let I presents the interval $[t_0,t_N]$,N is a positive integer，and $t_0<t_1<\cdots<t_N,t_i=(i-1)h,i=1,2,\cdots,N$. If the points $\{P_j\},j=1,2,\cdots,N,P_j=(t_j,B_H(t_j,\omega))$ are given, the orbit of FBM in the 2-D plane will be reconstructed.

Add points using step h，and let the interrelated region always contain N points. Then for $\forall i=1,2,\cdots,N-1,j=1,2,\cdots,N-1$, we can calculate

$$c_{ij}=\frac{h^{2H}}{2}(|N-i|^{2H}-|i-j|^{2H}+|N-j|^{2H})$$

$$k_i=\frac{1}{2}(|N-i|^{2H}-|N+1-j|^{2H}+1)h^{2H}$$

Let

* This paper is supported by the Fok Ink Tong Education Grant of China (91077).

$b = (b_1, \cdots, b_{N-1})^T = C^{-1}K$

$\sigma = \sqrt{S - K^T b}$

So the value of $B_H(t_{N+1}, \omega)$ can be obtained by

$$B_H(t_{N+1}, \omega) = \sum_{i=1}^{N} B_H(t_N, \omega) b_i + \sigma R$$

Where, R is a Gaussian random variable with expectation 0 and variance 1.

For $B_H(t_{N+,i}, \omega)$, $i = 2,3,\cdots$, the algorithm can be used recursively.

The main character of the method is that it can create some curves with arbitrary ideal resolution independently and that it guarantees these curves in accordance.

4. HIGHER-DIMENSIONAL FBM METHOD

In the case of 2-D，the subdivision model of FBM can be constructed by the analogous process. The vision window of a computer screen is defined as:

$D = \{(x, y) : x_{\min} \le x \le x_{\max}, y_{\min} \le y \le y_{\max}\}$

Let $\Delta x = \frac{x_{\max} - x_{\min}}{M}, \Delta y = \frac{y_{\max} - y_{\min}}{N}$,M,N are positive integers.

D is divided by steps $\Delta x, \Delta y$,so $x_i = x_{\min} + i\Delta x, y_j = y_{\min} + j\Delta y, i = 0,1,\cdots, M - 1, j = 0,1,\cdots, N - 1$ are gotten.

Let X(0,0)=0,and $X(\frac{1}{m},0), X(0,\frac{1}{n}), X\left(\frac{1}{m},\frac{1}{n}\right)$ are the samples of a Gaussian random variable with expectation 0 and variance σ^2 ,then $X\left(\frac{1}{2m},\frac{1}{2n}\right)$ is supposed to be the average value of X(0,0), $X(\frac{1}{m},0), X(0,\frac{1}{n}), X\left(\frac{1}{m},\frac{1}{n}\right)$ adding the Gaussian random excursion D_1 with expectation 0 and variance

$$\Delta_1^2 = \frac{\sigma^2}{2^{2H}}\left[\left(\frac{1}{m^2} + \frac{1}{n^2}\right) - 2^{2H-4}\left(\frac{1}{m^{2H}} + \frac{1}{n^{2H}} + \left(\frac{1}{m^2} + \frac{1}{n^2}\right)^H\right)\right]$$

In the 2-th recursion，the value of the center point of each child rectangle grid is the average value of the four corners adding a Gaussian random excursion with expectation 0 and variance

$$\Delta_2^2 = \frac{\sigma^2}{2^{2H}}\left[\left(\frac{1}{m^2} + \frac{1}{n^2}\right) - 2^{2H-4}\left(\frac{1}{m^{2H}} + \frac{1}{n^{2H}} + \left(\frac{1}{m^2} + \frac{1}{n^2}\right)^H\right)\right]$$

Apply FBM1D() to the edges of each grid to produce the value of the midpoint of each edge，the rest may be deduced by analogy, in the nth recursion, the value of the center point of each child rectangle grid is the average value of the four corresponding corners adding a Gaussian random excursion with expectation 0 and variance

$$\Delta_n^2 = \frac{\sigma^2}{2^{nH}}\left[\left(\frac{1}{m^2} + \frac{1}{n^2}\right) - 2^{2H-4}\left(\frac{1}{m^{2H}} + \frac{1}{n^{2H}} + \left(\frac{1}{m^2} + \frac{1}{n^2}\right)^H\right)\right]$$

When all the points are applied by random addition method, the produced fractal surface will be a self-affine fractal surface**.**

5. CONCLUSIONS

The area of the Yellow River's Loess Plateau of China is more than 630 thousands sq km, and more than 100 millions population live in this area. With arid climate and heavily desertification, this area is full of gullies and water, the soil erosion is very heavy. Thus, lots of natural disasters often happen. Modern spatial information technology such as geographic information system (GIS), remote sensing and Global Positioning System can allow us to build digital Loess Plateau for dealing with ecological problems and regional sustainable development.

With regard to the specific landscape on the Loess Plateau, the modeling methods given above were used to simulate and display landscape on the Loess Plateau.

In grid-based modeling mode, the terrain was created based on Digital Elevation Model (DEM) data, Remote Sensing (RS) Image and other images, and the terrain is made up of triangular network.

Digital Elevation Models (DEMs) are digital files consisting of points of elevations, sampled systematically at equally spaced intervals.

Grid DEM data is from papery map of scale 1:10000, which was issued by State Bureau of Surveying and Mapping of China. Firstly, the map was scanned into computer, and then become vector graph by handwork with GIS software, for example, ArcGis or MapInfo, mostly the vector graph is digital contour line, and then GIS software can produce Grid Dem automatically (see Table 1). There are another popular format of Dem, namely Triangular Irregular Network (TIN) can be created. For the following reason, Grid was selected in constructing terrain on the Loess Plateau:

Table 1. Grid DEM

75	85	97	82
81	99	86	77
92	88	95	85
99	75	86	79

1. Because of equally spaced intervals, the grid data can easily constructed into triangle network (Fig 1).
2. Random column and row of Grid can be read easily, so large scale terrain can be divided into some small pieces. With small pieces of grid file, not only the speed of constructing can be accelerated, but also the requirement for computer hardware can be easily fulfilled.

Comparing with Triangular Irregular Network, gird data have too many redundant triangles during constructing terrain, thus, data’s quantity of same area is larger than TIN.

After Normal of each triangle was calculated and light rendering, the 3D terrain was constructed (see Fig 2).

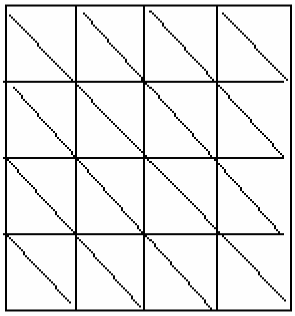


Fig.1. Trianglar network constructing

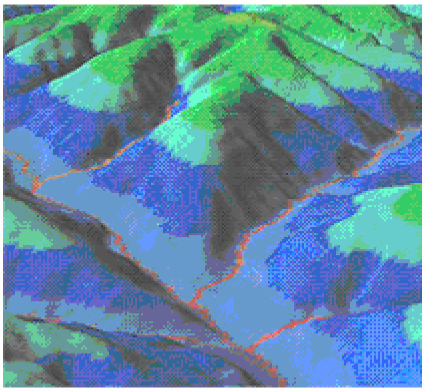


Fig.2. Terra in of gully lines overlaid

REFERENCES

[1] Jian Hua Gong and Hui Lin, "A Collaborative VGE: Design and Development", A Chapter to a forthcoming book titled *Collaborative Geographic Information Systems*, Shivanand Balram and Suzana Dragicevic, Simon Fraser University, Canada, 2005.

[2] Li,Zi-Cai, "New discrete techniques for 3D image transformations", *Computera & Mathematics with Applications*, 1998.

[3] Zhang,J. et al, "Wavelet-based multiresolusion statistical model for texture", *IEEE Transaction on Image Processing*, Vol.7, No.11, 1998, pp.1621~1627.

[4] Maragos,P., "Synthesis and application of lattice image operators based on fuzzy norms", *IEEE International Conference on Image Processing*, Vol.1, 521-524, 2001.

[5] Wittenbrink,Craig M., "IFS fractal interpolation for 2D and 3D visualization", *Proceedings of the IEEE Visualization Conference*, 77-84, 1995.

[6] Selim G. Akl, "Parallel Computation: Models and Methods", Prentice Hall, Upper Saddle River, New Jersey, 1997.

[7] N. Touheed, P. Selwood, P. K. Jimack, & M. A. Berzins, "Comparison of Some Dynamic Load-Balancing Algorithms for a Parallel Adaptive Flow Solver", *Parallel Computing*, Vol.26, 2000, pp.1535~1554.

[8] Dan Liu et al, "A Survey of Parallel Algorithms for Fractal Image Compression", *Journal of Algorithms and Computational Technology,* 2007.

[9] Liu Dan， *Practical Fractal Graphics,* Dalian：Dalian Maritime University Press，2001.

[10] Zeng Wen-qu, Wang Xiang-yang, Liu Dan et al，*Fractal Theory and Fractal Computer Simulation*, Shenyang: North-eastern University Press, 1993.

[11] M. F. Barnsley, *Fractals Everywhere*, Orlando: Academic Press, 1988.

[12] H. Wallin, "Interpolating and orthogonal polynomials on fractals", *J Constr Approxi*, No.5, 1989, 137~150.

Dan Liu is a Full Professor of Dalian Maritime University. She graduated from Jinlin University in 1991; from Dongbei University in 1994; from Dalian University of Technology in 1998. She is a holder of Fok Ink Tong Award of China (2004); was a Visiting Professor of University of Technology, Sydney (2003~2004), Visiting Professor of the University of Leeds, UK (2006~2007). She was the chairman of PDPTA 2004, co-chair of PDPTA 2005 and ICITA2004. She has published four books, over 50 research papers. Her research interests are in digital image processing, fractals, system simulation and CAGD.

Multi-channel Digital Image Capture System Based on TMS320DM642

Yun Lei[1], Xiaojun Tong[1], Qiuming Huang[1]
[1]Department of Mathematics and Physics, Wuhan Polytechnic University
Wuhan 430074, The People's Republic of China
Email: tongxiaojun1998 @yahoo.com.cn

ABSTRACT

In this paper, aiming at the needs of high stable and robust digital image capture system, considering the continuously upgrades of processor core, we introduce the architecture and functions of a real time multi-channel digital image capture System based on TMS320DM642, in the system, we emphasizes the hardware design and realization, Philips's SAA7121H and TI's TVP5150 are used to form video channel. We described the details of Video Port hardware design. Then analysis the system's application foreground, the system can be used in varied application field such as medical image process, video surveillance, the system is shown by experiments to have high performance and powerful expansibility.

Keywords: TMS320DM642, Video Port, Video Surveillance

1. INTRODUCTION

About real time image capture system design, there are varied methods to realize, nowadays, we have two chief methods. The first method is the use of video decoder/encoder, this approach's advantage is that people can save software development time, the disadvantage is that once the design finished, it is difficult to modify. The second method is the use of customizing DSP instrument, this approach's advantage is that it's easy to maintain and redevelopment.

Our system adopts the second method, the system adopts TMS320DM642 which is developed by TI Incorporated as the processor core, the system includes Image capture module, Image process module, and image display module.

2. INTRODUCTION TO THE SYSTEM MAIN FRAME

At present, the most popular system design approach is "Coder/decoder +DSP+FPGA" pattern, the function of DSP is to achieve collection, preprocessing, segmentation, target identification, tracking and estimation; and FPGA is used to overlap video signals and control, coder charges for converting analog signal to digital signal, decoder charges for converting digital signal to analog signal.

The essential part of our system is a PCB board based on TMS320DM642, The PCB board includes I/O module; DSP module; extern RAM module; Ethernet module; UART module; power module. With such modules, it's convenient to extend the function of our system. Considering the character of TMS320DM642 [1], there are three video ports integrated in DM642, so instead using FPGA, we can use the Video Port to form the video channel directly, to make the design a easier way, Our system's main frame is shown in Fig.1.

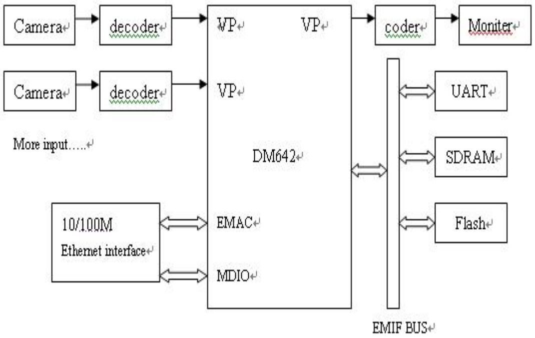


Fig.1. System main frame

There are three Video Ports [2] in DM642, it can support six interfaces of 8-bit BT.656. System's work flow is as below:

(1) When power up, Camera generate PAL format data, TVP5150 transfer the PAL format data into BT.656 format data.
(2) By the control of local clock, TVP5150 send the data into the Buffer of DM642 automatically.
(3) While the entire frame's data has been sent, DM642 generate a DMA interrupt event, in the DMA ISR, we do MPEG-4 algorithm or other customized algorithm.
(4) The processed data would be sent to SAA7121H via Video Port, SAA7121H transfer the data into PAL format data, then send the data to the monitor via S-port.

3. VIDEO PORT AND CODER/DECODER PART

3.1 Video Port

The video port [3] peripheral can be operated as a video capture port, video display port, or transport stream interface capture port. Video capture port provides the following functions:

(1) Capture rate of up to 80MHZ.
(2) Two channels of 8/10-bit digital video input in YCbCr 4:2:2 format.
(3) One channel of Y/C 16/20-bit digital video input in YCbCr 4:2:2 format on separate Y and Cb/Cr inputs.
(4) YCbCr 4:2:2 to YCbCr 4:2:0 horizontal conversion and 1/2 scaling in 8-bit 4:2:2 modes.

DM642 integrates three video port peripherals: vp0, vp1 and vp2. Each video port has 20-bit data bus interface, two clock signals VPxCLK0 and VPxCLK1, and three control signals VPxCLT0, VPxCLT1, VPxCLT2. Each video port consists of two channels: A and B, in DM642, VP2 only be operated as video display port, VP0 and VP1 can be operated as a capture port or display port. Two channels must work in the same pattern, each channel has 10-bit data bus, so each channel can receive or send data in YcbCr format.

VP's data transfer use DMAs, Capture rate is up to 80MHZ, there is a FIFO in each video port, the size of FIFO can be customized, when FIFO is full, DMA event is triggered. The optimized size of FIFO is the multiple of pixel per line in the picture.

3.2 DSP connect with TVP5150

TVP5150[4] is a video decoder which can transfer PAL format data into YUV4:2:2 format data, the block diagram of the connecting between DM642 and TVP5150 is shown in Fig.2.

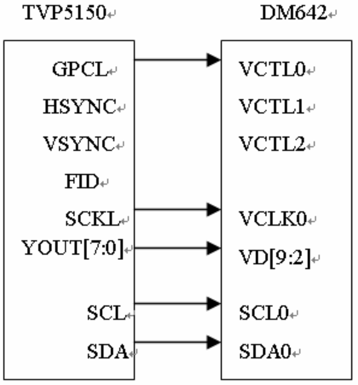


Fig.2. DSP connect with TVP5150 main frame

as Fig.2 shows, GPCL pin of TVP5150 is used to enable video port, DM642 control TVP5150 by I^{2}C host interface, TVP5150 functions as a slave device, DM642 functions as a master device. For TVP5150 output BT.656 format data, HSYNC, VSYNC, FID signal is not necessary.

DM642 configure TVP5150 as follow steps:

(1) DM642 initiates a write operation to the TVP5150 by generating a start condition
(2) DM642 present TVP5150 I^{2}C address, then followed by a 0 to indicate a write cycle.
(3) After receiving an acknowledge from TVP5150, DM642 presents the sub address of the register it wants to write
(4) After receiving an acknowledge from TVP5150, DM642 present the data
(5) After receiving an acknowledge from TVP5150, DM642 terminates the write operation by generating a stop condition

3.3 DSP connect with SAA7121H

SAA7121H [5] is a digital video encoder, the basic encoder function consists of subcarrier generation, color modulation and insertion of synchronization signals. it encodes digital luminance and color difference signals into analog CVBS and simultaneously S-VIDEO signals, by configuring SAA7121H's internal register, the output mode can be operated as RGB mode, S-Video mode or VGA mode.SAA7121H's features are as follow:

(1) Three Digital-to-Analog converters
(2) Fast I^{2}C –bus control port
(3) Accepts MEPG decoded data on 8-bit wide input, accepts NTSC or PAL output,
(4) Programmable horizontal sync output phase

The block diagram of the connecting between DM642 and SAA7121H is shown in Fig.3.

SAA7121H can be operate as master device or slave device, in our system, for the video port of DM642 generates clock signal, we configure DM642 as master device, SAA7121H as slave device. Video port generate PIXCLK, HSYN, VSYN, BLANK signal, then output to SAA7121H via VCLK1, VCTL0~2 pin, SAA7121H synthesize received signal to analog signal, then send to monitor via S-port.

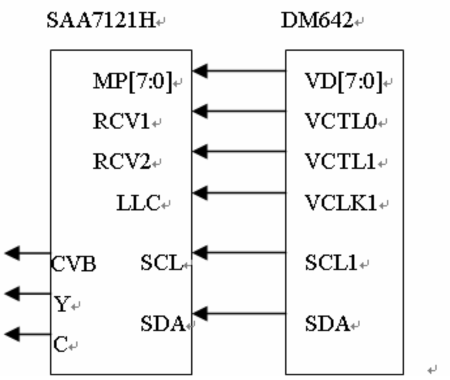


Fig.3. DSP connect with SAA7121H main frame

4. ETHERNET INTERFACE DESIGN

In DM642, HPI, PCI, EMAC peripherals share the same pins, except HPI16 is compatible with EMAC, the other peripherals are mutually exclusive. In our system, we specify EMAC peripherals by enable MAC_EN pin and disable PCI_EN pin, EMAC peripheral functions data link layer, so we use BCM5221 as a Ethernet transceiver to get 10/100M Ethernet interface, We choose 406549-1 chip of AMP Incorporated as RJ45 linker, as Fig.4 show:

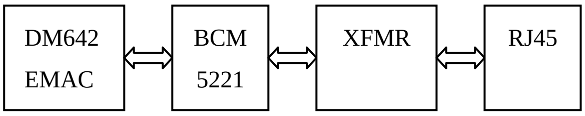


Fig.4. Ethernet interface main frame

5. CONCLUSIONS

Since we consider about the function to be extended, our system can be applied in embedded video communication terminals, medical image process, video surveillance such fields. The application of our system has a well prospect.

REFERENCES

[1] TMS320DM642 Technical Overview (SPRU615) http://www.ti.com.cn.

[2] TMS320C64X DSP Video Port/VCXO Interpolated Control Port Reference Guide (SPRU629) http://www.ti.com.cn.

[3] Wang jie yang, "The video port of TMS320DM642 in image process system", *Microcontrollers and Embedded systems*, 2006, 6:62-65

[4] *TVP5150APBS Ultra low Power NTSC/PAL/SE-CAM Video Decoder with Robust Sync Detector Data Manual* [Z].Texas Instruments, 2003

[5] *SAA7120H; SAA7121H Digital Video encoder Data Sheet*. Philips Semiconductors, 2002

Embedded System,
Hardware Design and Diagnosis

The DPK Scheduling Algorithm for CMP Hard Real-Time Applications*

Man Wang[1], Zhihui Du[1], Zhiqiang Liu[2], Song Hao[1]
[1]Department of Computer Science and Technology, Tsinghua University, 100084, Beijing, China,
[2]Hebei University, 071002, Baoding, China
Email: wangman05@mails.tsinghua.edu.cn

ABSTRACT

In order to boost the potential power of Chip Multiprocessor (CMP), the DPK (Dynamic priority and 0-1Knapsack) algorithm is proposed in this paper to handle the scheduling problem of multiple DAG-structure hard real-time applications. Although many DAG scheduling algorithms are created for heterogeneous computing environment, the DPK algorithm is mainly based on the unique characters of CMP, and provides three dispatch queues with different level to schedule the multiple DAG-structure applications as a whole. What's more, the DPK algorithm does not only use the deadline to define the priority of each application. Instead, the algorithm utilizes Laxity, a dynamic parameter to measure the current urgency of each application. Furthermore, at the end of each scheduling step, the algorithm finds other proper unscheduled sub-jobs in any application to fill the idle time slice generated in this scheduling step just like the classical 0-1 Knapsack problems. According to the algorithm analysis and simulation experiments, with the DPK algorithm, the Successful Rate can be increased a lot and the idle time of each processor is reduced.

Keywords: Chip Multiprocessor (CMP), Scheduling, Dynamic Priority

1. IMPORTANT INFORMATION

Chip multiprocessor (CMP) [1] is a relatively new micro-architectural paradigm in recent years, which means multiple processors, or "cores" on a single die. The purpose of CMP is to allow a chip to achieve greater throughput; however, because of the software which can not match current CMP hardware technology, the total power of CMP can not be reached. And how to schedule the tasks is still one of the key problems. In order to utilize the computing ability well, it is necessary for most applications to be separated as several sub-jobs, the relationship among which can be modeled as a DAG (directed acyclic graph), to be mapped on different cores. What's more, in many applications, the hard real-time is also required. When several such kinds of jobs come at the same time, how to schedule them is the main problem discussed in this paper. Although chip multiprocessor can be included in the heterogeneous computing environments, the CMP has its own characters. For example, because all processors are located in the same die and the shared cache exists, the data exchange is much quickly than in network. And different from other heterogeneous computing environments which based on network, the whole situation of CMP can be monitored more. Therefore, many current DAG or real-time scheduling algorithms can not be used directly in CMP.

As a result, in this paper, a DPK scheduling algorithm is proposed, which can be viewed as the hybrid of processor scheduling conception and other HPC scheduling conception. What's more, the DPK scheduling also considers the multi DAG condition as a whole to optimize the total throughout, reduce the reaction time and processors' idle time.

The rest of this paper is organized as follows. Section 2 presents an overview of scheduling problem discussed in the paper. Section 3 elaborates the DPK scheduling Algorithms. Section 4 presents the simulated experiments. Section 5 discusses the related work and Section 6 concludes the paper.

2. SCHEDULING MODEL

2.1 CMP Environment and Limitation

Currently, there are different architectures to describe chip multiprocessor, and here, a general one is considered in this paper. A CMP has n heterogeneous cores, each of which is an indispensable processor. Each core has its own cache and a shared cache also exists in the same chip. In the CMP, centralized scheduling policy is adopted. One core, is responsible for the scheduling, is called scheduling core.

There are still other limitations here. 1) The arrival pattern of discussed applications is aperiodic. 2) All sub-jobs in each application are non-preemptive. 3) When a sub-job is scheduled in a processor, it can not be migrated, or the sub-job can be called partitioned [4]. 4) Besides the processors, there are other resources always being needed. The same as [11], two kinds of resources are defined here: shared-resources EATs, and excluded resources EATe. 5) When the data is generated, the time of accessing data (no matter from local cache or shared cache) can be hided in the computing time and need not to be considered separately.

2.2 Application Model

This paper pays attention on the applications, which consist of, or can be separated, into a series of sub-jobs. What's more, there is a partial executing order on the sub-tasks. That is, such an application can be modeled as a directed acyclic graph (DAG), the nodes of which represent the sub-jobs and the edges show the dependence among the sub-tasks.

Here, we use the symbol A to represent an application. In an application, a sub-job is represented by the symbol SJ, and the edge is E. If SJ(j), which means the sub-job j, is the successor of SJ(i), E(i,j)=1. Then, an application is characterized by the following:

- ♦ AT: the arriving time of an application.
- ♦ HD: the hard deadline of an application.
- ♦ CET: the current earliest estimated finishing time of the application. This parameter is calculated dynamically according to the current execution situation.
- ♦ EST: the earliest starting time of an application.
- ♦ AFT: the actual finishing time of an application.
- ♦ SJ.WET: the worst execution time of sub-job SJ.
- ♦ SJ.EST(i): the earliest starting time of sub-job SJ on processor i.

* This paper is partly supported by National Natural Science Foundation of China(No.60503039 and No.10778604), Beijing Natural Sciences Foundation(No.4042018) and China's National Fundamental Research 973 Program (No. 2004CB217903)

♦ SJ.LST: the latest starting time of sub-job SJ.

According to the conceptions mentioned above, the scheduling problem is defined as follows:

Problem1.

INSTANCE: Set M of DAG-structure applications. For each application Mi∈M, there are Num(i) sub-jobs with their own execution length, and a start-time Mi.AT and a deadline Mi.HD.

QUESTION: Is there an n-processor CMP schedule for M that obeys the precedence constraints and resources constrains, and meets all the deadlines? That is, for each Mi, Mi.AFT<=Mi.AT+Mi.HD.

In order to analogy and compare this Multi-DAG applications scheduling and the independent tasks scheduling in multiprocessor, which has been already discussed for a long time by lots of literature[2][6], it is necessary and feasible to change Problem1 into Problem2:

Problem2:

INSTANCE: Set M of DAG-structure applications. For each application Mi∈M, there are Num(i) sub-jobs with their own execution length, and a start-time Mi.AT and a deadline Mi.HD. The total number of all sub-jobs is $\sum_{i=1}^{M} Num(i)$. Now, these $\sum_{i=1}^{M} Num(i)$ sub-jobs can be viewed as independent with each other. However, all sub-jobs are also viewed as resources which are required by other sub-jobs, if and only if there is dependence between them in original application. For example, if in application Mi, E(k,l)=1, Mi.SJ(k) is one of the shared resources of Mi.SJ(l). The EATs(Mi.SJ(k)) is the finishing time of Mi.SJ(k).

QUESTION: Is there an n-processor CMP schedule for $\sum_{i=1}^{M} Num(i)$ sub-jobs that obeys the resources constraints, and meets all the deadlines?

Problem2 has the similar description to the general hard real-time problem, and is a NP-Complete problem, the relevant proving processing of which can be found in [3]. In the following, a heuristic algorithm is provided based of the application model mentioned above.

3. DPK SCHEDULING ALGORITHM AND ANALYSIS

3.1 DPK scheduling algorithm

In DPK scheduling, three dispatched queues are kept by scheduler: the Main Queue, Schedule Queue and Ready Queue. Each queue has its own function to decide the priory of the elements.

Main Queue (MQ): a queue is used to store the application. The order is based on H_{MQ}, non-decreased. For each application in MQ, the function is:

$$H_{MQ}\text{= Laxity =A.HD-A.CET} \qquad (1)$$

Schedule Queue (SQ): a queue is used to store all SJs of the first application A_{first} in MQ. The order is based on H_{SQ}, non-decreased. For each SJ in A_{first} , the function is:

$$H_{SQ}\text{=SJ.LST} \qquad (2)$$

Ready Queue (RQ): a queue is used to store all SJs of the other applications in MQ, exception the first one. The order is based on H_{RQ}, non-decreased. For each SJ mentioned above, the function is:

$$H_{RQ}\text{=SJ.LST -SJ.EST-SJ.WET} \qquad (3)$$

In the algorithm, MQ, SQ and RQ are dynamic, even there is no new application coming. The main conception is that the scheduler adjusts the priority of each application according to the current situation of cores, to finish more application successfully.

Virtual Queue (VQ): For each core, there is a relevant VQ. In each scheduling cycle, the scheduled SJs are stored in VQ. In VQ, the schedule order can be changed. At the end of the scheduling cycle, the SJs in VQ are put into the relevant processors' local dispatch queue.

DPK algorithm:

Scheduling algorithms

♦ Step1: If there is a new application A_new comes, compute A_new.CET, and for all SJ in A_new, compute SJ.LST.

♦ Step2: Compute A_new. H_{MQ} and insert A_new into the proper place of MQ.

♦ Step3: Select the first application A_{first} in MQ.

♦ Step4: Send all A_{first}'s all SJs which have no any predecessor into SQ according to their H_{SQ}.

♦ Step5: Send all SJs of the application in MQ, except A_{fisrt}, which have no any predecessor into RQ

♦ Step6: Schedule all elements in SQ into the VQ based on the principle that each SJs are scheduled into the processor which makes it to execute earliest.

♦ Step7: For each processor which has been scheduled by Step6, schedule the elements in RQ to fill the idle time of each VQ using the method which is the same as 0-1 Knapsack problem.

♦ Step8: Transfer the VQ into the relevant processor dispatch queue.

♦ Step9: Delete all scheduled SJs from the relevant DAG, adjust MQ, and go to Step1.

Now, some explain of the algorithm is as follows.

(1) According to formula (1), H_{MQ} is decided by A.HD and A.CET. Generally, A.HD is defined by users and can not be changed during the whole process, and A.CET is defined by formula (4) here,

$$\text{A.CET=}t + \sum_{all\ SJ\ in\ critical\ path} SJ.WET \qquad (4)$$

The critical path here is mentioned as the longest length in the current DAG of the application. In Step8, the scheduled SJs are removed from the DAG, and the critical path may be changed.

(2) Like myopic algorithm [5], the resources are also considered here, but for simpler explanation, we just discuss that for each resource, there is only one instance. Similarly, We use EATs(k) to represent the earliest available time of shared resource R_k, while EATe(k) to represent the earliest available time of excluded resource R_k. When a SJ can be executed, all resources must be available, its all predecessors should be finished, and at least one core is idle. As a result, SJ.EST(i) means the earliest time of SJ begins to execute on processor i.

(3) In step6, it does not schedule the sub-jobs in the real cores dispatch queue. Instead, the scheduled SJs are stored in each core's VQ. The advantage of this method is to make changes in Step7. If a sub-job is scheduled into the real queue directly, it is inconvenient to let each processor to adjust the order of tasks. Therefore, before Step8, the SJs are put into the VQ, and it is good for the scheduler to reorder the SJs position to increase the resources utilizing efficiency and reduce the processors' idle time.

(4) In Step7, the idle time means, the blocking time of SJs scheduled in Step6. For example, assume SJ_1 is prepared to be scheduled in Processor 1 in Step6. Although Processor 1 provides the earliest start time, SJ_1 still needs to wait 5 time units more when the processor is free to wait for other resources. At this time, this 5 time units is called the idle time, and is like a Knapsack. So, how to select the rest unscheduled SJ to fill the idle time can be viewed as a 0-1 Knapsack problem, and the simplest widely used greedy algorithm is used here.

3.2 Algorithm Analysis

The main ideas of DPK scheduling algorithm are as follows.

(1) The DPK algorithm considers several applications as a whole. Different from other DAG scheduling heuristic algorithms which consider applications one by one, the DPK algorithm considers all sub-jobs which have no predecessors in all applications at each scheduling cycle. For example, in Fig.1, assume application A has the higher priority than application B now. In the first scheduling cycle, node0 and node1 are firstly scheduled according to the algorithm. However, node0 in application B is also considered subsequently. If there is a proper processor to schedule, this node will be scheduled right now.

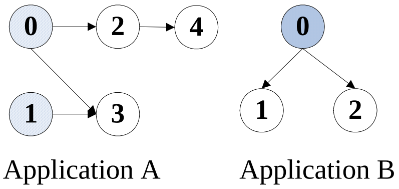


Fig.1. A example of scheduling order

(2) The DPK uses a dynamic priority, which is discussed in details later, to represent the urgency of the application. For example, in Fig.2, assume application A has the earlier deadline than application B, and A and B have the same AT. At the beginning time, showed in Fig.2(a), by computation, A's priority is higher, so, some of SJs in A are scheduled first. When CET is computed, it uses the WET of each SJ, but in real scheduling, the true execution time of a SJ may be less. Therefore, it is possible that after some times, according to the current execution situation, A.CET may be reduced, as Fig2(b) shows. In Fig.2(b), A is anticipated to be finished ahead a much longer time than B. So, although the deadline is not changed, in DPK algorithm, at this time, B's priority is higher than A, and the scheduler considers SJs in B first now. Contrarily, if deadline is considered only and the priority is stable, the extreme condition, showed in Fig.2(c) may occurs. As a result, this dynamic priority makes the scheduler more flexible and then increases the success rate.

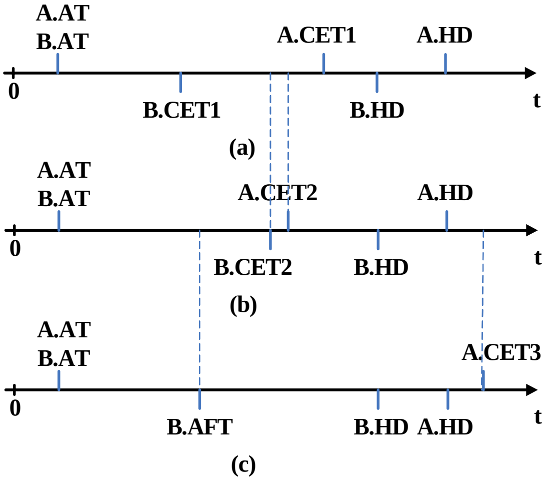


Fig.2. An example of priority changing condition

(3) The DPK algorithm introduces the Virtual Processor Queue and the classical 0-1Knapsack conception here to reduce the processors' idle time. Once a sub-job is scheduled in a real processor dispatch queue, it is hard and improper for the central scheduler to change the order of this sub-job. So, if the sub-job is blocked because of waiting resources ready or its other factors, the processor idle time comes. Fig 3 shows the advantage of adopting the 0-1Knapsack conception. Use Application A and B described in Fig1 as the example. On the one hand, in Fig.3(a), in the second scheduling cycle, A_2 and A_3 are scheduled in VQ first. However, due to that A_3 can start only after A_0 finishing, there is an idle times slice between A_1 and A_3 in processor 2. So, DPK algorithm views the idle time slice as a 0-1Knapsack, and finds a proper "goods" B_0 to fill in. Similarly, in the third scheduling cycle, A_4 is used to fill in the idle time. One the other hand, in Fig.3(b), although Application A and B both meet the deadline, there actual finish time is late and the idle time is more. If there are more applications or the deadline is earlier, the situation in Fig.3(b) has the higher possibility to be failed .

A concrete example serves to show the advantages of DPK scheduling algorithm. There are 2 processors and 2 resources R1,R2. For the sake of simplicity, a resource only has one instance. Two applications A and B with the DAG-structure described in Fig1, and the WET of each SJ in these applications is showed in table1. A and B come at the same time, and their deadline are: A.HD=18, B.HD=17. Based on the deadline of each application, each SJ.HD can be computed and showed in tabel1 also. Using the general DAG scheduling method, which just considers the deadline of each application, Fig5(a) shows, at least one application can not be finished before its deadline. Fig5(b) shows scheduling these sub-jobs according to Problome2 using myopic algorithm with K=3, W=1 and Backslid=1[5], the schedule is also failed. But Fig5(c) uses DPK algorithm provided in this paper, the two applications both meet the deadline.

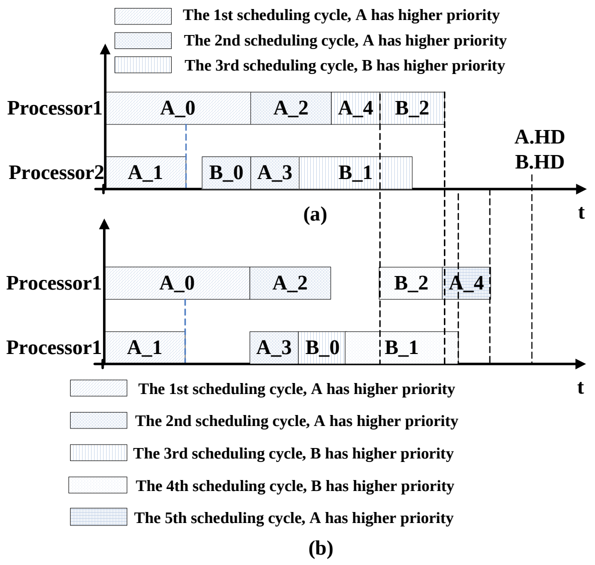

The 1st scheduling cycle, A has higher priority
The 2nd scheduling cycle, A has higher priority
The 3rd scheduling cycle, B has higher priority
Processor1
Processor2
A_0
A_2
A_4
B_2
A_1
B_0
A_3
B_1
A.HD
B.HD
(a)
t
Processor1
Processor1
A_0
A_2
B_2
A_4
A_1
A_3
B_0
B_1
t
The 1st scheduling cycle, A has higher priority
The 2nd scheduling cycle, A has higher priority
The 3rd scheduling cycle, B has higher priority
The 4th scheduling cycle, B has higher priority
The 5th scheduling cycle, A has higher priority
(b)

Fig.3. The comparison of introducing 0-1Knapsack conception or not

Table 1.

	WET	HD	Resource
A0	5	18	R1/e
A1	2	18	
A2	4	18	R2/e
A3	3	18	R2/e
A4	3	18	
B0	2	17	
B1	6	17	R1/e
B2	5	17	R2/e

4. SIMULATION EXPERIMENT

In the simulation, we compare the DPK algorithm and the general method that scheduling each application according to their deadline.

There are three kinds of DAG structures in the experiment, which are showed in Fig.6. 200 applications are generated randomly as one of these three structures. But each SJ.WET is generated and in different application they are different. Moreover, the deadline of each application is also provided randomly. 8 cores are considered in the experiments and the time of data transferred among cores is ignored.

In the experiment, we mainly compare DPK scheduling algorithm and the deadline-driven algorithm (DDA), which sets the priority of each application according the deadline.

Formula (5) defines SR (Successful Rate). There, N_{suss} means the number of applications which are successfully finished before deadline.

$$SR = \frac{N_{suss}}{total\ \ number\ \ of\ \ applicatio\ \ ns} \qquad (5)$$

To show the advantage of DPK scheduling algorithms, a variable k is introduced for the deadline of each application. With other parameters stable, the deadline of each application multiplies different k in experiments. If k is bigger, this means there is more time for each application. Fig.7 shows SR of DPK and DDA according to different k. It is easy to say that when the applications are dense, DPK shows greater advantages. The similar condition is also showed in Fig.8, which describes the total cores utilization.

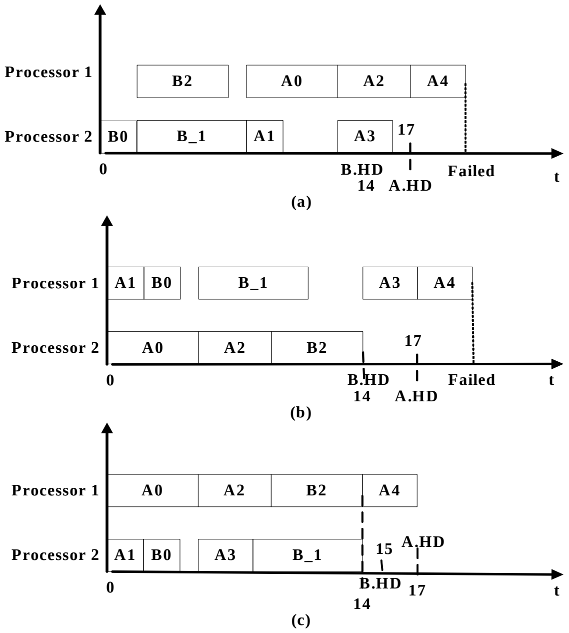


Fig.5. The scheduling result for table 1

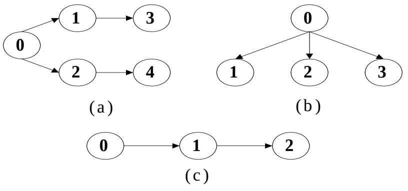


Fig.6. The structures of applications

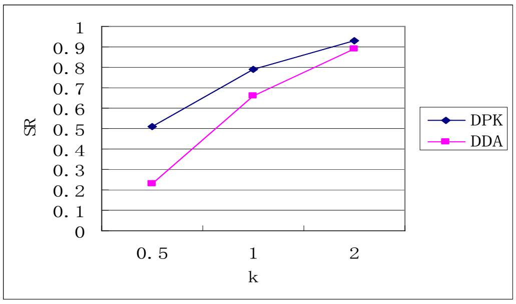


Fig.7. The SR of two algorithms according to k

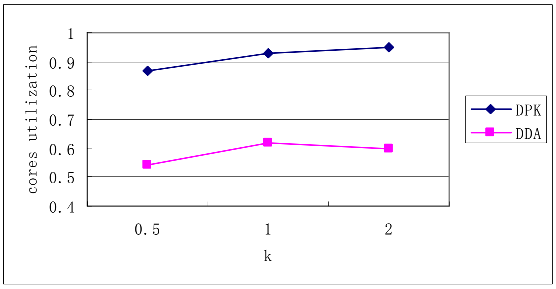


Fig.8. The cores utilization of two algorithms according to k

5. RELATED WORK

The DAG scheduling algorithms for non real-time applications in heterogeneous computing environment has been studied by many literatures [7][8].

The real-time scheduling for indispensable applications on heterogeneous computing environment have also been researched for many years, and many literature [4][6][9][10][5] give different solutions. In this paper, the aperiodic applications are considered only.

However, the research about real-time DAG scheduling algorithms is relatively fewer. [11][12] use real-time DAG to research the scheduling problem of dependent tasks, but they all static scheduling algorithm. Other literatures [13] provide the algorithm which can be only used in the homogeneous environment. [14] presents a dynamic real-time scheduling algorithm DEFF to deal with multiple parallel applications which are modeled by DAG in heterogeneous environments. [15] proposes an algorithm to schedule the sub-jobs by utilizing the spare capability left by the periodic real-time jobs and an Early Deadline First policy. But the DPK algorithm considers both of the priority of application level and sub-jobs level to make a more precise solution, and when the idle time considered, DPK introduces the method of solving 01Knapsack problem. Besides, all algorithms mentioned above pay less or even no attention to CMP, while DPK algorithm is focused on the CMP heterogeneous environment to evoke more power of CMP.

6. CONCLUSIONS AND FUTURE WORK

In this paper, the DPK scheduling algorithm for multiple DAG-structure applications in CMP is discussed. This algorithm considers all current applications as a whole, rather than scheduling them one by one, and introduces the conception of solving 01Knapsack problem, to achieve higher SR and lower processor idle time. The algorithm analysis and simulation experiments show the advantages.

This paper represents our first and preliminary effort to research this complex problem in CMP, and there are still some open problems to be solved. The next steps of the planned work include the followings:

First, based on the current algorithm, the fault-tolerant should be considered in the next step. Second, for more complex situation, we consider the network in CMP and the data communication is taken into account.

REFERENCES

[1] L. Hammand, B. A. Nayfeh, and K. Olukotun. “A single-chip multiprocessor”. *IEEE Computer*, 30(9):79–85, Sep. 1997.

[2] C. L. Liu and J. W. Layland, “Scheduling Algorithms for Multiprogramming in a Hard Real-Time Environment,” *J. ACM*, Vol.20, No.1, pp.46-61, 1973.

[3] M.R. Garey, and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, New York. 1979.

[4] BJO¨ RN ANDERSSON, *Static-priority scheduling on multiprocessors*, Thesis for the degree of Doctor of philosophy, Chalmers University of Technology, G¨oteborg, Sweden 2003.

[5] K. Ramamritham, J. A. Stankovic, and P.-F. Shiah, “Efficient Scheduling Algorithms for Real-Time Multiprocessor Systems”. *IEEE Trans. Parallel and Distributed Systems*, Vol.1, No.2, pp.184-194, Apr. 1990.

[6] H. Jin and P. Tan, “A Novel Dynamic Allocation and Scheduling Scheme with CPNA and FCF Algorithms in Distributed Real-time Systems”, *Proceedings of the 2005 11th International Conference on Parallel and Distributed Systems* (ICPADS'05).

[7] T.D. Braun, et al. « A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems”. J*ournal of Parallel and Distributed Computing*, 61:810-837. 2001

[8] J. Blythe, et al. “Task scheduling strategies for workflow-based applications in grids”. *Cluster Computing and the Grid, 2005. IEEE International Symposium* on Volume 2, Page(s):759 – 76. 2005.

[9] S. K. Dhall and C. L. Liu, “On a Real-Time Scheduling Problem”. *Operations Research*, Vol.26, No.1, pp.127-140, 1978.

[10] M. L. Dertouzos and A. K. Mok, “Multiprocessor On-Line Scheduling of Hard Real-Time Tasks”. *IEEE Trans.Software Eng.*, Vol.15, No.12, 1989, pp.1497-1506, Dec.

[11] X. Qin, H. Jiang, C.S.Xie, and Z.F.Han, “Reliability driven scheduling for real-time tasks with precedence constraints in heterogeneous distributed systems”. *In Proceeding of 12th International Conference Parallel and Distributed Computing and Systems*, 2000.

[12] X. Qin, Z.F.Han, H.Jin, and L.P.Pang, “Real-time fault-tolerant scheduling in heterogeneous distributed systems”. *In proceeding of the 2000 International Conference on Parallel and Distributed Processing Techniques and Applications*, June 26-29, Vol.I, pp. 421-427. 2000.

[13] M.Y. Wu, W.Shu, and Y. Chen. “Runtime parallel incremental scheduling of DAGs”. *In proceeding of the 29th International Conference on Parallel Processing*, 2000, pp. 541-548.

[14] D. Ma, W. Zhang, Q. Li, “Dynamic Scheduling Algorithm for Parallel Real-time Jobs in Heterogeneous System”, *Proceedings of the Fourth International Conference on Computer and Information Technology* (CIT’04), 2004.

[15] L. He, Stephen A. Jarvis, Daniel P. Spooner and Graham R. Nudd. “Dynamic Scheduling of Parallel Real-time Jobs by Modelling Spare Capabilities in Heterogeneous Clusters”. *Proceedings of the IEEE International Conference on Cluster Computing* (CLUSTER’03).2003.

DSCA: A Coarse-Grain NUCA For CMP System*

Song Hao[1], Zhihui Du[1], Man Wang[1], Zhiqiang Liu[2]
[1]Department of Computer Science and Technology, Tsinghua University, 100084, Beijing, China,
[2]Hebei University, 071002，Baoding, China
Email: haos06@mails.tsinghua.edu.cn

ABSTRACT

Chip Multiprocessors （CMPs） will have more cores and larger on-chip cache capacity in the future. The design of on-chip cache hierarchy has significant impact on processor's performance. Non-Uniform Cache Architecture (NUCA) is a shared cache architecture supporting data migration among cache banks according to access frequency.

In this paper, we propose a coarse-grain NUCA — Distributed Shared Cache Architecture (DSCA) . In this architecture we separate the shared L2 cache into slices which are connected by on-chip networks. A cache-to-cache mechanism is utilized to support data sharing among slices, and an improved R-R (Round-Robin) scheduling algorithm is adopted to exploit data reuse in a TLP (thread level parallelism) program. According to the experiment, DSCA provides a reduction in average memory access latency of up to 23% over traditional "Dance Hall" design.

Keywords: Chip Multiprocessor (CMP), Distributed Shared Cache Architecture (DSCA), Thread Level Parallelism (TLP), Data Reuse

1. INTRODUCTION

Chip multiprocessor has become a main direction of processor design both for academy and industry. Small-scale CMPs, with two or four cores per chip, are already commercially available [1, 2]. With the transistor densities increase, more and more cores will be integrated in a single chip [3, 4]. At the same time, on-chip memory hierarchies are summoning innovative designs. The increasing clock frequency makes the costs of off-chip misses more disastrous to the system performance.

Most of the CMPs available on the market use the private L1 cache structure. But the proposals for the organization of the on-chip L2 cache are different. A shared L2 cache architecture was widely used now, but with the growth of the number of cores and the size of L2 cache, the average L2 cache access latency is heavily influenced by the latency of accessing remote cache banks, which in turn is influenced by on-chip wire delays. Private L2 cache has the advantage that most L1 misses can be handled locally, which could reduce remote on-chip L2 cache access, but this will result in many more off-chip accesses than a shared L2 cache.

Wire delay plays a significant role in cache design. For example, in the 65-nm technology (2004), transmitting a data 1 cm requires only 2-3 cycles, but in the 32-nm technology (which will be achieved around 2010 according to the Moore Law), this will necessitate over 12 cycles [5]. So the data in the shared cache should be placed close to the core(s), which is using the data, to minimizing the access latency. This makes it difficult to provide uniform access latencies to all the shared cache banks. A Non-Uniform Cache Architecture (NUCA) was provided in [6], which allow nearer cache banks to have lower access latencies than further banks. This idea could hide the wire delay effectively. As a development of NUCA, Jichuan Chang and Gutindar S. Sohi present the Cooperative Caching (CC) [7], which could achieve the benefits of both private and shared cache designs. But the shortage of CC is its lack of support to TLP which could improve the performance of CMPs significantly.

CMP is attractive for thread-level parallelism (TLP) which can make full use of computing resource and provide higher throughput while reducing energy consumption. TLP will bring more locality than before, which will also direct the design of on-chip cache hierarchy. In a TLP program, the data and code will be highly reused among different threads, so place them in a shared cache will be an optimal choice.

In this paper, we present a Distributed-Shared Cache Architecture (DSCA), which is more suitable for future CMPs and could effectively support the TLP while keeping the benefits of both private and shared cache designs. The DSCA divides a whole shared cache into several slices which are connected by on-chip networks, and each slice is shared by 4(or 8) cores directly. In such architecture, the threads of the same process could be scheduled to cores which are directly sharing the same cache slice, and this will increase the reuse of data and code. So we improved the R-R scheduling algorithm which is widely used for CMP task scheduling. At the same time, because the area of each slice is much smaller than a centralized cache, the wire delay could be reduced. To minimize the off-chip misses, we import some mechanisms used by CC to realize cache-to-cache data transfer.

The rest of the paper is organized as follows. In section 2 we explain the Distributed-Shared Cache Architecture and the policy used to support TLP while sustaining the capacity advantages of shared caches. Section 3 covers our experiment and the result. Related work is discussed in Section 4 and the conclusion and future work are presented in section 5.

2. DISTRIBUTED SHARED CACHE ARCHITECTURE

This section describes the baseline design for the architecture and how we support the TLP while achieving the benefits of both private cache and shared cache.

2.1 Architecture Overview

Most CMPs available in the market adopt a centralized structure to organize the on-chip shared L2 cache. We call this kind of architecture the "Dance Hall" CMP. Fig.1 depicts a 64-core "Dance Hall" CMP, the shared L2 cache is aggregated together, and the processor cores access the L2 cache through on-chip networks. The main shortcoming of this kind of

* This paper is partly supported by National Natural Science Foundation of China(No.60503039 and No.10778604), Beijing Natural Sciences Foundation(No.4042018) and China's National Fundamental Research 973 Program (No. 2004CB217903)

organization of on-chip cache comes from the wire delay. That means the latency of a processor core accessing a cache bank physically near it is much shorter than the latency of accessing a bank far from it. This results in two disadvantages. First, when the capacity of on-chip cache grows, the average access latency grows, too. So it may become another speed bottle net of System; Secondly, stabilization of data access latency couldn't be guaranteed. This will bring negative influence on the pipeline.

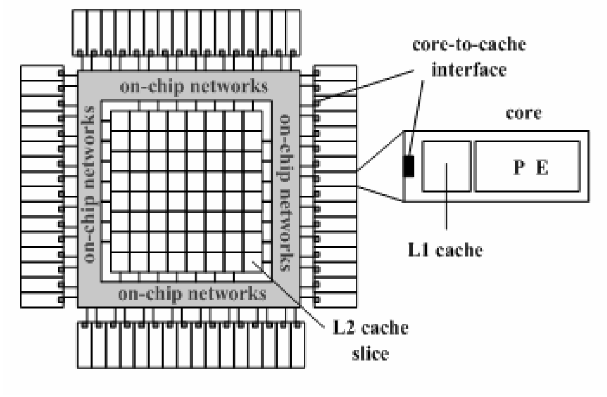


Fig.1. A 64-core "Dance-Hall" CMP

Being different from the "Dance Hall" CMP, the DSCA adopts a distributed structure. Fig.2 depicts the DSCA concept for a 64-core CMP. The on-chip L2 cache is not centralized, but divided into small slices, and each slice is shared directly by 4 processor cores. The detailed organization of the 4 cores and the small cache slice is not discussed in this paper, but as a part of future work. The small cache slices are connected by on-chip networks (such as a mesh shown in fig.2). The advantage of such organization is that, wire delay is hidden, because the size of each cache slice is much smaller than the centralized cache. On the other hand, the L2 capacity seen by each processor core is the sum of all the small cache slices. Because the on chip networks connecting the small slices could achieve the cache-to-cache data transfer.

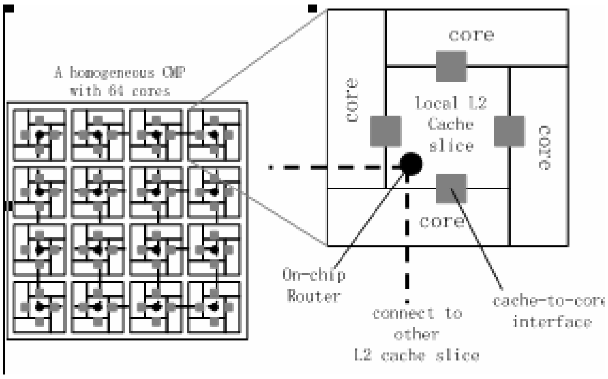


Fig.2. A 64-core DSCA CMP

2.2 Shared Scheme

As a shared design, all of the L2 slices are managed as a shared L2 cache with a uniform address space. The shared design is used by a number of existing CMP designs [8, 9, 10, 11], where several processor cores share a banked L2 cache. Fig.3 shows the flow chart of an L2 access transaction. In the uniform address space design, a global cache directory is adopted to record the blocks available in the L2 cache and its location. When a L1 cache miss happened, the fetch request is processed by the local L2 cache slice. On a local L2 miss, a request is sent to the Global L2 Cache directory to check if the data is available in a remote on-chip L2 cache slice. And if it does, a cache-to-cache transfer is initiated. Else, an off-chip data access transaction is initiated.

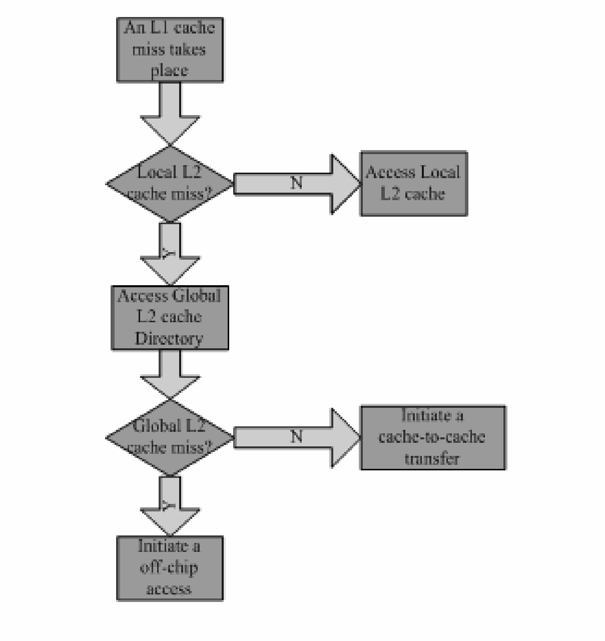

An L1 cache miss takes place
Local L2 cache miss?
N
Access Local L2 cache
Y
Access Global L2 cache Directory
Global L2 cache miss?
N
Initiate a cache-to-cache transfer
Y
Initiate a off-chip access

Fig.3. Flow Chart of L2 Cache Access

Latency to the cache-to-cache transfer varies according to network congestion and the number of network hops between the requesting processor core and the home tile.

2.3 Scheduling Optimization

Most existing CMPs adopt a simple Round-Robin algorithm to schedule tasks, because it is simple in hardware implement. A lot of other algorithms are also studied, such as DAG algorithm [14][15], DEFF algorithm [16] for real-time CMP.

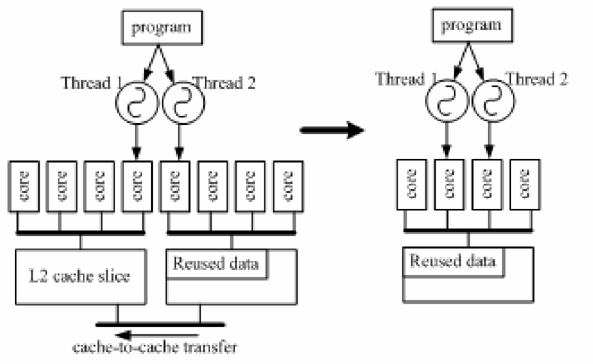


Fig.4. Scheduling Optimization

In the DSCA CMP, we claim that the threads derived from the same task should be scheduled on the processor cores sharing the same L2 local cache. Data and code are highly reused among the threads in a multi-thread program. And data reuse has significant impact on the performance and processor efficiency. As shown in fig.4, if a program with two threads is scheduled onto cores without sharing the same local L2 cache slice, the reused data has to be transferred through the on-chip networks. But if we schedule the two threads onto cores sharing the same local L2 cache slice, the cache-to-cache transfer latency can be discounted.

This idea may guide us to improve the scheduling algorithm. In our experiment, we compared an improved R-R algorithm and a traditional R-R algorithm. The result shows that the improved R-R gains 10% achievement on average memory access latency and 5% performance improvement.

3. EXPERIMENT AND RESULT

We evaluate the performance impact of the architectural extension and policies described above using a simulation program which is written all by us in C language. The processor cores, cache hierarchy, interconnection network, memory subsystem are modeled in detail (Table 1 shows important system parameters and contentionless access latencies).

Table 1. System Parameters

Processors	64, single-thread in-order processor
pipeline depth	20
cache line size	64B
L1 I-cache size/Associativity	16KB/16-way
L1 D-cache size/Associativity	16KB/16way
L1 load to use latency	1 cycle
L2 cache size/Associativity	1MB/16way
L2 load to use latency	DSCA: 6 cycles ("Dance Hall": 24cycles)
cache-to-cache transfer latency	36 cycles
External memory access latency	128cycles

To present a clearer picture of memory system behavior, we use a simple single-thread in-order processor model and focus on the average raw memory latency seen by each memory request.

We assume that wire delay has significant influence on cache latency, so in the DSCA using a local small L2 cache, the latency is smaller than that in the "Dance Hall" architecture using a big L2 cache.

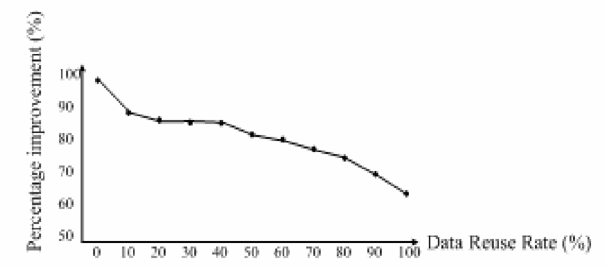


Fig.5. Percentage of Off-chip Access Frequency when data reuse is counted in Relative to the frequency when no data is reused

In a multi-thread program, data reuse among different threads is frequent and is a significant performance factor for the data access latency. We know that off-chip data access has great negative influence on total data access latency, so we ran experiments to understand how the off-chip access frequency changes along with the reuse rate among threads. As fig.5 shows, with the increase of the data reuse rate, the percentage of off-chip access frequency relative to the frequency when no data is reused decreases up to over 30%.

To compare the performances of DSCA and the baseline "Dance Hall" CMP, 2000 applications were fed into the simulating program. As fig.6 shows, with the data reuse rate increase, the average data access latency decreases. This is because, off-chip data access has significant influence on average data access latency, and increasing the data reuse rate can reduce off-chip data access frequency (shown above), as a result increasing data reuse rate among threads will reduce average data access latency.

Also we can see from fig.6 that, in DSCA CMP, the latency is much smaller than in the "Dance Hall" CMP. For in the DSCA CMP, most L2 cache access is taken placed in local L2 cache. The local L2 cache latency is smaller than the centralized L2 cache. The experiment shows, in DSCA CMP, average data access latency is over 23% smaller than in the "Dance Hall" CMP. On the other hand, we can see that the improved R-R scheduling could reduce data access latency by average 10%, because in the DSCA CMP with traditional R-R scheduling, the cache-to-cache transfer is much more than the improved R-R DSCA CMP.

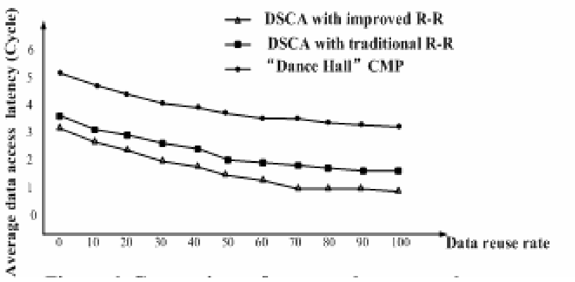


Fig.6. Comparision of Average Data Access Latency

Fig.7 shows the comparison of runtime. In the simple in-order processor, the operand fetching and write back stage are the bottle net of the pipeline. So reducing data access latency can reduce total runtime significantly.

In order to present a clear comparison of runtime between the DSCA and "Dance Hall" CMP, we assume the total runtime of the applications in the "Dance Hall" CMP without data reuse is 1. The result of experiments is shown in fig.7. In each of the three architectures, with the increase of data reuse rate, the total runtime decreased by up to 30%.

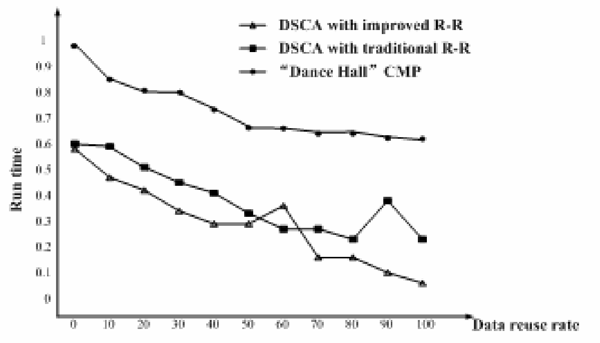


Fig.7. Comparision of Runtime

At the same time, we can find that the total runtime on the DSCA CMP is much shorter than the "Dance Hall" CMP. According to the experiment, the DSCA CMP is over 30% faster than the "Dance Hall" CMP on average. And also, the improved R-R could improve the performance by 5% on average.

The processor efficiency is compared too. The result is shown in fig.8. We can find similar improvement with above.

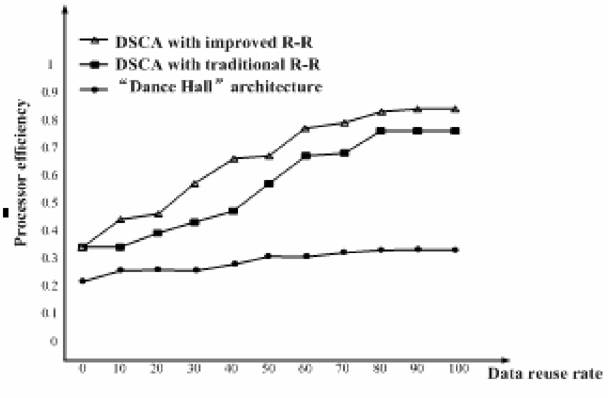


Fig.8. Comparision of Processor Efficiency

4. RELATED WORK

There are two impulsions to the DSCA. First, the increasing wire delay makes the physical position of data in the cache very important to the access latency. If a core in the CMP wants to use the data far from it, the access latency will be significantly high compared to the data near the core.

The NUCA [6, 8] was first presented to hide the wire delay. In the NUCA, the cache was divided to several banks, if a core wants to use data at a far bank, the data will be provided to it and at the same time, the data will be moved to a bank near the core for the next access.

As a development, Huh et al. [9] design a CMP cache to support a spectrum of sharing degrees, denoting the number of processors sharing a pool of their local L2 banks. The average access latency can be decreased by partitioning the aggregate on-chip cache into disjoint pools, to fit the run application's capacity requirement and sharing patterns. DSCA is similar in partitioning the aggregate cache into disjoint pools, but achieves it through dividing the cache into slices which is more suitable for the growing number of cores in a single chip.

Lastly, CC uses a private cache based architecture which can reduce the number of expensive cross-chip and off-chip accesses.

The second impulsion is that, the CMPs provide support for TLP [10]. In the future, there will be more and more parallel codes in a program to make full use of the computing resource in a CMP, which will result in more and more reuse of data. So organizing the threads to share the same cache slice directly will help the reuse of data preventing remote cache accesses.

The TLP has been applied since CMP was born. The IBM Power5 [1] can process 4 threads currently. The Sun Niagara [3] is a 32-way multithread processor. And in the cell processor [4], 10 threads could be processed simultaneously. A lot of TLP technologies have been discussed to be used on these chips.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have examined both architectural and policy decisions regarding the use of a Distributed Shared Cache Architecture (DSCA) in the context of a chip multiprocessor. To hide the wire delay, we divide the L2 cache into small slices, and the data is placed in the slice which is physically neighbored to its users. Because the area of each slice is smaller than the "Dance Hall" design, the wire delay is reduced, which will in turn reduce the average access latency. At the same time, a cache-to-cache transfer mechanism is adopted to increase the data sharing among slices in order to reduce off-chip data access frequency.

The trend of placing more cores on a single CMP will result in more TLP applications. In a TLP program, the data and code will be highly reused among threads. According to this condition, we realized an improved R-R scheduling in the experiment. Its main idea is to schedule the threads of the same program on the processor cores directly sharing the same cache slice. In virtue of such scheduling, cache-to-cache transfer, which will increase L2 access latency, can be reduced.

According the experiment, DSCA provides a reduction in average memory access latency of up to 23% over traditional "Dance Hall" design.

Currently, we are investigating the on-chip networks connecting cache slices in order to reduce cache-to-cache transfer latency, including topology, protocol, on-chip router, and routing algorithm.

REFERENCES

[1] Ron Kalla, B. Sinharoy, and Joel M. Tendler, "IBM Power5 Chip: A Dual-Core Multithreaded Processor. Micro", *IEEE Volume 24,* Issue 2, Mar-Apr 2004 Page(s):40 - 47

[2] C. McNairy, and R. Bhatia. Montecito, "A Dual-Core, Dual-Thread Itanium Processor", *Micro, IEEE Volume 25,* Issue 2, March-April 2005 Page(s):10 - 20

[3] P. Congetira, K. Aingaran, and K. Olukotum. Niagara, "A 32-Way Multithreaded Sparc Processor", *Micro, IEEE Volume 25,* Issue 2, March-April 2005 Page(s):21 - 29

[4] D. Pham, S. Asaon, M. Bolliger and etc, "The Design and Implementation of a First-Generation CELL Processor. Solid-State Circuits Conference, 2005. Digest of Technical Papers", *ISSCC. 2005 IEEE International Volume 1,* 6-10 Feb. 2005 Page(s):184 - 592

[5] B. M. Bechmann, and D. A. Wood, "Managing Wire Delay in Large Chip-Multiprocessor Caches. Microarchitecture", 2004. MICRO-37 2004. 37th *International Symposium on 04-08 Dec.* 2004 Page(s):319 - 330

[6] C. Kim, D. Burger, S. W. Kechler, "Nonuniform Cache Architectures for Wire-Delay Dominated On-Chip Caches. Micro", *IEEE Volume 23,* Issue 6, Nov.-Dec. 2003 Page(s):99 - 107

[7] J. Chang, and G. S. Sohi, "Cooperative Caching for Chip Multiprocessors", *Computer Architecture*, 2006. *ISCA '06. 33rd International Symposium on 2006* Page(s):264 - 276

[8] C. Kim, D. Burger, S. W. Kechler, "An Adaptive, Non-Uniform Cache Structure for Wire-delay Dominated On-Chip Caches", *Proceedings of the 10th international conference on Architectural support for programming languages and operating systems*, volume 5, issue 30, Dec. 2002, Page(s): 211-222

[9] Michael Zhang, Krste Asanovic, "Victim Migration: Dynamically Adapting Between Private and Shared CMP Caches", *Technical Report of Computer Science*

and Artificial Intelligence Laboratory, MIT. Oct. 10, 2005.

[10] Michael Zhang, Krste Asanovic, “Victim Replication: Maximizing Capacity while Hiding Wire Delay in Tiled Chip Multiprocessor”, *Computer Architecture*, 2005. *ISCA '05. Proceedings. 32nd International Symposium* on Jun. 4-8 2005 Page(s):336 - 345

[11] Evan Speight, Hazim Shafi, Lixin Zhang, Ram Rajamony, “Adaptive Mechanisms and Policies for Managing Cache Hierarchies in Chip Multiprocessors”, *Computer Architecture*, 2005. *ISCA '05. Proceedings. 32nd International Symposium* on Jun. 4-8 2005 Page(s):346 – 356

[12] J. Huh, C. Kim, H. Shafi, L. Zhang, D. Burger, and S. W. Kechler, “A NUCA substrate for flexible CMP cache sharing”, In the 19th *ICS*, Jun. 2005, page(s): 31-40.

[13] L. Spracklen, and S. G. Abraham, “Chip Multithreading: Opportunities and Challenges”, *High-Performance Computer Architecture*, 2005. HPCA-11. 11th International Symposium on Feb. 12-16 2005 Page(s):248 – 252

[14] T.D. Braun, et al, “A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems”, *Journal of Parallel and Distributed Computing*, 61:810-837. 2001

[15] J. Blythe, et al, “Task scheduling strategies for workflow-based applications in grids”, *Cluster Computing and the Grid*, 2005. *IEEE International Symposium on Volume 2*, Page(s):759 – 76. 2005.

[16] D. Ma, W. Zhang, Q. Li, “Dynamic Scheduling Algorithm for Parallel Real-time Jobs in Heterogeneous System”, *Proceedings of the Fourth International Conference on Computer and Information Technology (CIT’04)*,2004.

Song Hao is a Master Degree student of the Department of Computer Science and Technology, Tsinghua University. He graduated from Nankai University in 2005 with specialty of Computer Science and Technology. He is awarded the scholarship twice for his achievement in study and research, the scholarship is awarded to the Top 10 student in each department; He is awarded the honour of outstanding student in academic, moral and health, and the honour of excellent leader to commend his work in the student council. His research interests are in Parallel Computing and High Performance Computing, especially in Chip Multiprocessor technology.

Application and Realization of RFID in Auto Theftproof System Based on MC9S12D64

Chunnian Zeng, Shuanghua Li, Hongmei Huang
School of Automation, Wuhan University of Technology
Wuhan, Hubei, 430070, P.R.China
E-mail: waishi@whut.edu.cn

ABSTRACT

The radio frequency identification device based on Multi Protocol Transceiver IC S6700 is introduced in this paper, and the working principle of RFID applied in auto theftproof system as well as the design of software and hardware of auto theftproof system combining with the 16-bit microcontroller unit MC9S12D64 of Motorola automotive electronics MCU have been introduced emphatically.

Keywords: Multi Protocol Transceiver, S6700, RFID, Theftproof System

1. INTRODUCTION

RFID (Radio Frequency Identification) is a self-identification technology that became mature from the 1980's. It uses Radio Frequency to realize intercommunication by noncontact way. RFID can identify high speed mobile and multi-card at one time and with convenient operation, at the same time, with the characteristics of unafraid of bad environment as oil stain or dust contamination, it is suitable for realizing the system automatization and not easy to be damaged. The RFID system introduced in this paper is a successful experiment in the application of auto theftproof base on MCU.

2. THE SUMMARIZATION OF RFID AUTO THEFTPROOF SYSTEM

Along with the development of technology, the auto theftproof device is becoming more and more strict and perfect. The auto theftproof device can be currently divided into four major kinds according to its configuration and function, which are mechanical style, electronic style, chip style and network style and each of them has its own advantages and disadvantages, however, the development trend of the auto theftproof device is towards to the chip style and network style which are more intelligent.

RFID auto theftproof system that belongs to a chip style system is the new application of RFID. Since the enough small tag was developed, it can be integrated into the auto key with given code. The read-device is fixed below the steering wheel and the distance between the reader and the key must be less than seven centimeters, while the key inserting and switching to “M”, the auto theftproof system goes into operation, and the reader gets valid UID, then the system self-acting will open the engine computer and tell the key is valid by the voice system, otherwise, it will give an alarm by voice and close the engine computer, while the EMS will close the oil access and the engine, and in this way, to realize the theftproof function.

3. THE SYSTEM BUILDUP AND WORKING PRINCIPLE OF RFID

The RFID system is the core part of the auto theftproof system. Generally, It is made up of tag, reader and radio frequency antenna. The tag is made up of coupling element and chip, which contains inner-antenna used for the communication with the radio frequency antenna, and the reader is used to gain the information of the tag, while the antenna is used to transfer radio frequency information between tag and reader. The general working principle is that: the reader transmits certain frequency radio-frequency signal through the radio frequency antenna, and when the radio frequency card enters the radio frequency antenna working space, it produces the induced current and the radio frequency card obtains energy to be activated; then, the radio frequency card transmits the information as its own code, etc. through its inner-antenna; and when the radio frequency antenna receives the carrier signal transmitted from the radio frequency card, it transmits the signal to the reader through the regulator, the reader demodulates and the decodes to the received signal and then delivers it to the backstage host system for correlative processing; the host system will judge the validity of the card according to the logic operation, and make corresponding processing and control in view of different settings, and then send out the command signal to control the movement of executing agency.

4. THE HARDWARE DESIGN OF RFID AUTO THEFTPROOF SYSTEM

The RFID auto theftproof system takes the RFID system as its core composition, the auto theftproof system hardware control unit selects 16-bit microcontroller unit MC9S12D64, and the radio frequency identification system is made up of reader S6700, responder TAG-IT and the radio frequency antenna, besides, the system also includes the storage circuit (AT24C01), detecting circuit, sound circuit and the CAN bus communication circuit. The diagram of hardware designing of RFID auto theftproof system is shown in Fig.1.

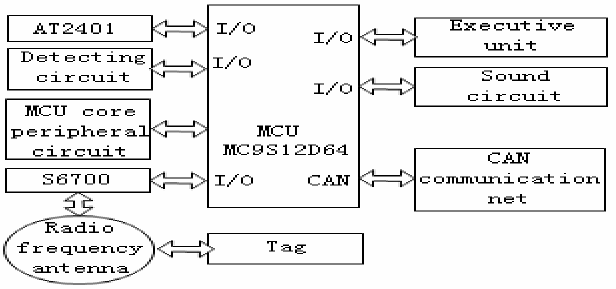


Fig.1. ardware frame of RFID theftproof system

(1) The control unit microcontroller MC9S12D64 that inherits the fine tradition of Freescale semiconductor applied in the field of auto microcontroller, is the member of MC9S12 series with the core of the quicker speed S12 (Star Core), and the pin is compatible, the memorizer can

be upgraded, besides, there are a lot of peripheral equipment for choice inside the microcontroller. MC9S12D64 altogether has 8 kind of working patterns, the setting of patterns is realized through the BKGD, MODB and MODA three pins' status gathered during the reset periods [2][3]. This strengthens the selectivity of application.

(2) S6700 IC card read-write multi-protocol transceiver and responder TAG-IT compose the radio frequency system, and S6700 using 13.56MHz as its operating frequency, has the mechanism of guarding against conflict, it uses the Manchester encoding method and can allow many cards read-write at the same time without conflict. The model transmission power is 200mW. The supporting protocols are as following: TI TAG-IT protocol, ISO/IEC15693-2 protocol and ISO/IEC14443-2 protocol. The interface between S6700 and the CPU is synchronized serial interface (SPI). And SCLOCK, DIND, OUT are respectively the clock line, the data input line and the data output line. The clock line is bidirectional, DOUT is used to output data during the period of accepting data, while used to express FIFO register situation of S6700 during the period of transmission data.

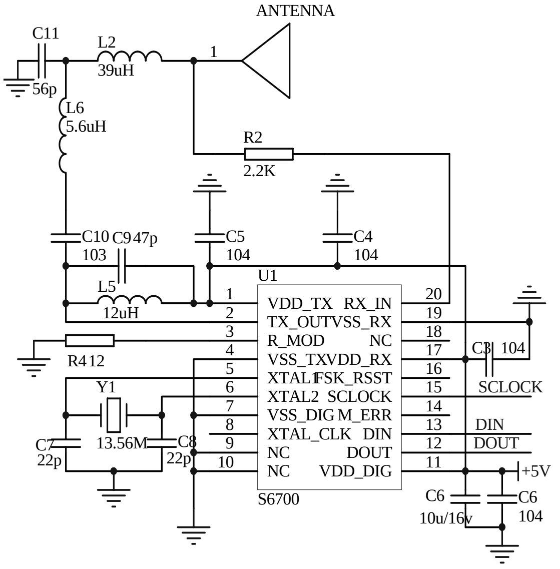


Fig.2. Circuit design of reader

Responder TAG-IT is completely compatible with the ISO/IEC15693 protocol. Inside the card, there are 64-bit UID (card serial number), 8-bit AFI (application identification serial number), and 8-bit DSFID (data storage form), among which the UID cannot be revised. Furthermore, there is 2Kbit EEPROM inside the card, which is divided into 64 blocks, each of them is 32Bit and can be locked, which can protect the data avoiding revision. The circuit design drawing of Radio frequency system reader is shown in Fig.2.

(3) AT24C01 has I2C bus and 1K bit EPROM with independent write-cycle (biggest 10ms), it can be programmed on-line with power on, and can preserve data for long-term when losing power, which can prevent the man-made destruction to auto power effectively. AT24C01 saves corresponding TAG-IT UID number, which is used to check with the responder's UID. The sound circuit takes the ISD5216 integration sound chip as its core. ISD5216 has the ability of recording and broadcasting and 4MB digital data memory function, which realizes multistage sound recording and broadcast combining with modems and power magnifying, in order to realize of the security and alarm function of RFID auto theftproof system easily. Detecting circuit is used to detect the diversified information of the automobile status, which includes door signal, power signal, braking signal, etc. MCU makes the judgment and decision according to the status information detected by detecting circuit, and controls turning lamps, power, gate magnetism locks and wheel hub locks through executing agency.

(4) (CAN communication network module is responsible for transmitting the start signal and the examination signal to the central processor of automobile through the CAN net and the central processor makes decision according to signal received. With flexible communications and strong anti-jamming ability, the CAN bus is widely applied in the automobile control system at present. CAN communication connection hardware design is shown in Fig.3, in which 82C250 is the interface between the CAN controller and the physical bus, and it uses P113 to separate with the CAN controller to enhance the system's anti-jamming ability.

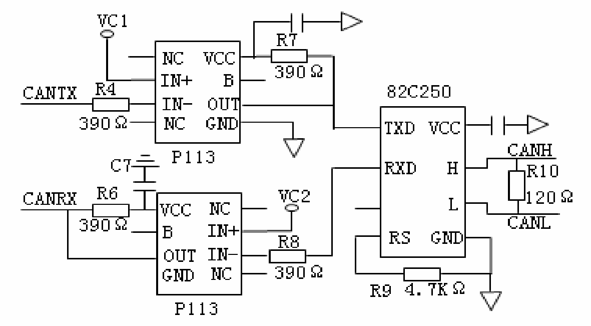


Fig.3. Circuit design of CAN communication interface

5. THE REALIZATION OF THE RFID AUTO THEFTPROOF SYSTEM SOFTWARE

The development environment of the RFID auto theftproof system software designs is Code Warrior for S12, which is a software package developed for the application of MCU facing to CPU with HC12 and S12, including integrated development environment IDE, processor expert data-base, entire chip simulation, visualization parameter demonstration tool, project manager, C across compiler, assembler, link and debugger. Its debugging way is the way of BDM (Background Debug Mode) which is a system debugging way of Freescale Corporation, and which has the basic debugging function, including the resources visit and the operating control, and can realize many important development function with the instruction license and the break point logic coordination.

5.1 S6700 working flow

Software design introduces emphatically the programming of S6700, which needs to follow its communication protocol and the working schedule strictly. S6700 has three kinds of operation patterns: ordinary pattern, register pattern and direct pattern. Under direct pattern, CPU must face radio-frequency signal processing directly, which is quite complex and is not used generally. Under ordinary pattern, each instruction includes parameters as protocol, modulation way, transmission speed, etc. But in register pattern, the series does not include

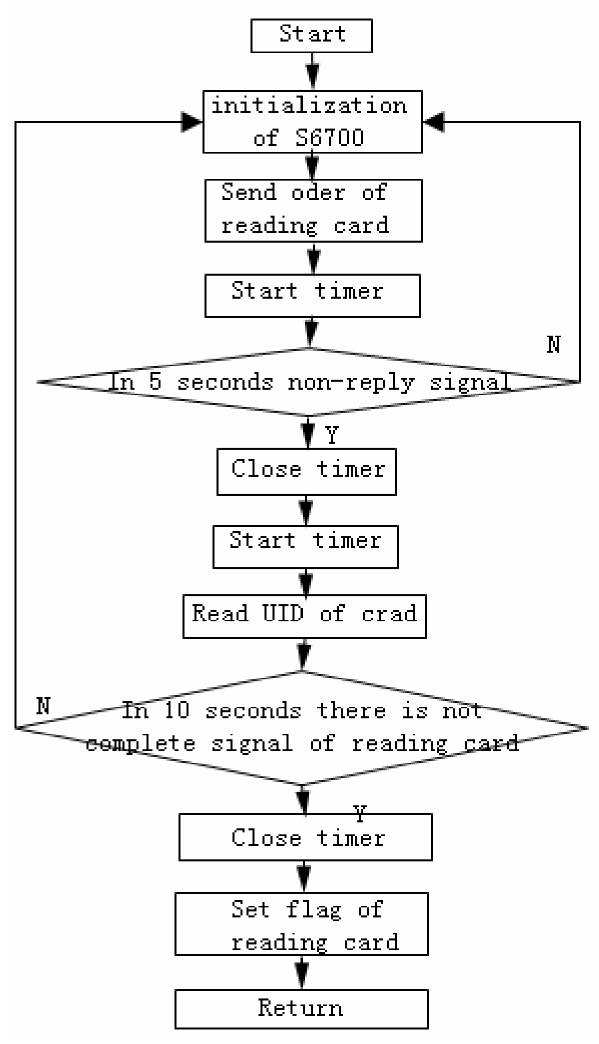


Fig.4. Working flow of S6700

these parameters, and it decides by the parameters written into the register in advance. The system selects the ordinary pattern for S6700 operation, under which, firstly, MCU must transmit the closure order to prevent the miscarriage of justice of reset pulse, and initialize time checks, then transmit order parameter of ordinary pattern, before TAG-IT replying, MCU must give up the clock line domination and delivers it to S6700, at the same time, waits for the reply signal of the responder, and after accepting the reply signal, MCU reads responder's UID and judges whether there are errors or not for the reading card , and finally the reply is over, MCU takes back clock line domination.

5.2 Initialization

All operations of radio frequency reply start from the S6700 initialization. First, the time register must be initialized during the communication process of MCU and the TAG-IT. According to ISO/ IEC15693 protocol, initialization time series S1 01111011 1000000011000 ES1 must be first written in. S1 and ES1are respectively for the start bit and the stop bit, and in ordinary pattern, the order byte is 8 bit, its transmission order is that the top bit is in front; whereas for data stream, the low bit is in front.

Order structure

Start bit S1	Order byte	Data stream （random bit）	Stop bit ES1

The start bit's wave is when SCLOCK at the high level, DIN has a rise pulse, and which can only happens after SCLOCK breaks to the high level of 300ns. The stop bit wave is when the SCLOCK breaks to the high level of 400ns, DIN will have a drop pulse.

5.3 Read responder UID

While MCU reading TAG-IT, S6700 holds clock line domination, and after it reads the data which it transmits to MCU through the pin of DIN. During the period of reading data, MCU must simulate response schedule of TAG-IT strictly, and confirms the correctness of the data through FLAG, only when FLAG is entirely accurate can continue to accept the response content, otherwise, the reading card process will be ended.

The TAG-IT response format is as the following order: start bit S2, FLAG, the response content, CRC16, stop bit ES2, its basic request of reading card and reply schedule are shown in Fig.6. TRAN1 and TRAN2 respectively are the MCU giving up clock and MCU obtaining the clock.

6. CONCLUSIONS

The problem of automobile theftproof attracts wide attention in the world, to solve this problem must start from the technology of high-tech theftproof, and the RFID auto theftproof system has the following merits: ① Using the radio frequency identification technology which can distinguish UID accurately and instantaneous complete the status recognition. ② The responder contains a unique UID number and digitized password, which makes extremely low rate of repeat code so as to enhance the security performance. ③Using MC9SD64 as the controller of the theftproof, which enhances the anti-jamming ability, and guarantees the normal operation of the security system. ④Using the CAN bus to realize the communication of central computer of the auto, which guarantees the fluency of the correspondence and enhances the anti-jamming ability of the RFID theftproof system.

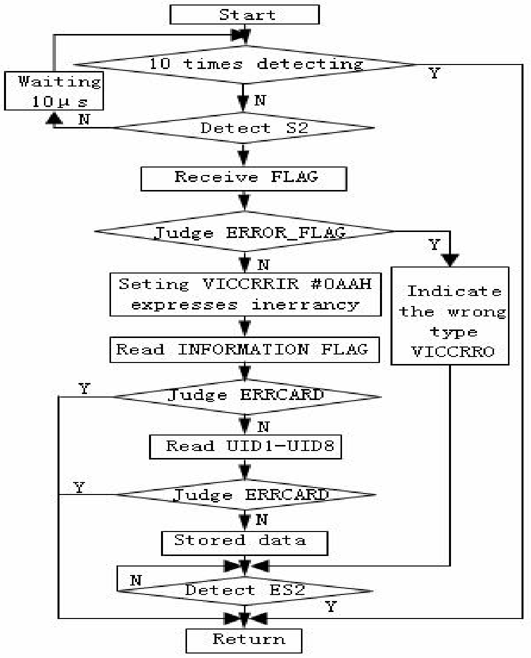


Fig.5. Subprogram of reading UID

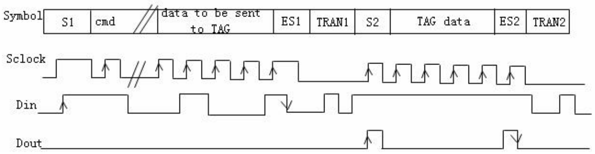


Fig.6. Schedule of read card/ responsion

REFERENCES

[1] Wang Aiying. Smart Card Technology—IC card (Secondedition)[M].Beijing: Qinhua University Press, 2000

[2] Shao Beibei. Online development of MCU embedded application [M]. Beijing: Qinhua University Press, 2005

[3] Liu Huiyin. Micro-controller MC68HC08 Principle and Embedded Application [M]. Beijing: Qinhua University Press, 2005

Low Power Design in VLSI*

Pengyong Ma, Shuming Chen
7th team, Computer School, National Unicersity of Defence Technology, ChangSha, 410073, P.R.China
Email: mpy9608@yahoo.com.cn

ABSTRACT

The chip's power-consume doesn't lower with the development of process. People pay more attention to the low power design. There are many methods in low power design. From top to down, there are low power design of architecture level, design level and material level. The mobile device request low power critically, this demand designer trying their best to reduce the system's power. It is difficult to reduce power consume largely with one method. None but adopting methods from architecture level to material level can reduce the system's power.

Keywords: Integrate Circuit, Power Consume, Low Power Design, System's Power, Architecture Level

1. INTRODUCTION

With the development of integrate circuit technology, more and more transistors are integrated on one chip and the chip's performance become higher. But the chip's area becomes larger and the frequency is higher than before. So the power consume becomes lager. It results in heat dissipation, chip package and other problems. Even more, it maybe leads circuit parameter drifts and reliability drop. Fig.1 shows the AMD series chip's power.

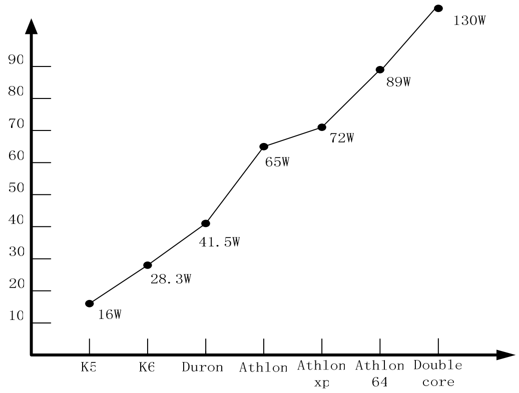


Fig.1. the AMD series chip's power

Compared with area optimization and speed optimization, low power design is a new field. But it is a hot research now. Today, the designers attach importance to low power. A distinct trend is that they pursue MIPS before, but now MIPS/mW. For example, many international conferences' theme is "Power-aware Systems" and so on.

2. ARCHITECTURE LEVEL LOW POWER DESIGN

With the development of integrate circuit and computer system, the low power become more and more important. This paper mostly discuss the low power design and the develop trend. There are many methods of low power design. From top to down, it is architecture level, design level material level and so on.

Architecture level low power design. The chips with difference architecture consume difference power and runtime when they complete a same task. Those chips with simple architecture and function unites need long runtime and the average power is low, but this doesn't means that the total power consume is low. On the other side, the chips with complex architecture and function unites need short runtime and the average power is high, but this doesn't means that the total power consume is high.

2.1 Cache Architecture

Now the cache takes up larger and larger proportion in chip, the cache power consume is an important parts of the whole power. The different architecture Cache has different miss rate in the same application, and the power is not same too. Table 1 is the miss rates of several different way-set caches in an application.

Table 1. cache miss rates of different architecture

capacity	8KB	8KB	8KB	8KB
way	1 way	2-way	4-way	8-way
Miss rate	0.046	0.038	0.035	0.029

As table 1 show, those cache sizes are 8KB. But the miss rate of 1-way cache is near upon 2 times of 2-way caches. Appropriate sizes and ways-set of cache can get low miss rate at the cost of little hardware. Low miss rate means short runtime and low power consume. Of course, high ways cache will cost large hardware.

2.2 Redundancy Structure

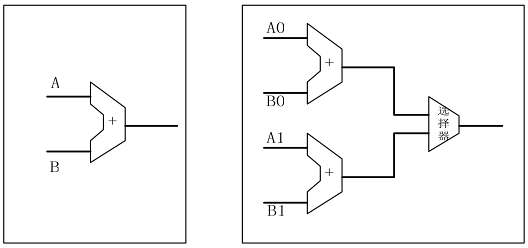


Fig.2. Redundancy Circuit

Some people propose redundancy structure to reduce power consume. Fig.2 shows that two adder can reduce clock frequency and power.

As Fig.2 shows, the left is one adder and two adder right. If they work at the same voltage, the right's power is more 2 times than left's. But if we reduce the voltage to 1/2 in right graph, the circuit speed will reduce to 2/1 too, but it can complete a task use same time of one adder. We know that the power is in direct proportion to with voltage square, so it will get low power too. For example, if we reduce the adder's voltage from 3v to 1.5v, by rule P＝CV2f, the power will reduce to about 1/2 compared with it of only one adder. This

* This work is supported by National Natural Science Foundation of China(No. 60473079) and the Research Fund for the Doctoral Program of Higher Education (No. 20059998026)

power decrease is at the cost of a additional adder hardware.

2.3 Asynchronous Circuits

Some researcher had proposed asynchronous circuit at begin of 1950s. Muller and Partky complete some theory in 1956. from that time on, the asynchronous circuit research does not interrupted, but they are only rest on theory, rarely product present on the market. With the rigor low-power require of mobile device, many people focus on asynchronous circuits. Now most chip use synchronous circuits, all transistors are turned at one or several clock rising edge. Clock consume about 15%～35% power of the whole chip. It includes clock create, PLL, driver, clock tree, register and other devices. These will be saved if we adopt asynchronous circuits. But the control circuits are too complex in asynchronous circuits, and the EDA tools can't support asynchronous circuits well. It is a large challenge to designer to design a very large chip. They usually divide a chip to several modules. Synchronous circuits are used in module but asynchronous circuits between modules. If a module is idle, its clock will be closed. This method may reduce whole chip's power.

2.4 Dynamic Voltage and Frequency Management

For example, in Montecito processor, Foxton technology attempts to tap unused power by dynamically adjusting processor voltage and frequency to ensure the highest frequency within temperature and power constraints. The 6-Mbyte Itanium 2, for example, consumes 130W for worst case code1 and has power limits that keep it from using a higher voltage, and thus a higher frequency or performance point. But low-power applications, such as enterprise or integer workloads, often consume only 107W. Consequently, 23W of power—and hence performance—remain untapped.

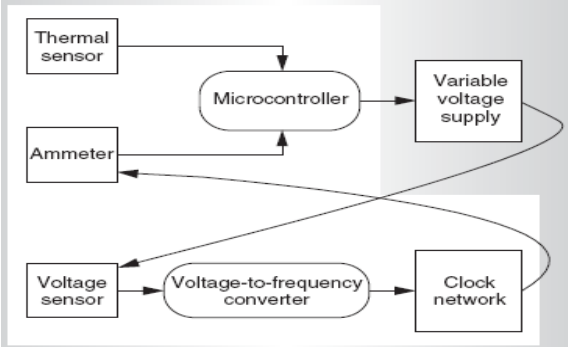


Fig.3. Foxton Technology in Montecito(dual-core itanium)

When the microcontroller detects a need, it will change the voltage it requests of the variable voltage supply in 12.5 mV increments. The voltage control loop responds within 100 ms.

3. DESIGM LEVEL POWER CONTROLS

3.1 Low Power Circuits on Uncritical Path

There are many methods to control power in design level, for example, we may select small and slow circuits which consume low power without influencing processor's performance. Now the chip is mostly synchronous circuits. It is always one or several paths limiting chip's frequency. Other paths are not the bottleneck of the chip, so we may select low power circuits on these uncritical paths. Then the chip's whole power will down.

As Fig.4 shows, the critical path of this module is 1 —> 2 —> 3 —> 4, the and-gate A1, A2 and the inverter I1 is not on the critical path, so these circuits may be low power circuits. Though these circuits sizes are small and they speeds are slow, but these don't influence the chip's frequency.

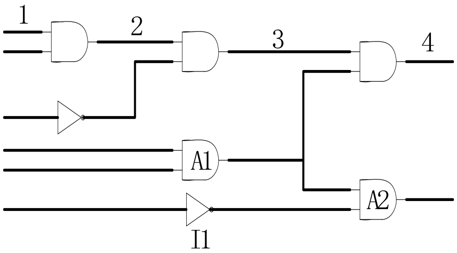


Fig.4. Low Power Circuits on Uncritical Paths

3.2 Gate Clock and Clock Negedge

Now the clock power consume take up large part of the whole chip's power. Gate-clock can control clock turn of inactive units to reduce power consume. But the designer must take care of gate-clock at the time of stopping and starting clock. Because of signal turning, it maybe brings glitches, and these glitches possibly result in false turn. In addition, we should adequately use clock negedge. Mostly chip only use clock rising edge to control circuit turn, the descend edge is waste. If the chip apply double-edge-triggered flip-flop, the clock frequency may reduce to 1/2 without reducing processor's performance.

3.3 Reducing Chip's Area

Reducing chip's areas can make the whole power down. So the designer should try their best to reduce the chip's area on the condition of place and route and the chip's cost decrease too. In addition, the finished product rate is in inverse proportion to with area power.

Chip finished product rate
= wafer finished product rate
× (1+fault density × chip area/a)-a.

If a is about 3, suppose that fault density is 0.8/cm2 and wafer finished product rate is 100%, if chip's area is from 1cm2 to 1.5cm2, the Chip finished product rate will decrease from 0.49 to 0.24.

Now designer always apply semi-custom ASIC. We may use full -custom ASIC in critical path and regular units, it will reduce area and power markedly. For example, we full custom design a 10 read ports and 6 write ports register file, the area and power reduce to 1/2.

3.4 Bus Coding

In placing and routing, the long line should route with high layer metals or top layer metal. Because the high layer metals' capacitance is low and this will reduce the power. The buses and FSM should use Gray code and other bus-code with lesser jump instead of ordinary binary code. For example, a 6bits bus jump from 31(6'b011111) to 32(6'b100000), all the six lines are turned. It will enlarge the power. If we use Gray code, there is only one line jumped and the power consume will be low. So when we design bus code, we should place the frequently jumping states together, thereby it will reduce the power. There are many bus coding that we may select, such as one-hot code, bus-invert code and fixed step change code, and so on.

On the side, dynamic logic can enhance the circuit's speed and lessen the transistors number, thereby it can reduce the power consume. But the dynamic logic's control is complex and the

EDA tools support it insufficiency.

Table 2. Binary code and Gray code

Binary code	Gray code	Decimal
0000	0000	0
0001	0001	1
0010	0011	2
0011	0010	3
0100	0110	4
0101	0111	5
0110	0101	6
0111	0100	7
1000	1100	8
1001	1101	9
1010	1111	10
1011	1110	11
1100	1010	12
1101	1011	13
1110	1001	14
1111	1000	15

4. MATERIAL LEVEL POWER CONTROLS

The power of VLSI and is close correlation with material and technology. New material and advanced technology can enhance the processor's performance greatly. For instance, TI Co. produced the 64Mbits Ferroelectric Random Access Memory (FRAM). It adopt 0.13um copper technology and 1.5V Supply Voltage. One cell's area is only 0.54um2, on the other side, it is 1.95um2 before. The FRAM needn't periodic refresh, it will reduce the memory's power greatly.

Double or more threshold voltages can control power hugely. With the development of technology, the transistor's size becomes smaller and smaller. The circuit speed becomes high, but the leak current becomes high too. It takes mostly power of the whole chip. Double or more threshold voltages can control leak current. We can use high threshold voltage transistors on uncritical paths, then the leak will become low.

Other new material can achieve low power too. For example, BiCOMS has the CMOS and bipolar integrated circuits' advantage. It's power is low, but the technology is complex and the cost is high. GaAs has low power and high speed characteristic too.

SOI is a new technology in the last years. Now CMOS technology produce P and N transistors on silicon wafer, when signal change, the charge near the source and drain must be exhausted, it will consume large power. If the chip adopt SOI technology, the transistors is isolated by oxide layer. The transistor's switch speed is high and it may use low threshold voltage. The capacitance may reduce 30% if SOI is adopted, so the chip's power maybe reduced largely.

5. CONCLUSIONS

With the development of transistor integration, larger and larger heat is produced by per unit area. It increases the cost of chip encapsulation. The mobile device request low power critically, this demand designer trying their best to reduce the system's power. It is difficult to reduce power consume largely with one method. None but adopting methods from architecture level to material level can reduce the system's power.

REFERENCES

[1] LUCA BENINI, System-Level Power Optimization, ACM Press, 2000

[2] Ansgar Stammermann, System Level Optimization and Design Space Exploration for Low Power，ACM Press, 2001

[3] Yanbo Wang, VLSI low power design analyze, electronic device, 2006.6, p152-p156

[4] Ying Yu, VLSI Design and Implementation of A Low Power Microcontroller Using Asynchronous Logic, Chinese Journal of Semiconductors, 2001.10, p1346-p1351

[5] Yuanqi Zhang, The Research of the Technology and Method Reducing Power Consumption in High-performance Microprocessors, Computer Engineering and Applications, 2002. 11, p54-56

[6] Zhonghe Guo, Design of CMOS double-edge- triggered flip-flop, Semiconductor Technology, 2003.4, p65-p67.

[7] Chengxi Zhang, computer architecture, High Education Press. BeiJing, 2005.6

[8] Richard A. Hankins, Gautham N. Chinya, Multiple Instruction Stream Processor, ISCA2006

[9] Cameron McNairy;micro; MONTECITO: A Dual-Core, DUAL-THREAD ITANIUM PROCESSOR, Published by the *IEEE Computer Society*,2005

USB Chip CH375's Principle and Application in Vehicle's Black Box

Chunnian Zeng, Jie Zhang, Hongmei Huang
School of Information Engineering, Wuhan University of Technology
Wuhan, Hubei, 430070, P. R. China
E-mail: waishi@whut.edu.cn

ABSTRACT

The article introduces the function and the operating principle of chip CH375 in details. In addition, it includes a design scheme about embedded USB host system for Vehicle's Black Box, and specifically analyzes the correlative protocol, the design method of hardware interface circuit, and software design of the embedded USB host system. By using USB mobile storing equipment, the mass storing of MCU system can be realized, and also the prompt data exchange between MCU and the computer system can be realized.

Keywords: CH375, USB Host, U Disk, Vehicle's Black Box

1. INTRODUCTION

With the rapid development of computer technology, using USB moving storage device is very common nowadays. Therefore, some USB interface chips are produced one after another which are used in some equipment or instruments needing to store data with USB moving storage device. The Black Box of vehicle is the kind of intelligent equipment, using to monitor, record and store vehicle's various data in the moving state. Since the massive real-time online data need immediate transmission and storing, and USB interface has the advantage of convenience and high speed of data transmission, the USB interface chips are widely used. As a general interface chip of USB bus, CH375 can support the HOST and SLAVE mode.

2. THE OPERATING PRINCIPLE OF CH375

2.1 Internal Structure

CH375 is the SOP28 seal chip. The chip's internal part integrates PLL frequency multiplier, host/slave USB interface SIE, data buffer, passive parallel interface, asynchronous serial interface, order interpreter, protocol processor for controlling transmission and general firmware software. The chip's pin chart is shown in Fig.1.

CH375 has 8-bit data bus and reading, writing and chip selection controlling line as well as interrupt output, which can be conveniently integrated with system bus of MCU. Under the host way, CH375 has also provided serial communication ways, and connect MCU by serial input, serial output and interrupt output. CH375's UBS host can support all kinds of full speed devices, and the exterior MCU can communicate with USB device through CH375 according to the corresponding USB protocol. Under the slave way, conforming to USB specification, CH375 has full speed device interface and contains USB communication protocol of bottom layer. It has two ways of application and development, such as, simple built-in firmware mode and flexible external firmware mode; it can also support 5V and 3.3V power voltage. Besides, CH375A chip also supports low power consumption mode.

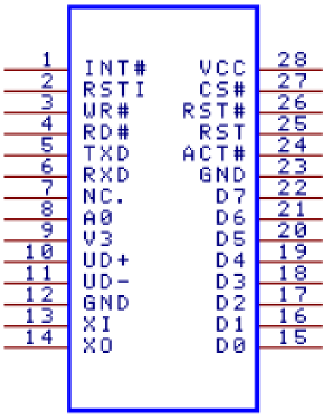


Fig.1. CH375's pin chart

The interior of CH375 chip has 7 physical endpoints. CH375 endpoint 0, 1, 2 can only be suitable for USB device mode. Under host way, it can support all kinds of common full-speed device of USB. The USB device endpoint numbers may be 0 to 15. Two directions support 31 endpoints at most. The USB device packet length may be 0 to 64 bytes.

CH375 built-in firmware can process the communication protocol of mass storage device and requests USB storage device to support Bulk-only transmission protocol.

2.2 Software Interface

AS to application of USB storage device, CH375 provides reading and writing interface of data block directly, and takes physical area of 512 bytes as basic reading and writing unit so as to simplify USB storage device to the kind of exterior data storage equipment. Therefore, MCU can freely read and write data stored in USB, and can also define its data structure freely.

CH375 provides file interface of USB storage device by the sub-procedure database of C Language. These application-programming interfaces (API) include usual file operations, and can be transplanted and embedded into various common single chip procedures.

The sub-procedure of U disk file database of CH375 has the following characteristics: supporting common FAT12, FAT16 and FAT32 file system, reaching more than 100 GBs of disk capacity, supporting multilevel subdirectory, supporting capital letter of file name of 8.3 formats, supporting document's opening, setting up, deleting, reading and writing and searching etc.

The sub-procedure database has two reading/writing modes to U disk file, which are physical area mode and bytes mode. Under the mode of physical area (usually of 512 bytes each area), the sub-procedure operates U disk reading/writing, which uses physical area as basic unit. The speed of operating is comparatively fast. However, it is usual that the extra file data buffer area is needed, and the data buffer area must be integral time of the length 512 of physical area. Therefore, the mode of physical area is suitable for the single chip procedure that has a big number of RAMs, data, and frequently reading/writing data. Under the mode of bytes, the sub-procedure database takes bytes as basic unit to read/write files of U disk, and the speed is slower which makes no need of extra file data buffer area so that it is convenient for using, and can be available to most kind of single chip procedures. Whenever constructing a new file or opening a file, the database acquiesces it as physical area and supports the area as the basic unit to read/write. After executing an order each time, which takes the bytes as basic unit, the database will come into the mode of bytes automatically. The system adopts the mode of physical area while reading/writing U file.

The file interface of API sub-procedure of CH375 needs nearly 600 bytes RAM as buffer area. All of the APIs have operation state to return after using, but do not necessarily to have answering data.

3. FILE SYSTEM

Since many systems will finally exchange data with personal computers with windows operating system, the data in the U disk should match the format of the Windows file system in order to facilitate the data exchanges. If the embedded system needs to conform USB storage device into a file system, it can use the API interface provided by CH375 file subroutine library, which will process the file system. The U disk's subprogram library of CH375 supports FAT12, FAT16 and the FAT32 file system, and supports the biggest capacity of 100 GBs. The MCU needs not to consider the file system and only needs to know the basic knowledge, such as files names and lengths. There can be several files in one U disk, each file is composed by a set of data and identified by its name.

The U disk (including FLASH, USB FLASH and mobile disk) has become a very common storage device and its file system is usually FAT file system. It is mainly composed of the following kinds of data: MBR main boot record, DPT disk partition table, DBR Dos boot record, FAT file allocation table, BD directory and DATA region.

4. SYSTEM DESIGN

The embedded USB host system applied in the vehicle's black box transfers data through CAN bus, and stores data from local collection to the U disk in the form of file so as to realize data storage. Thus data collected can be analyzed by connecting U disk to any computer. The total design diagram is shown in Fig.2:

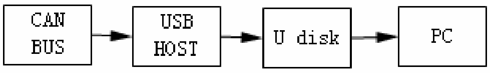


Fig.2. Total design diagram of the embedded USB host system.

4.1 Hardware Design

The embedded USB host system takes ATMEL MCU AT89C55 as the core controller, and the host interface chip adopts the USB host/slave interface chip CH375 produced by Nanjing Qinheng Company.

CH375 works under the host mode in the entire system, and links with MCU system bus in bus mode. MCU works with CH375 by integrated control of RD, WR, chip selection CS and address bus A8. The IRQ by INT pin output is effective in low level and this pin can connect to interrupt input pin INT0 of MCU. MCU uses interrupt mode to get interrupt request. Because MCU need larger data buffer, which is at least 600 bytes while MCU reads/writes to U disk, the system expands 32K SRAM62256 as data buffer, considering the influence of file system application and data transfers. Since 62256 is 32K of capacity (15 address buses), lower 8 bits of address is provided by flip-latch 74ALS573, and higher 7 bits is provided by P2.0-P2.6 of MCU, while P2.5-P2.7 is used as 3 address bus for selecting external chip. The system principle frame diagram is shown in Fig.3. CH375's TXD pin connects to ground directly or through by 1kΩ R so that CH375 works in parallel interface mode. The USB chip interface design diagram is shown in Fig.4.

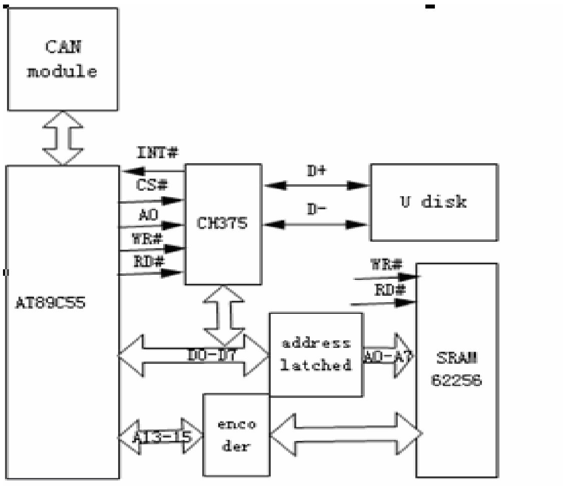


Fig.3. system principle frame diagram

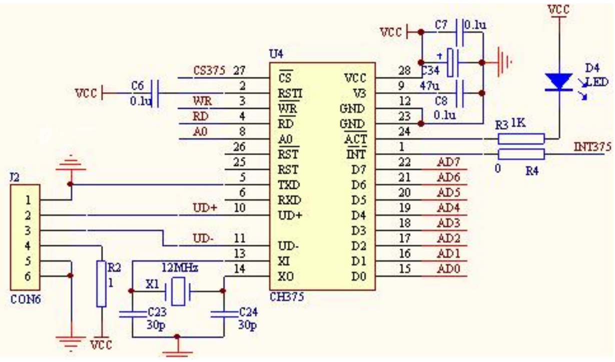


Fig.4. USB chip interface design diagram

4.2 Software Design

Since CH375 includes special communication protocol of firmware to process mass storage device, the embedded system MCU can take U disk as moving mass storage, and reading/writing data only needs several instructions but not knowing USB communication protocol in details. CH375 is specially used to deal with USB communication. While detecting the state change on USB bus or after finishing command, CH375 informs MCU to process in interrupted mode. MCU must be accord with CH375 sequence request while sending correlative commands.

Here is an example for explaining the operating principle, that is, storing the outside data collected by monolithic in the U disk: after receiving a series of communication order with outside controller, local monolithic will initialize and set the work style of CH375. Word sending mode and fan area sending mode are both available, fan area sending mode is adopted here. Then

CH375 can automatically detect the linking condition of U disk to USB port and inform the MCU in interrupt condition. After connecting with U disk,

MCU will inquire whether the state of U disk satisfy the memory request, such as checking the capacities, the mode of sub area or the form of document system etc. (CH375 supports the FAT document system only). MCU will transfer data after finishing all above operations and store all data received in 62256, then set up a new file in U disk, at last write all data in this file from 62256. The flow chart is shown in the Fig.5.

In the software design, we use MCU programming tool Keil C to write computer program. For simplifying programming, CH375 provides packed library function CH375HF6.LIB including some macro definition so as to greatly accelerate software programming. The following are some examples for the critical operational functions and programming function:
Setting CH375 command port parameters:
CH375_CMD_COMMAND_ADDR;
Setting CH375 data port parameters:
CH375_CMD_DAT_ADDR;
Setting CH375 initialize function:
CH375LibInit ();
Deciding U DISK connecting variable:
CH375DiskStatus ();
Justifying CH375 disk information:
CH375DiskQuery ();
Writing bytes to U disk:
CH375FileWrite ();
Data saved in U DISK is already FAT system file; which makes no need to any hardware device when connecting it to computer. At the same time, programming procedure is very convenient and files can be read directly.

5. CONCLUSIONS

This article has introduced the CH375 chip, and explored the design plan of the embedded USB host controller applied in the vehicle's black box, which is quite worthy for references while realizing USB host controller in the other embedded microprocessor systems. The result of experiment has indicated that this system can store the data in U disk accurately with high speed. It has proved the correctness and practicality of this system design plan. Further more, U disk is used as moving storage device in this system, which has general availability and cannot be confined to the unitary host controller. In addition, it is a cheap and convenient intermediary for collection and storage of data, which solve the problem of bringing heavier equipment on the spot for carrying out renewal and collecting in the past. With the fact that a great quantity of USB equipment popularizes, USB device communication has already become the inevitable trend, many interface chips such as CH375 will certainly have the vast application prospect. Moreover, the exploitation of interface chip with more powerful function is also urgent affairs.

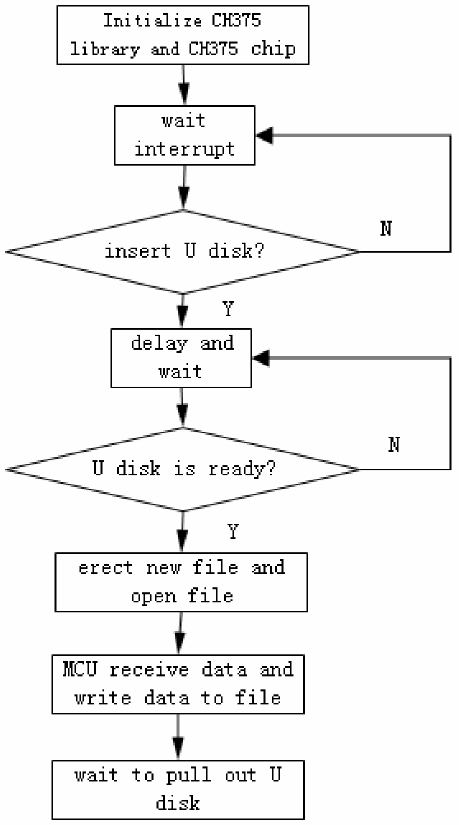

Initialize CH375 library and CH375 chip
wait interrupt
insert U disk?
N
Y
delay and wait
U disk is ready?
N
Y
erect new file and open file
MCU receive data and write data to file
wait to pull out U disk

Fig.5. Software design flow chart

REFERENCES

[1] Xiao Juxiong, Weng Tiecheng, Song Zhongqing
[2] *USB Technology and Application Design* [M]. Tsinghua University Press, 2003.
[3] Nanjing Qing Heng Co.Ltd.CH375 Interface Chip Data handbook (second edtion)[R]. 2004. 12.
[4] Nanjing Qing Heng Co.Ltd. *Instruction of USB Module Design Based on CH375* [Z].2004.
[5] Ma Wei. *USB system principle and Host/Slave Design.* Beijing Aerospace University Press.2004.
[6] Shao Beibei, Ma Wei. Application Area on "Embedded USB Host" Development of Expanding Moving Data Storage And Exchange.[J]. *Electrics Today*.2003.

Hardware/Software Co-design Methodology of SOPC Based FPGAs*

Wei Tang [1], Baojian Ge [2]
[1]School of Mathematics & Computer Science, Jianghan University, Wuhan, 430056, P.R.China
[2]School of Computer Science, Wuhan University of Science & Technology, Wuhan, 430081, P.R.China
Email: tomvea03@yahoo.com.cn

ABSTRACT

Co-design method of software and hardware becomes one of the most important methods for the embedded system development. This method not only overcome the limitation of the tradition method but also put up many advantage in the develop risk, periodicity, and stabilization. In this field, SOPC (system on programmed chip) become as a representative method. This paper discussed the model building, designing, simulation and integration of the SOPC system based on the development kit of Altera Company, and gives an example on this kit.

Keywords: FPGA, Co-design of SW/HW, SOPC, IP CORE.

1. INTRODUCTION

Currently, the trends of the embedded system design presents like as: the growth is being continued in integration of the system design; it is become realizable in integrating the CPU, memorizer and I/O devices into one chip. For the embedded system, SOC (system on chip) is the basic object and the application field is being extended. Different application has different requirement in the system function, consumption, real-time and size of the embedded system. Multiple aims design is used in-stead of single aim design. The reuse method which can help to assure the quality of the product, advance the efficient, short the development cycle is emphasized in the system design and the skill includes design reuse, hardware IP core reuse and software components reuse. With the requirement of the system complexity and size is continued enhanced, the objects of embedded system design tend from singled system towards distributed system**.**

2. CO-DESIGN METHOD OF THE EMBEDDED SYSTEM DEVELOPMENT

Usually, the traditional method for the application system development is first hardware system constructing then software modules programming. The flowchart of the traditional method is shown as Fig.1. In the process of embedded system design, it is to be modified the system construction and the codes of module more and more, until realize the anticipated purpose at last. If using the traditional method to design the system, it is maybe need more time, more cost, and more experiences of the designer. In the repeated modification process, one or more function or capability of the system may be deviated from the original purpose. These limitations presents more and more distinct with the continued growth of system’s size and complexity**.**

The embedded system is a system consisted of hardware and software and its development is different with soft-ware development in the all-purpose computer. Both the software and the hardware should be taken into account carry out the design process. Because they contact with each other so closely, making the Co-design become more and more necessary. HW /SW Co-Design method is a method which aim at overcome the above problems and challenges in the embedded system development. The main issue of Co-Design method is consisted the hard-ware subsystems with software subsystem making them work better.

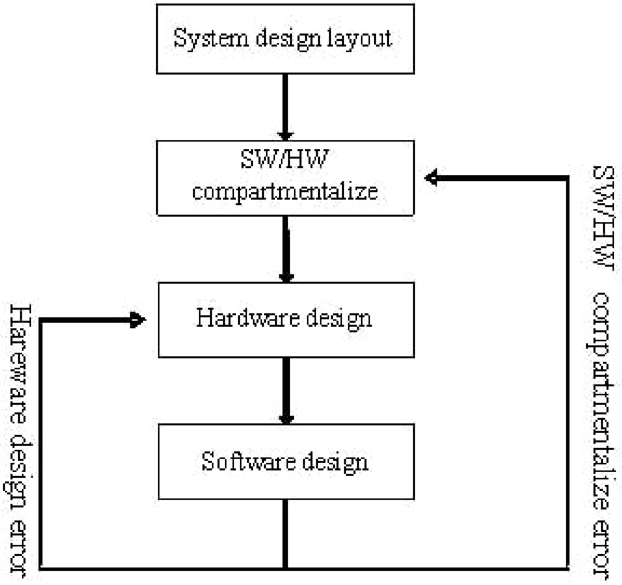


Fig.1. Traditional design flow

The dissimilarity of Co-Design method and tradition method is that it emphasize on parallel and mutually feedback, this character make it easier to overcome the limitations come from the traditional method which brings many abuses we cannot anticipated because of the design of the software and hardware separately. We con-firm that the relationship between the software and hard-ware via the Co-design method, make the system synthesis correctly and ensure the efficiency of the design. This method heightens the system design abstraction levels, and expands the scope of the design overlaid. At the same time, this method can realize the reuse of resource efficiently; especially software components reuse and IP core reuse. SW/HW Co-design method presents many advancements in the fields like shorten the develop cycles, reduce the system costs, enhance the system capability, and ensure the system quality.

3. PROCESS OF CO-DESIGN METHOD

The embedded system is consisted of appropriative hardware and component software alternant with the hardware. Usually, it includes three parts: software component, hardware unit and appropriative communication channel. The main challenges that the embedded system development face on include:

* Science and technology development project of Wuhan City government[20043007077-19].

systems modeling, describing the main function of the system; compartmentalizing the system felicitously into hardware and software, balancing on the satisfying requirements with reducing the cost of the system resources; checking the system's realization and criterion and so on.

The goal of Co-design method is making every part of the system can concerted with each other to accomplish the function, making the design of the both parts evolvement and integration synchronized. The design flow is shown in Fig.2.

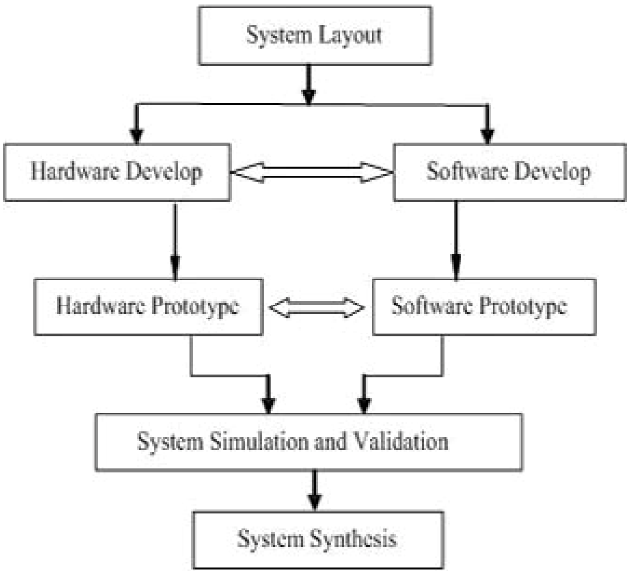


Fig.2. Co-design method flow

Hereon, we divide the process of Co-design method into four phases:

1) Layout of the system:

Descript the function model in system level according to the anticipated requirements. It means that one or more system levels description languages should be used to define the system function. After that, the map of the function and the capability will be gotten and a holistic system model will be formed. For the idiographic design, there are kinds of description language such as natural language, HDL (hardware description language) or System C etc. As another choice, UML2.0 can be used for hardware modeling.

2) System design:

After the first step, we make use of the model just built to compartmentalize the system into subsystems of software and hardware and mapped the function requirement to system design.

Compartmentalizing the system into subsystems of hardware and software is one of the pivotal steps of the design process. The characters of hardware subsystem determine the capability of the system and the cost. The characters of software subsystem make the function of the system come true and determine the agility of the system. A proper compartmentalization of the system is base on the goal we originally wish, can maximum reduce the cost and increase the agility of the system, so it is one of the most important steps in embedded system design. The mapping of system from function to design is based on the proper compartmentalization of the system. It means we should properly define the interface between the hardware and software subsystem, educe the best method of the system design. After that, we can determine the architecture and structure of the system and integrate these modules at last.

3) Simulation and validation:

Simulation is a course which evaluates the validity of the system function. According to the phases of the design process, we divide the simulation into system levels simulation, behavioral levels simulation, RTL levels simulation and gate levels simulation. We can use one or more of them to accomplish the simulation of the system base the different requirement. System levels simulation is used to check the correction of the algorithm and evaluate the entire function. The typical SW/HW Co-simulation is run the software on simulative hardware. If we have validated the function of the IP Core or other sources independently, we can skip over the gate levels and behavior levels simulation.

4) System synthesis:

Now we can begin the feasibility design of the system based on the above works. The feasibility design of the system includes hardware design and software design, and the interface between them. For the hardware subsystem design, we need do something like the choice of chips, interface define, and hardware program design corresponding to different chips. In this part, we should pay attention to the function optimize of the every modules. For the software subsystem design, we need do things like the definition of the interface of hardware and software, choice of programming languages, and the compiler tools. Here, we should pay attention to the necessary code optimization. Design the modules of the system parallelized, and intercommunicate the information of software and hardware. Validate the function of every module, integrate the modules together and synthesis the entire system, and then validate the function of the system. If reusable components are used in the system, we should check the validity of them.

4. DEVELOP KIT OF CO-DESIGN BASED ON FPGAs

The development of FPGAs satisfied the requirement of hardware capability. The goal of SOPC is put an integrated digital electronic system into one chip. The system includes processor, coprocessor, interface, DSP unit, memory and peripheral and so on. FPGAs provide a good foundation for SOPC, provide a hardware and software synthesize solution based on the system logic design. The solution can optimize the system in scale, reliability, vol., cost, function, capability, time and hardware update etc.

The FPGA develop tools Quartus II provided by Altera company is one of the most popular develop kit support SOPC Co-design method. In the environment of Quartus II, we can use kind of input modes to design modules of the hardware system.

Micro controller IP core can be used as a good choice for SOPC Co-design, such as 8051 IP core, 8086 IP core. Here we give the design flow of 8051 IP core as an example. We first custom the components of the IP core according to the function we want to realize, then synthesize the IP core in QuartusII environment, simulate the function of it, and download the configuration file into FPGA, the hardware of the system is finished. We can develop the software program in the Keil C51 environment use C or assembly language to realize the function of the system, compile the software program and create the

HEX files, use the HEX files as the configuration files for the ROM program data of the IP core. It is a good solution for SOPC SW/HW Co-design. Fig.3 shows the develop flow of the IP core.

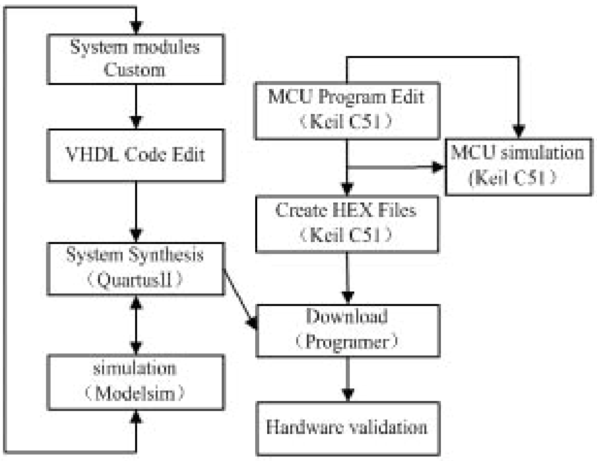

System modules Custom
VHDL Code Edit
System Synthesis (QuartusII)
simulation (Modelsim)
MCU Program Edit (Keil C51)
MCU simulation (Keil C51)
Create HEX Files (Keil C51)
Download (Progrmer)
Hardware validation

Fig.3. MCU IP Core Co-design flow

SOPC Builder, as one of the components of Quartus II, provides a powerful platform for creating memory mapped systems based on processors, peripherals, and memories that are internal or external to the FPGA. SOPC Builder generates the Avalon switch fabric that contains logic to manage the connectivity of all modules in the system. The NIOS II processor is a RISC processor adopts pipeline and Harvard architecture. We can integrate the modules such as CPU, memories internal or external, peripherals and reusable IP CORE into a system module via it. We will get a whole function system on one chip by synthesize these modules at last.

NIOS II processor provides 32-bit instructions, interface of system and peripherals. The NIOS II integrated development environment (IDE) is the software development graphical user interface (GUI) for the NIOS II processor. It is based on the popular Eclipse IDE framework and the Eclipse C development tool kit (CDT) plug-ins. The NIOS II compiler tool chain is based on the standard GNU GCC compiler, assembler, linker, and make-file facilities.

By the support of Quartus II, We can add custom components into system, design the modules by HDL, make communication with CPU via Avalon bus. If DSP Builder is installed, it can be a toolbox of Matlab, we can design some hardware modules on algorithm levels, create the custom components, integrate the custom components into the system.

Ever since the system was created, we should synthesize and simulate it, create the files which can configure the FPGA chip. We put up the software program in the integrated development environment NIOS EDS, it support C/C++ program language. The software program is based on the hardware system we created, and we can modify the hardware system if the software design needed. So, the convenience of this method is evidently.

Here we give a design flow of Quartus II developmentenvironment based on the system modeling by Matlab,the chief steps of the flow is shown in Fig.4.

In this development environment, we whish control the stepping motor by NIOS II processor. First, we design the component of stepping motor for the NIOS II processor which is described by VHDL. We define the bus interface of NIOS II processor and the stepping motor component. The component description include the input signals like clock, read, write and output signal like motoout[3..0] which control the motor etc. Second, we define the HAL by C language head file; it is the foundation of the communication between software and hardware. Third, we develop the software by the NIOS II IDE development environment. We choose C as the programming language,and descript the function accomplishing the micro-step driving of the stepping motor. We added the display module into the system, and the precision of the micro-step displayed on the numeral tubes.

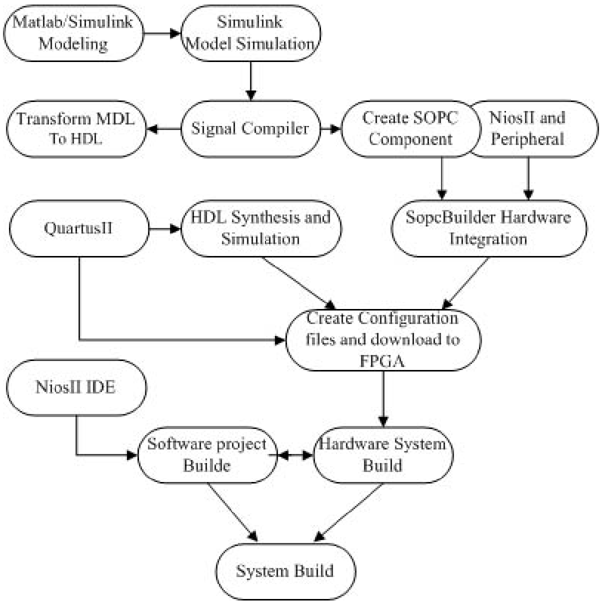


Fig.4. Representative FPGA Co-design develop environment design flow

Compared with the traditional design method, this method has some advantages like system customize agilely and conveniently, software programming simple. This experiment has validated the agility of the Co-design method.

5. CONCLUSIONS

SW/HW Co-design method has become indispensable trend in the embedded system design. This method predigest the complexity of system design, shorten the develop periodicity; ensure the system quality. With the continued development of digital logic technology, SW/HW Co-design method for the embedded system design will be developed. It will become more and more important in the embedded system design.

REFERENCES

[1] Jingzhao Ou and Viktor K. Prasanna, “MATLAB/Simulink Based Hardware/Software Co-Simulation for Designing Using FPGA Configured Soft Processors,” *IPDPS*’05. 2-3

[2] WANG Shaoping, WANG Jingqian, QIAN Wei, “Hardware/Software Collaborative Design of Embedded System,” *Modern Electronics Technique*.2005.2

[3] Pan Song, Huang Jiye, Zeng yu, *SOPC Technique Practically Tutorial*, Tsing Hua Universtity Press. 2005.3

[4] M.Haldar, A.Nayak, A.Choudhary, and P.Banerjee, "Automated Synthesis of Pipelined Designs on FPGAs for Signal and Image Processing Applications Described in MATLAB," in *Proceedings of the 2001 Asia South Pacific Design Automation Conference*, vol. 1, pp. 645-648, 2001.

[5] Yan Yingjian, Liu Mingye, "A Hardware-Software Co-Verification Method for SOC Design," *Journal of Electronic & Information Technology*, 2005.2.

Wei Tang is an Associate Professor, in Math and Computer Department, Jianghan University. His research interests include information security and real-time embedded systems. He has published two textbooks, more than ten Journal papers.

Baojian Ge is a master graduate in School of Computer Science, Wuhan university of Science and Technology. He is interested in embedded system development.

Research and Design of Embedded GUI System Based on Linux

Tianhuang Chen, Yanli Zhang
College of Computer Science & Technology,.Wuhan University of Technology, Wuhan
Email: thchen@whut.edu.cn

ABSTRACT

This paper analyses the realization methods of some known embedded GUI, as well as their advantage and disadvantage. Considering the characteristic of linux embedded system, the author proposes some betterment measure aiming at Client/Server mode of linux embedded GUI system in order to improve the stability of system.

Keywords: Embedded System, Graphical User Interface(GUI), Client/Server Mode

1. INTRODUCTION

Graphical user interface(GUI) is system-level bottom software. GUI together with file system and operation system kernel, can structure a integrated operation system[1]. A excellent operation system should provide good GUI, or else it will bring some trouble for user's operation and make program developer difficult to design out good application program with friendly interface[2-3]. So, GUI is very important to an embedded operation system. Considering the sensitivity of consuming production to cost, Windows CE, QT Embedded and so on embedded GUI system is more strictly with resource. Embedded system suffers the restricting of various resources, so it needs a light-level GUI support with efficient and friendly.

2. FAMILIIAR GUI PRODUCTIONS AND THEIR CHARACTERISTICS

With the development of embedded system, there come forth numbers of embedded GUI system. At present, embedded GUI system have mostly WinCE, Qt/Embedded and MiniGUI.

2.1 WinCE

WinCE is based on Windows , and provide a stable real time operation system to intelligence mobile system and small memory device. WinCE is a tidy Windows95, but at the part of technology, it is not a excellent embedded operation system. Embedded operation systems demand efficiency and saving resources, but WinCE cannot satisfy this request. It engrosses overmuch memory, and its application program is too large.

2.2 QT/Embedded

QT/Embedded is produced by the known QT library develop businessmen, and is the GUI system of QT version oriented to embedded operation system. Its main characteristic is good portability. Many X Window programs based on QT can be ported very expediently into embedded devices. But, it has many disadvantage. Firstly, QT is a C++ library, and must maintains a set of language-level data type and data structure for more high portability and platform independence. So, this makes the whole system very large. Secondly, the controls style follows the PC style, and don't adapt to operation request of embedded devices. Finally, the structures of QT/Embedded is too complex, especially the MOC file realization signal/slot mechanism, so it are difficult to expand, tailor or port.

2.3 MiniGUI

MiniGUI is system GUI system which is oriented embedded system and real time. It run mainly at the Linux console, also can run at POXIS compatible system with a POXIS thread support. MiniGUI have many advantages. Firstly, the storage space occupied by MiniGUI is very small. The whole space occupied by system should be between 2MB and about 4MB. For some system, the space occupied by MiniGUI with self-contained function can be reduced to within 1MB. Secondly, for meeting the various need to embedded systems, MiniGUI must be configurable. By the Automake and Autoconf interface in Linux, it can achieve large numbers of compiling configure options. These options can specify what functions are included in MiniGUI library. Finally, The abundant function and configurable trait of MiniGUI make it can run not only low-end products but also high-end products based on ARM9. Developer can use the advanced control style and skin interface technology and so on to create gorgeous user interface.

3. THE SYSTEM STRUCTURE DESIGN OF EMBEDDED GUI SYSTEM

To realize the portability and configurable trait of the system, it adopts the design idea of layer. As show at Fig.1. the system module can be divided to the device driver module, window management module, message management module and API application interface module. The message management module and window management module are the core module. These modules cooperate to realize the base function of GUI.

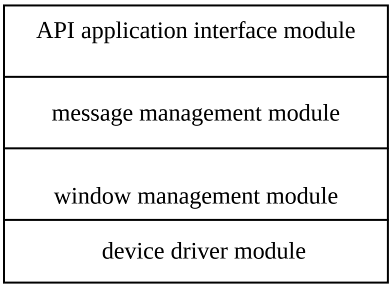


Fig.1. System structure of GUI

3.1 Message Management Module

This module manages all message of GUI system. The program with message driver will still be at loop state. Different with traditional program, the program with message driver still be at loop state after startup. The program in loop state captures some events, such as user's key-down or the movement of mouse, then makes some response to these events and achieves relevant functions. The loop will be over until the program capture certain message. So, the message transfer and processing is the core of the GUI system with message driver. The system must ensure that the messages do not lose and have the correct executing order after coming into the core module. These will be managed and controlled by message management module.

3.2 Window Management Module

Most embedded system request the multilevel windows. So, this GUI system design the three-level windows: main window, child window, control. Main window could contain certain child windows and controls, and each child windows also could contain certain controls. This three-level windows form a window tree relation, as shown at Fig.2. The child windows and controls belonging to same parent window connect together to chains. Except main window, all child windows and controls have a pointer to point at the parent window. This can provides convenience for each window finding each other. The data structure of window provides the mutual operation and attribute of all window unit. The main attributes have the position, size, style, border of window unit, and the pointer to the window member of the chain.

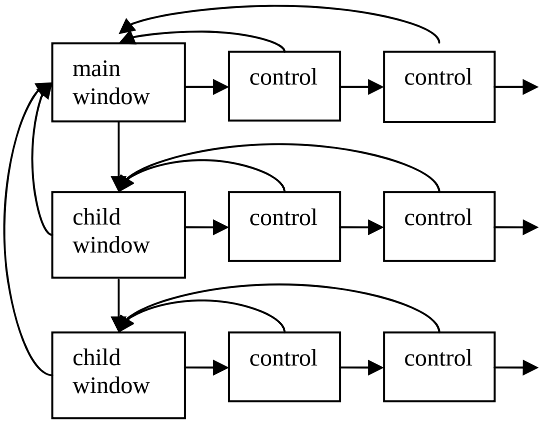


Fig.2. The structure of window tree graphic

3.3 Device driver module

The devices supported by GUI system commonly are input or output devices. The input devices are the main operator of GUI system, and GUI system will send in the form of message the event of input devices to the core module of this system, and be managed by message management. GUI system provides mainly graphic windows, control and so on to user, and they are all displayed on the output devices.

This GUI system uses Frame buffer based on Linux as the display driver. This driver program interface abstracts the display devices into frame buffer, and it provides the abstract interface of graphic hardware based on frame buffer. Frame buffer are seen as the mapping of display memory, and user can map frame buffer into user process space to make directly read and write operation to achieve the function of controlling graphic display. This method shields the detail of bottom graphic hardware, and provides compatible interface of accessing graphic hardware in different platform. It can implements some basic graphic operation, such as read or write pixel, paint horizontal and vertical line, rectangle and block copying and so on, and support upper advanced graphic operation, such as font model, color model, and paint image and so on. This reduce redundant code , and predigest the realization of display driver, and ensure the good portability.

4. IMPROVEMENT TO TRADITIONAL CLIENT/SERVER MODULE

MiniGUI supports Client/Server module and multi-thread, and it supports multi-process after Lite edition. But, MiniGUI only displays a process window at a time. This is not true multi-process, so difficult satisfy the request of user for high quality GUI[4]. Using process as unit realize the multi-task Linux. Linux system provides safe running environment for process to ensure the independency of each process, in order to the crash of a process will not effect other process. So, realization the process-level GUI system will provides the stability of system, at the same time, satisfy the request of user.

4.1 Micro-Client/Server Module Description

At this system environment, desktop process and application process form the micro Client/Server relationship. Any application will be started up by GUI desktop process. After application starting up, it will communion firstly with desktop process, and send itself information to desktop process. So, at any time, the desktop process knows how many processes at present are running, and holds some description information of running process. And, application processes send some request in order to get current information of system, or make desktop process achieve some else computing task, such as current full clipping regions. As for the client process, not all output request need be sent to server and be outputted by server. Client process only need to send necessary information to server, and else output and clipping regions are maintained by client itself. This improve greatly the performance of system, and reduce the process communication and the operation of complex windows. So, the design of Micro-Client/Server reduces the cost of system resources.

In the system, the communication between client and server are realized by Domain Socket[5]. As shown at Fig.3, as for the connection of each client, server will create a single channel to send data, in order to make any one client can not disturb the communication between server and other client, and insure better stability of system. Firstly, server and client run in different process space. Because the processes have independent address space, the problem appearing at client process do not lead to the crash of server process. In addition, any bad process do not affect else process. Finally, regardless of process whether close normally, server process will always receive a closing socket message, then release the resource held by client process in the course of processing messages. The server process has a description list to store the information of client processes. The each node in the list contains the name and process ID of application process, as well as the rectangle borderline of window.

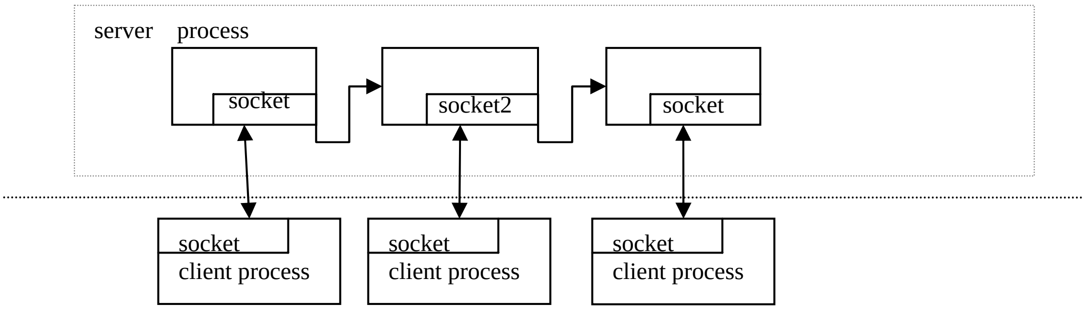


Fig.3. The sketch map of Client/Server Domain Socket.

4.2 Management Of Server And Client Process

Server process uses a list to express the information of current connecting client process. In the list, a node is as a client process of connecting to current server process. Its structure is shown at Fig.4.

```
    char          pAppName[256];     //Application
name
    RECT       rc;          //main window position
    PClipRegion   pClipRgn; //main win cut field
    BOOL           bVisible;       //visible
    int        fdSocket;       //Socket description
    pthread_t    tdSocket;      //socket thread ID
```

Fig.4. The structure graphic of node

The messages transferred between server and client contain some parts as follow.

1) When client create main window, send message to server: LMSG_IPC_CREATEAPP. Client create Domain Socket to connect with server, then send creating message to server. The append data of message is the description of client process.
2) When client show main window, send message to server: LMSG_IPC_SHOWMAINWIN. Client send message to notify server to do some computing work, but output work still be achieved by client itself.

5. CONCLUSIONS

This paper designs a embedded GUI system antetype based on Linux. After analyzing the Client/Server module of MiniGUI, the author proposes some advanced measure and designing Micro-Client/Server module. The measure can enhance the stability of system and reduces the cost of system memory.

REFERENCES

[1] WANG Bo-yun，LI Sheng-yang，BAI Lin，LUO Yup-ing. “Light-weight Embedded GUI System and Its implementation”. *Computer Application* 2006, 26(9): 2244-2247.

[2] Bejing Feymman Software Technology Co. MiniGUI Technology White Paper [EB/OL]. http://www.minigui.com/whitepaper/MiniGUIThchWhitePaper-2.0E.pdf ,2005.9

[3] Richard N T. “A component and message based architectural style for GUI software”.*IEEE Trans on Software Engineering*.1996, 22(8):39-406.

[4] Alesssandro Rubini. *Jonathan Corbet Linux device drivers*.New York: OReilly, 2001.

[5] Gray,J.S. *Interprocess Communication in UNIX , Second editor BeiJing*: Publishing House of Electronics Industry，2001.3

Research on Dual-interface SIM and Unauthorized Using Protection for Mobile Terminal

Meihong Li [1], Qishan Zhang [2]
[1] School of computer and information Technology, Beijing Jiaotong University
[2] School of electronic information and engineering, Beihang University
Beijing, 100044, China
Email: mhli1@bjtu.edu.cn

ABSTRACT

A novel dual-interface SIM card is designed to take the place of the original one with contact mode in a mobile terminal. Here we focus on designing an antenna for SIM. In addition, a scheme on using a mobile terminal as an electronic purse is presented. To protect unauthorized using, an independent key card is proposed by integrating a RF module in the mobile terminal.

1. INTRODUCTION

The development of mobile terminal is so fantastic in the world, especially in China. It is reported by China government that the number of users using mobile phone is more than 300 millions. Similarly, smart card industry that has a close relationship with mobile terminal also grows rapidly. It is well known that each mobile terminal has a piece f SIM card that conforms to the standard of smart card named ISO/IEC 7816. The most successful application field of smart card is the one of public transportation. Therefore, a new problem is to be faced. A user should hold a mobile terminal and a few pieces of cards at the same time. In this paper, a solution is presented that only one mobile terminal can provide the functions of phone and smart cards, in which a new dual-interface SIM and a RF module are to be designed. Also, an independent key card is adopted for unauthorized using protection.

In section 2 of this paper, a schematic structure of mobile terminal is presented. In section 3, a total solution on dual-interface SIM is designed and implemented. An independent key card is introduced in section 4. Finally, in section 5 some conclusions are drawn. Also, the current issues are to be presented.

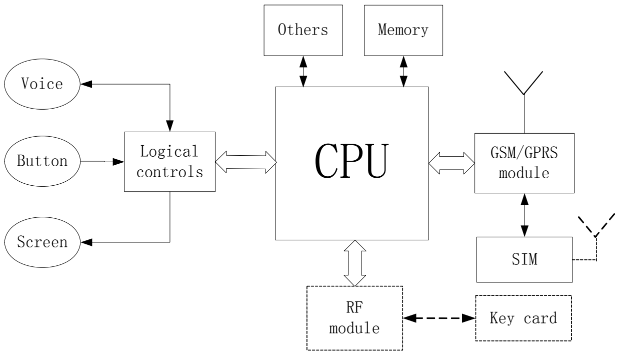


Fig.1 schematic structure of system

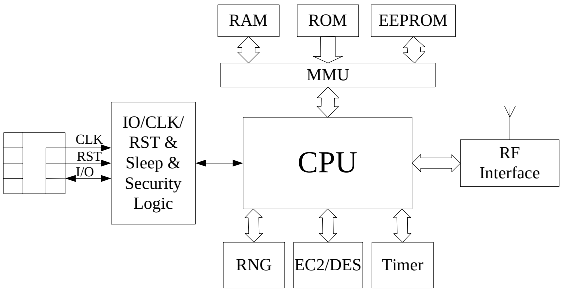


Fig.2 structure of a dual-interface SIM

2. SCHEMATIC STRUCTION OF SYSTEM

Our scheme is based on an existed platform of mobile terminal. A detail solution of a mobile terminal is not covered in this paper. The schematic structure of the system is shown in the Fig. 1[1]. The following mainly refers to the parts of a dual-interface SIM, an antenna of SIM, RF module and an independent key card.

As is shown in the fig. 1, a dual-interface SIM with its antenna is designed to replace the original one for the application of an e-purse. An independent key card is the contactless one for unauthorized using protection. A RF module is adopted to identify the personal key card of the cardholders. A piece of key card is a unique one for a mobile phone. The relationship resembles the one between a key and a lock.

3. SOLUTION ON DUAL-INTERFACE SIM

In this section, firstly a prototype of a dual-interface smart card is given. Secondly, the technique of antenna design based on the original SIM card is presented. Lastly, a method on using mobile terminal to perform transaction is proposed.

3.1 Implementation of A Dual-interface SIM

A kind of typical structure of a dual-interface chip is shown in Fig 2, in which the dual interfaces conform to ISO/IEC 7816 and ISO/IEC 14443A&B respectively[2].

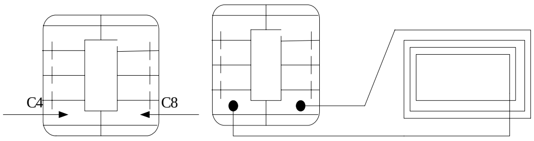


Fig.3. Connection of SIM with external antenna

As for the original SIM with contact mode, we would design an antenna for the new dual-interface SIM. But it would be better to be compatible with the original system as possible. The detail design is the following.

An original SIM card has two unused reserved points named

C4 and C8. Here we could use the two points to connect an external antenna of SIM. The antenna of SIM is designed to locate in the back or front of a mobile terminal. The connection of antenna for SIM is shown in Fig 3. This kind of antenna is different from the one of mobile terminal. Its size is very small, which has little interference to the structure of a mobile phone. In addition, a power consumption of a contactless application is obtained from terminal device, which doesn't consume the power of the battery. At last, the problem of EMC should be considered. As the work frequency of SIM equals to 13.56MHz, and that of mobile phone equals to 900/1800/1900 MHz, a mutual EMC interference in theory doesn't exist. But the shield design should be considered.

3.2 Application for a Dual-interface SIM

A dual-interface SIM is mainly designed for an application of e-purse. The application for the bus transportation is taken an example. To comply with an original system, a dual-interface SIM card should adopt compatible software as before. Thus the original transaction terminal should not be changed. But several procedures as below need to be updated in dual-interface SIM card.

At first, we should create an environment of application according to the requirement of a bus transportation, which is called card issuing. Then a user would charge for his e-purse. He could perform the operation not only by transferring account from the balance of mobile terminal, but also from the appointed charging points. Finally a user could perform a transaction in a bus terminal, even when his phone is calling. However, a user should remember that a valid distance between his phone and a terminal is kept for less than 10 centimeters.

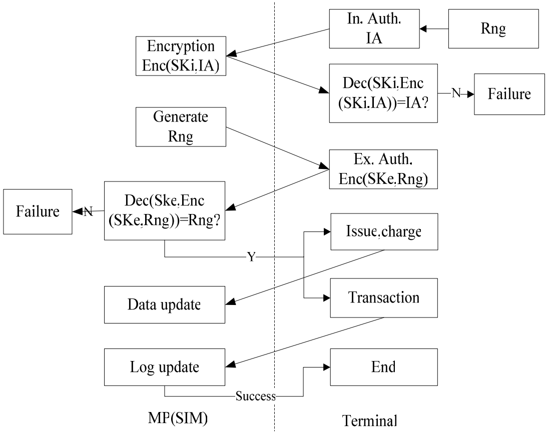


Fig.4. Flowchart for authentication

A general authentication should be performed between the both. A typical scheme for mutual authentication is shown in the Fig.4[3][4] . The detail introduction is the following.

The first procedure is internal authentication. The authentication object is the SIM card. Firstly, a terminal transfers authentication data IA to SIM card. Secondly, SIM card encrypts IA using DES and transmits the cipher data to the terminal. Finally the terminal would decrypt the cipher data and compare with the original one.

The second procedure is external authentication. The object is the bus terminal. Firstly, a terminal issues Get challenge command to SIM card for random number. Secondly, it encrypts the random number from SIM card using DES and transmits the cipher data to SIM card. Finally the SIM card would decrypt the cipher data and compare with the original random number.

The last one is the transaction procedure. The two parties should backup the log record of transaction every time. As for limited memory, a SIM only stores the records of recent transactions. The terminal should be equipped with a larger memory, or transmit the recent transaction records to the supervising center by network periodically. In addition, the two parties should support data restoring mechanism to prevent important data from losing while abnormal power-down. The transaction time may be different while different applications. But it should be less than 200 milliseconds for the bus transportation.

4. UNAUTHORIZED USING PROTECTION

At present, there are three kinds of method as below in access control field. The first is ID and password, the second is smart card technology, and the last is biometric recognition technology. Here a dual-interface SIM is a kind of smart card in itself. Therefore, we present a method adopting an independent key card to protect unauthorized using. The detail introduction is the following.

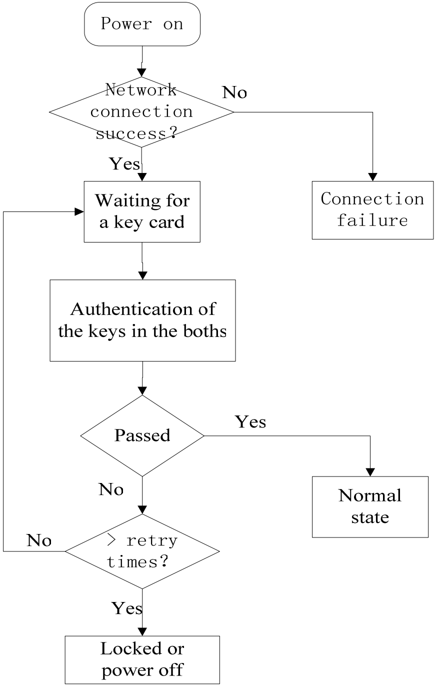


Fig.5. Workflow of a key card

As is shown in Fig 1, a RF module embedded in mobile terminal is designed to identify the unique key card. The workflow of a key card is illustrated in Fig.5[5]. In which, the key card is integrated with the contactless chip which is compatible with ISO/IEC 14443. It could be designed to several kinds of handy figures such as general IC card, watch, and button etc. If the independent key card is lost or corrupt, the user should report to the authorized points for maintenance.

5. CONCLUSIONS

Based on the above solution, we have simulated the procedure of authentication and transaction between a dual-interface SIM and a terminal using a dual-interface chip of SLE66CLX320P in Infineon Corp. Firstly, we developed a dual-interface operating system in the chip. Secondly, an application of e-purse with personal information of a cardholder is created. Finally, the mean time of one-time transaction (debit operation) and the one of verification time between a pair of keys in two reader devices are shown in table 1(Unit: milliseconds). These indicate that it is feasible to implement the dual-interface SIM and an independent key card in the mobile terminal.

The above solution would implement one-phone-multiple-functions for users. But, at present several issues still exist. Firstly, a dual-interface SIM card would bring more cost. Secondly, the number of manufacturers of dual-interface SIM card is very few. Lastly, because memory space of SIM card is limited, a new e-purse application would influence the original system.

Table 1. Time of transaction and verification

Items/Readers (milliseconds)	**Debit operation**	**Verification**
Infineon reader	174	30
Solomon reader	181	42

ACKNOWLEDGEMENTS

Thanks to smart card research group in Beihang University for supplying the hardware devices. The research is supported by: Science and Technology Fund in Beijing Jiaotong University (No. 2005RC048).

REFERENCES

[1] Zhao Changkui,*Application system of GSM digital mobile communication[M]*, Bejing: national defense industrial press, 2001.

[2] Wang Aiying. *Smart card technology [M*], Beijing: Tsinghua University press, Jul, 1996.

[3] Isobe, Y.; Seto, Y.; Kataoka, M.; “Development of personal authentication system using fingerprint with digital signature technologies[C],” *in Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pp3-6, 2001.

[4] Bing, H.; Zheng-Ding, Q.; Dong-Mei, S.; “Secure authentication system incorporating hand shapes verification and cryptography techniques[C]”, *TENCON '02*, Oct. 2002.

[5] Bing H., Zheng-Ding Q., Dong-Mei S., “Secure authentication system incorporating hand shapes verification and cryptography techniques”, *2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering,* Oct. 2002.

A Parallel Sorting Scheme of 50 Numbers and its Hardware Implementation on FPGA*

Bing Zhang, Jinguo Shi, Mingcheng Zhu
Faculty of Information Engineering, Shenzhen University
Shenzhen, Guangdong 518060, China
Email: zhangb@szu.edu.cn

ABSTRACT

This paper presents a parallel sorting scheme on *50* numbers which is an extension of our previous work of sorting *10* numbers based on harmony theory neural network. The network is derived from the densest graph with degree *7* and diameter *2* which can accommodate *50* nodes. The nodes of the network are connected in an incomplete way, and the connections between nodes are greatly reduced. The proposed sorting algorithm has the characteristics of simple in function and easy for hardware implementation, and the sorting operation can be terminated in a finite number of steps, with capability of finding the maximum or minimum number in a single step. A prototype of the proposed sorting algorithm of *50* numbers is implemented in FPGA (Field Programmable Gate Array), and test result shows that the basic sorter can sort *50* numbers in *50+1* clock cycles, with maximum or minimum sorted in the second clock cycle.

Keywords: Parallel Sorting Algorithm, Sorting Network, VHDL Simulation

1. INTRODUCTION

Sorting is one of the basic operations in the application of computers. It has a wide application in the area of VLSI design, digital signal processing, network communication and database management systems, in which large data optimization and processing algorithms are required. For example, sorting plays a crucial role in the selection and evaluation of chromosomes in genetic algorithm. It has been estimated that over *25%* of the total processing time in artificial evolution of optimal VLSI design using evolutionary algorithms like genetic algorithms involves sorting operation. The nature of evolutionary algorithms and time-consuming sorting operation makes the evolution process very slow, which is a big problem in the research and application of evolvable hardware. Especially, when the complete hardware evolution approach is adopted, the evolution and evaluation process has to be implemented in hardware, which requires the evolution algorithm easy and cost effective for hardware implementation. This paper presents a parallel sorting scheme of *50* numbers with the aim of easy for hardware implementation on FPGA as a sorting engine in research on complete hardware evolution. The parallel sorting algorithm presented in this paper is an extension of our previous algorithm of sorting *10* numbers which has been implemented in hardware, and it can find the maximum or minimum of 10^n numbers in *n* clock cycles[2]**.**

The paper is organized as follows: After introduction, the sorting network and the parallel sorting algorithm of *50* numbers is presented. Then, issues in hardware implementation of the algorithm is discussed with VHDL simulation result shown. Finally, conclusions are drawn.

2. THE PROPOSED SORTING SCHEME

2.1 The Architecture

The parallel sorting algorithm presented in this paper is based on a network architecture which is derived from harmony theory neural network which is briefly introduced in [1] and the most dense graph with degree 7 and diameter 2 which can accommodate 50 nodes[3]. A 3-layer neural network with each layer have 50 nodes is constructed and is shown in Fig. 1.

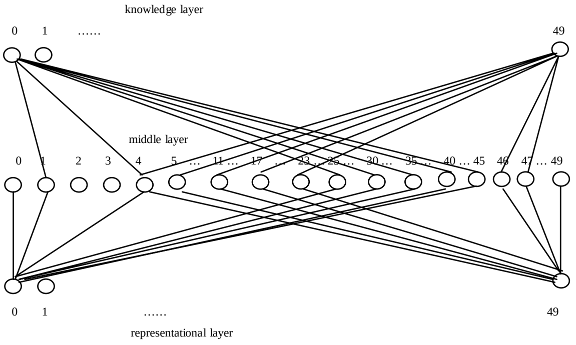


Fig.1. A neural network model to sort 50 numbers

A sequence of 50 numbers $\{n_i\}$(*i=1,2,…,50*) to be sorted are fed into the nodes of the first layer which is called the representational layer, the second layer is called the middle layer, and the third layer is knowledge layer. The *50* nodes in the representational layer represent the *50* numbers to be sorted. The activation of *i*th node of the knowledge layer suggests whether number n_i should be put into the sorted sequence.

Table 1 shows the detailed connections between representation layer nodes, middle layer nodes, and knowledge layer nodes.

Table 1. Detailed Connections

Representation layer node no.	Middle layer nodes	Middle layer node no.	Middle layer nodes
0	0,1,4,25,30,35,40,45	0	1,4,25,30,35,40,45
1	1,0,2,26,31,36,41,46	1	0,2,26,31,36,41,46
2	2,1,3,27,32,37,42,47	2	1,3,27,32,37,42,47
3	3,2,4,28,33,38,43,48	3	2,4,28,33,38,43,48
4	4,0,3,29,34,39,44,49	4	0,3,29,34,39,44,49
5	5,6,9,25,31,37,43,49	5	6,9,25,31,37,43,49
6	6,5,7,26,32,38,44,45	6	5,7,26,32,38,44,45
7	7,6,8,27,33,39,40,46	7	6,8,27,33,39,40,46
8	8,7,9,28,34,35,41,47	8	7,9,28,34,35,41,47
9	9,5,8,29,30,36,42,48	9	5,8,29,30,36,42,48
10	10,11,14,25,32,39,41,48	10	11,14,25,32,39,41,48
11	11,10,12,26,33,35,42,49	11	10,12,26,33,35,42,49
12	12,11,13,27,34,36,43,45	12	11,13,27,34,36,43,45
13	13,12,14,28,30,37,44,46	13	12,14,28,30,37,44,46
14	14,10,13,29,31,38,40,47	14	10,13,29,31,38,40,47
15	15,16,19,25,33,36,44,47	15	16,19,25,33,36,44,47
16	16,15,17,26,34,37,40,48	16	15,17,26,34,37,40,48

[*] This work is supported by Shenzhen University, project no:200516.

17	17,16,18,27,30,38,41,49	17	16,18,27,30,38,41,49
18	18,17,19,28,31,39,42,45	18	17,19,28,31,39,42,45
19	19,15,18,29,32,35,43,46	19	15,18,29,32,35,43,46
20	20,21,24,25,34,38,42,46	20	21,24,25,34,38,42,46
21	21,20,22,26,30,39,43,47	21	20,22,26,30,39,43,47
22	22,21,23,27,31,35,44,48	22	21,23,27,31,35,44,48
23	23,22,24,28,32,36,40,49	23	22,24,28,32,36,40,49
24	24,20,23,29,33,37,41,45	24	20,23,29,33,37,41,45
25	25,0,5,10,15,20,27,28	25	0,5,10,15,20,27,28
26	26,1,6,11,16,21,28,29	26	1,6,11,16,21,28,29
27	27, 2,7,12,17,22,25,29	27	2,7,12,17,22,25,29
28	28,3,8,13,18,23,25,26	28	3,8,13,18,23,25,26
29	29,4,9,14,19,24,26,27	29	4,9,14,19,24,26,27
30	30,0,9,13,17,21,32,33	30	0,9,13,17,21,32,33
31	31,1,5,14,18,22,33,34	31	1,5,14,18,22,33,34
32	32,2,6,10,19,23,30,34	32	2,6,10,19,23,30,34
33	33,3,7,11,15,24,30,31	33	3,7,11,15,24,30,31
34	34,4,8,12,16,20,31,32	34	4,8,12,16,20,31,32
35	35,0,8,11,19,22,37,38	35	0,8,11,19,22,37,38
36	36,1,9,12,15,23,38,39	36	1,9,12,15,23,38,39
37	37,2,5,13,16,24,35,39	37	2,5,13,16,24,35,39
38	38,3,6,14,17,20,35,36	38	3,6,14,17,20,35,36
39	39,4,7,10,18,21,36,37	39	4,7,10,18,21,36,37
40	40,0,7,14,16,23,42,43	40	0,7,14,16,23,42,43
41	41,1,8,10,17,24,43,44	41	1,8,10,17,24,43,44
42	42,2,9,11,18,20,40,44	42	2,9,11,18,20,40,44
43	43,3,5,12,19,21,40,41	43	3,5,12,19,21,40,41
44	44,4,6,13,15,22,41,42	44	4,6,13,15,22,41,42
45	45,0,6,12,18,24,47,48	45	0,6,12,18,24,47,48
46	46,1,7,13,19,20,48,49	46	1,7,13,19,20,48,49
47	47,2,8,14,15,21,45,49	47	2,8,14,15,21,45,49
48	48,3,9,10,16,22,45,46	48	3,9,10,16,22,45,46
49	49,4,5,11,17,23,46,47	49	4,5,11,17,23,46,47

As shown in Fig. 1, the nodes in the proposed neural network model is not fully connected, every node in representational layer is connected with every other *8* nodes in the middle layer, and every node in middle layer is connected with every other *7* nodes in the knowledge layer. The connection parameter between the representational layer and the middle layer is denoted as c_{ij}. If node *i* in the representational layer is connected with node *j* in the middle layer, then $c_{ij} = 1$, otherwise, $c_{ij} = 0$.

2.2 The Algorithm

The 50 sequence numbers in the representational layer are fed into the nodes in the middle layer via the connection parameter c_{ij}. Node *j* in the middle layer will receive *8* numbers: n_j, n_p, n_q, n_r, n_s, n_t, n_u *and* n_v, sent from the *8* nodes *j*, *p*, *q*, *r*, *s*, *t*, *u* and *v* in the representational layer. For example, node *0* in the middle layer will receive n_0, n_1, n_4, n_{25}, n_{30}, n_{35}, n_{40} and n_{45}. The connection between knowledge layer and middle layer also adopts an incomplete connection scheme, in which every node in the knowledge layer is connected with the other *7* nodes in the middle layer. The weight between node *j* in the middle layer and node *k* in the knowledge layer is calculated according to the following equation:

$$w_{jk} = t_p + t_q + t_r + t_s + t_t + t_u + t_v \qquad (1)$$

where:

$$t_p = \begin{cases} +1 & ifn_k > n_p \\ 0 & ifn_k = n_p \\ -1 & ifn_k < n_p \end{cases} \quad t_q = \begin{cases} +1 & ifn_k > n_q \\ 0 & ifn_k = n_q \\ -1 & ifn_k < n_q \end{cases} \quad t_r = \begin{cases} +1 & ifn_k > n_r \\ 0 & ifn_k = n_r \\ -1 & ifn_k < n_r \end{cases}$$

$$t_s = \begin{cases} +1 & ifn_k > n_s \\ 0 & ifn_k = n_s \\ -1 & ifn_k < n_s \end{cases} \quad t_t = \begin{cases} +1 & ifn_k > n_t \\ 0 & ifn_k = n_t \\ -1 & ifn_k < n_t \end{cases}$$

$$t_u = \begin{cases} +1 & ifn_k > n_u \\ 0 & ifn_k = n_u \\ -1 & ifn_k < n_u \end{cases} \quad t_v = \begin{cases} +1 & ifn_k > n_v \\ 0 & ifn_k = n_v \\ -1 & ifn_k < n_v \end{cases} \qquad (2)$$

In the sorting algorithm, the output of representational atoms are tied to the sequence numbers to be sorted. Every update will activate one knowledge atom a_k, and the number n_k in the corresponding representational atom *k* is added into the sorted sequence. The knowledge atom *k* is then disabled and will not be considered in the subsequent updates. If there are more than one knowledge atoms are activated in one update, the numbers in the corresponding representational atoms are added to the sorted sequence at the same time. In this way, the *N* numbers in the sequence will be sorted after a maximum of *N* updates. The output of knowledge atom *k* is given by:

$$a_k = \begin{cases} +1 & ifI_k \geq 0 \\ 0 & ifI_k < 0 \end{cases} \qquad (3)$$

where I_k is the input of knowledge atom *k*, which is calculated according to:

$$I_k = Comp_k - \mathrm{K}_m$$

$$Comp_k = \sum_{j=1}^{N} w_{jk} r_j \qquad (4)$$

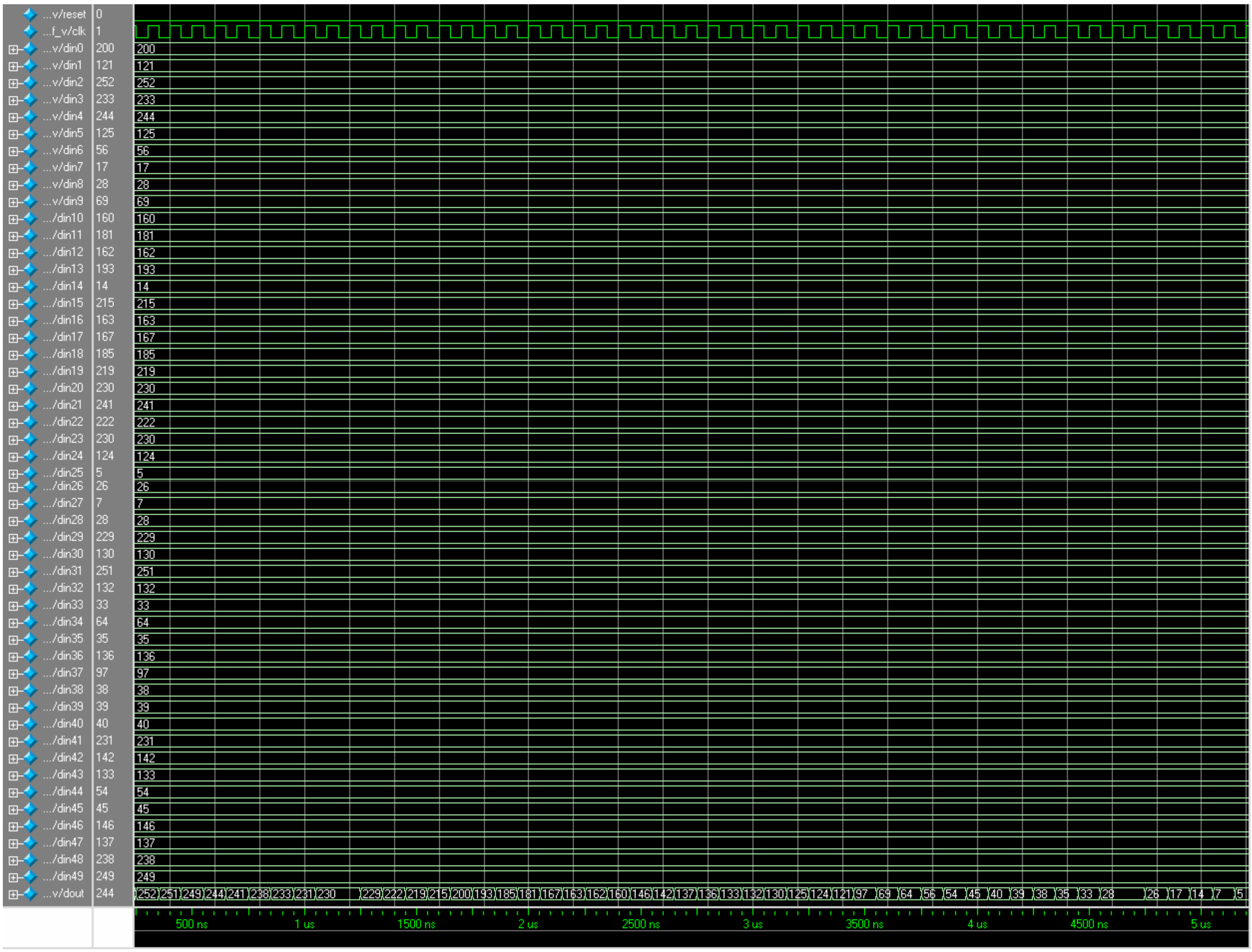
where w_{jk} is the weight between middle layer node *j* and knowledge atom *k*, which is given by equation *(1)*. r_j is the output of middle layer node *j*(always *+1*), *Km* (*m=1,2,…N*) is the progressively decreasing threshold parameter used in the network updates and is given by:

$$K_m = N - 2m \qquad (5)$$

where *m* is the number of network updates.

3. VHDL SIMULATION OF THE PROPOSED SORTING ALGORITHM

The proposed sorting algorithm is implemented on Xilinx Virtex 4 series FPGA. The core of the sorting circuit consists of two types of comparators and two types of adders. One type of comparator is a 2-bit comparator used in calculating sub-rank value *t* in equation *(2)*, a total of about *2500* such 2-bit comparators are required. The other type of comparator is a 7-bit comparator to calculate the output of knowledge atom according to equation *(3)*. The number of 7-bit comparators used equals to the number of knowledge atoms, which is *50*. To calculate the weight w in equation *(1)*, a kind of 4-bit full adder with *7* inputs is used. Since every representation atom has *7* weights, a total of *350* 4-bit adders will be needed in the 50-number sorter. Another kind of 7-bit adder with *7* inputs is used to calculate $Comp_k$ in equation *(4)*. A total of *50* such adders are needed. The proposed sorting algorithm has been described with VHDL to verify its operation. Fig. 2 shows the simulation result of the sorting operation of the 50-number sorter. As can be seen from Fig. 2, a total of 51 clocks is needed to sort a vector of 50 numbers, with the first clock calculating the weights of the network, and every number is sorted out in each clock thereafter.



```
...v/reset  0
...f_v/clk  1
...v/din0   200
...v/din1   121
...v/din2   252
...v/din3   233
...v/din4   244
...v/din5   125
...v/din6   56
...v/din7   17
...v/din8   28
...v/din9   69
.../din10   160
.../din11   181
.../din12   162
.../din13   193
.../din14   14
.../din15   215
.../din16   163
.../din17   167
.../din18   185
.../din19   219
.../din20   230
.../din21   241
.../din22   222
.../din23   230
.../din24   124
.../din25   5
.../din26   26
.../din27   7
.../din28   28
.../din29   229
.../din30   130
.../din31   251
.../din32   132
.../din33   33
.../din34   64
.../din35   35
.../din36   136
.../din37   97
.../din38   38
.../din39   39
.../din40   40
.../din41   231
.../din42   142
.../din43   133
.../din44   54
.../din45   45
.../din46   146
.../din47   137
.../din48   238
.../din49   249
...v/dout   244   252 251 249 244 241 238 233 231 230 229 222 219 215 200 193 185 181 167 163 162 160 146 142 137 136 133 132 130 125 124 121 97 69 64 56 54 45 40 39 38 35 33 28 26 17 14 7 5
```

Fig.2. VHDL Simulation of the proposed 50 number sorting algorithm

By cascading 50 the basic 50-number sorters into a 2500-number sorter, the maximum of 2500 numbers can be found in two clock cycles in parallel. Theoretically, the maximum of 50n numbers can be found in n clocks. This means that the software exponential search problem can be converted into a linear search problem using hardware.

Performance comparison of the basic 10-number sorter was made with other artificial neural network based parallel sorting algorithms[1]. Compared with sorting networks, the main advantage of the proposed sorting algorithm is that its delay is fixed to 3 which is the layer number of the neural network, while the delay of the sorting network is propositional to the number of numbers to be sorted. Another advantage is that the maximum or minimum can be sorted in one step without waiting for the whole series of number to be sorted, which is very useful in implementing the replacement strategy in genetic algorithms, where if the best genome of the new population is not as good as that of the old population, the worst genome of the new population is replaced with the best genome of the old population. The parallel sorting algorithm presented here provides a fast way of implementing hardware replacement operator.

4. CONCLUSIONS

The parallelism of the sorting algorithm presented in this paper lies in two aspects. First in network architecture, the most dense graph of degree 7 diameter 2 is adopted, which ensures that the processing nodes in the network are symmetrical to each other and the number of processing nodes is optimal. Second, the computing load is evenly distributed and highly synchronized in every processing node of the network. Inter-processor communication is kept to a minimum and a high parallel speedup can be achieved.

The parallel sorting algorithm presented in this paper is an extension from 10-number sorter of our previous work[1][2] to 50-number sorter, which in turn proves that the basic parallel sorting algorithm can be scaled very well and easy for hardware implementation. A chip that functions as a maximum/minimum finder of 10/50 analog signals can be designed based on the algorithm presented. Such chips can be cascaded into even larger maximum/minimum signal finders. Applications such as evolvable hardware, image filters can be expected.

REFERENCES

[1] Zhang Bing, Xu Gang, Zhu Ming-Cheng, "A Parallel Sorting Algorithm Based on Harmony Theory Neural Networks", *Proc. Of the 4th IASTED International Conference on Modeling, Simulation and Optimization*, August 17-19, 2004, Hawaii, USA, pp.165-169.

[2] Bing Zhang, Yuan Xu, Mingcheng Zhu, "A Parallel Sorting Algorithm and its Hardware Implementation Based on FPGA", *Journal of Information and Computational Science*, Vol.1, No.3, December 2004, pp.417~422.

[3] F. Comellas and J. Gómez, "New Large Graphs With Given Degree and Diameter", *Graph Theory, Combinatorics and Algorithms*, *Vol 1*, *New York:John Wiley & Sons, Inc.*,1995， pp. 221-233.

Bing Zhang is an Associate Professor in Faculty of Information Engineering, Shenzhen University. He graduated from Beijing University of Aeronautics and Astronautics in 1982; from University of Strathclyde in Glasgow, UK in 1991. He has published one book, over 20 research papers. His research interests are in parallel and distributed processing, embedded system and artificial neural networks.

32 bit Multiplication and Division ALU Design Based on RISC Structure

Yuehua Ding, Kui Yi
Department of Computer Science and Information Engineer, WuHan Polytechnic University,
Wuhan, HuBei Province 430023, China
Email: ykll1903@126.com

ABSTRACT

This paper analyses structure and algorithm of Floating-Point ALU, and implements multiplication and division operation in the homo-hardware circuit. The Floating-Point multiplication and division ALU supports Floating-Point number according with IEEE-754 standard. This ALU adopts 4-Level pipelining structure: ‘0’ operation number check、exponent addition and subtraction operation、fraction multiplication and division operation、result normalization and round. Each step can act as a single module. Among these modules, there are some registers which can prepare necessary data for next operation.

Keywords: Pipelining, IEEE-754 Standard, VHDL

1. INTRODUCTION

On digitalization and information dynasty, digital integrated circuit application is very popular. With the development of micro-electronics technology, craft digital integration runs to ASIC (Application Specific IC) now, but radio tube、transistor、middle-small-scale integration 、 VLSI (Very-Large-Scale Integration) is out of time gradually. ASCI reduces production cost, enhances production physical size and accelerates society digitalization tenor. But for long cycle length of design、high investment of version change and low flexibility, ASCI application range is restricted. With the growth of market demands, Very-Large-Scale、high speed、low consumption new pattern FPGA/CPLD is renovated constantly. New pattern FPGA integrates CPU or DSP (Digital Signal Processing) even. This kind of FPGA can allow you implement software and hardware in the homo-EPGA chip cooperation design, which provides strong hardware support for SOPC (System On Programmable Chip)[1].

This paper introduces multiplication and division ALU can implement CPU ALU Floating-point multiplication and division operation in QuartusII with VHDL language. In addition, this ALU in this paper is designed corresponding to RISC design characteristic. Relation among each module in arithmetic unit and each module function are also illuminated well and truly. This article provides top structure of multiplication and division arithmetic implementation. Pipelining structure can make each module implement time parallelism on running process and enhance the whole CPU arithmetic speed.

2. MULTIPLICATION AND DIVISION WORKING THEORY AND STRUCTURE

To complete 32 bit multiplication and division ALU design, there are 4-Level pipelining: 0 operation number check、 exponent addition and subtraction operation、 fraction multiplication and division operation、result normalization and rounding; Floating-Point number accords with IEEE754 standard. Fraction is expressed by sign magnitude. Exponent is expressed by biasing. This ALU can enhance CPU arithmetic speed and the whole system performance.

2.1 Key Questions to be Resolved

(1) In the process of arithmetic, there are some different details between multiplication and division operation, although they are similar. So INOP operation signal is designed to distinguish the arithmetic type.
(2) Separate multiplication and division arithmetic process correctly. Divide this process into 4 pipelining-segments and assure the 4 modules have each dependent function.
(3) Because of adopting pipelining structure in the design, we must assure mission in the pipelining is in sequence and enter the pipelining uninterruptedly, so the advantage of pipelining--enhancing the CPU performance, can be reflected.

2.2 Key Technology and Complexity Analysis

2.2.1 Adopted Key Technology

The key technology adopted in the design is pipelining structure. Pipelining in computer is just like factory assembly line. To implement pipelining, input mission must be divided into a series of sub-mission at first, which can make sub-mission run parallel at each step of pipelining. Mission entering pipelining uninterruptedly can realize sub-mission parallel. Therefore, pipelining process improves computer system performance by leaps and bounds. It is very economic method to realize time parallel in computer.

2.2.2 Complexity Analysis

(1) For avoiding overflow, It must check divisor is 0 or not. If divisor is 0, error message must be cautioned.
(2) Because of adopting IEEE 754 standard floating-point number in the design, exponent is expressed by biasing. On doing exponent addition or subtraction, [x + y] biasing = [x] biasing + [y] complement;[x - y] biasing = [x] biasing + [-y]complement, so it must assure [y] complement and [-y] complement are correct on subtraction arithmetic [7].
(3) While doing fraction multiplication and division, fraction structure is the 1.m form actually, so the lost “1” is filled in actual arithmetic. The result after arithmetic must be 1.m form too, so the fraction of final result must be processed.

2.3 Structure and Design of Floating-Point Number

2.3.1 Structure of Floating-Point Number

Because multiplication and division ALU is for 32 bit Floating-Point and Floating- Point numbers adopts IEEE 745 standard, the Floating-Point numbers structure in the design shows as Fig.1:

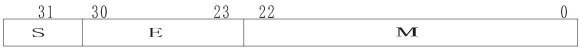


Fig.1. Structure of Floating-Point Numbers

(Caption: In the Floating-Point number structure, M is value of fraction, which is expressed by sign magnitude adopting “hidden bit” expression; E is value of exponent, which is expressed by biasing; S is fraction sign.)

2.3.2 Floating-Point Numbers Multiplication and Division Arithmetic Flow

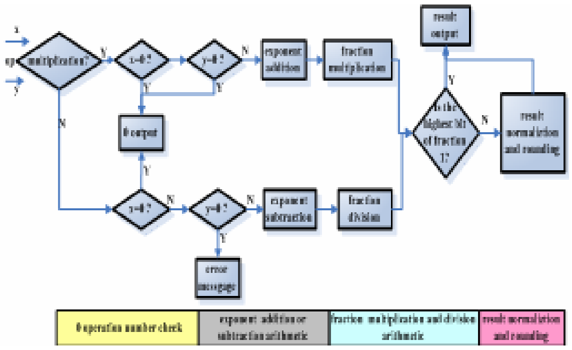

x
y
multiplication?
x=0?
y=0?
exponent addition
fraction multiplication
result output
Is the highest bit of fraction 1?
result normalization and rounding
0 output
y=0?
x=0?
exponent subtraction
fraction division
error message
Y
N
0 operation number check
exponent addition or subtraction arithmetic
fraction multiplication and division arithmetic
result normalization and rounding

Fig.2. Floating-Point Numbers Multiplication and Division Arithmetic Flow

Through analysis of two Floating-Point numbers multiplication and division arithmetic, the multiplication and division ALU adopts pipelining structure and is divided into 4 steps in the whole arithmetic process: 0 operation number check、 exponent addition and subtraction operation、fraction multiplication and division operation、result normalization and rounding. Between each module there are registers designed to store last operation result. The four modules can act as four different sub-missions. As long as input numbers to do multiplication and division arithmetic into the pipelining, the missions can implement each operation parallel and time parallelism is realized too. Adding register among each module meet the pipelining structure demand, because realizing every sub-mission time parallelism in the pipelining structure asks every sub-mission in the process can input pipelining uninterruptedly. But each sub-mission implement time is not same in the whole pipelining. Considering not affect arithmetic of early entering pipelining data, store the data which inputs pipelining later, into every register temporarily. Abstract structure graph designed shows as Fig. 3.

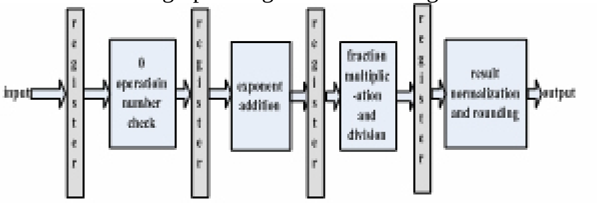


Fig. 3. ALU Design Structure

3. TOP CONNECTION GRAPH AND MODULE DESIGN

3.1 Top Connection Graph

The top connection graph of Floating-Point number multiplication and division ALU shows as Fig. 4、5、6、7:

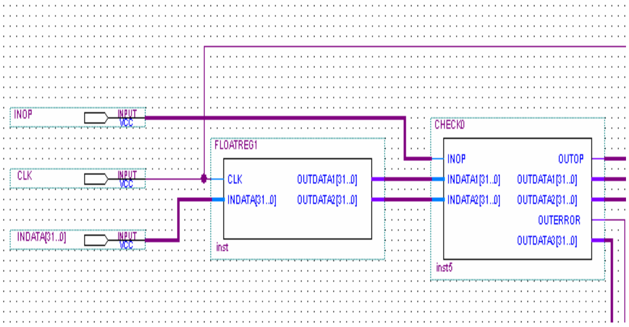


Fig. 4. The First Part of Connection Graphic

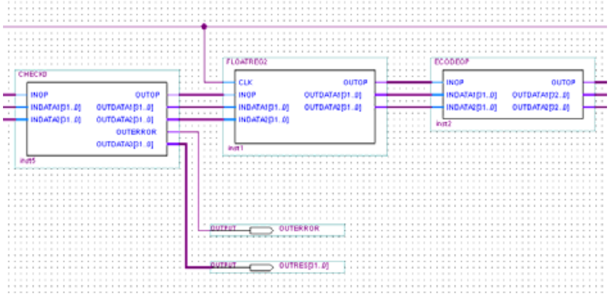


Fig. 5. The Second Part of Connection Graphic

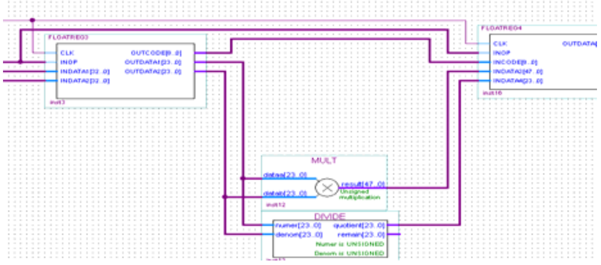


Fig. 6. The Third Part of Connection Graphic

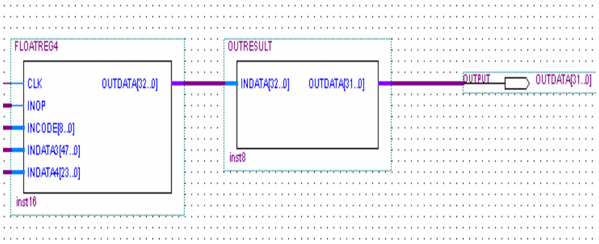


Fig. 7. The Fourth Part of Connection Graphic

(Caption: because of the connection graphic is too big, the whole connection graphic is divided into 4 parts orderly.)

3.2 0 Operation Number Check Module (CHECKO)

This module charges of checking whether two operation numbers 0 or not. On doing multiplication arithmetic, if one operation number is 0 the result must be 0, no matter what another number is. On doing division arithmetic, if divisor is 0, overflow happens. If divisor is not 0 but dividend is 0, the final result must be 0. The abstract chip design shows as Fig.4:

- OUTDATA1、OUTDATA2, shows first data and second data which result is not 0 respectively;
- OUTERROR, shows divisor is 0, that is activating error alarm;
- OUTDATA3, shows final result is 0;
- INOP: because there is difference between multiplication arithmetic and division arithmetic, which can result in difference in the operation of each module, a pin must be designed to distinguish the operation type. It is supposed that on INOP = 0 expressing multiplication arithmetic and on INOP = 1 expressing division arithmetic.
- OUTOP, shows output operation sign;
- INDATA1 、 INDATA2, shows two source operation numbers which comes from upper register.

3.3 Exponent Addition and Subtraction Operation Module (ECODEOP)

The main function of this module is to realize exponent addition and subtraction arithmetic of two numbers. Multiplication rule of Floating-Point number is exponent addition and fraction multiplication; Division rule of Floating-Point number is exponent subtraction and fraction division. Exponent is usually expressed by biasing in computer. Data bit of biasing and data bit of complement is same but sign bit of them is on the contrary. Converting the exponent to complement form can realize exponent addition and subtraction arithmetic. It is supposed that there are two Floating-Point numbers x and y. Complement form of y is taken by means of negating highest bit of 8 bit exponent of y. [x + y]biasing = [x]biasing + [y]complement, which is multiplication operation; After highest bit of 8 bit exponent of y is negated and complement form of 8 bit exponent of y is taken, the complement of [–y] is got. [x – y]biasing = [x]biasing + [–y]complement, which is division operation. When INOP equals 0, [y]complement outputs and adds to [x]biasing, which is multiplication operation. When INOP equals 1, [–y]complement outputs and adds to [x]biasing, which is division operation. If result of exponent overflows, the above mentioned rules does not establish. Exponent of double sign bit ALU can be used and the second sign bit of biasing is defined. Highest bit is always 0 taking part in addition and subtraction arithmetic, so the overflow condition is that highest sign bit of result is 1. If the low sign bit is 0, result overflow shows; if the low sign bit is 1, result underflow shows. If the highest sign bit is 0, overflow does not happen; If low sign bit is 1, result is positive; If low sign bit is 0, result is negative. The two input data in the chip are INDATA1 and INDATA2. On implementing addition and subtraction arithmetic of exponents, taking bits from exponent of each number are getting from the 23rd bit to 30th bit of each number. As for the first number, 8 bits data can be got to take part in arithmetic which are INDATA1(30 DOWNTO 23); But for the second number, DATA2(30) is negated at first. Combine with other bits which are from 23rd bit to 29th bit to take part in arithmetic. The 8 bits are INDATA2(29 DOWNTO 23). In IEEE 754 standard fraction of every Floating-Point number is 1.m default form, but input number dose not show the "1". To ensure correctness of next multiplication and division arithmetic of fraction, output data in this module must show the "1". The "1" adds to fraction of output number, and then 1.m form shows. In this way, fraction takes part in arithmetic with 24 bits on next multiplication and division arithmetic. Detailed chip design shows as Fig.5.

3.4 Resister 3(FLOATERG3)

This register prepares correct data for next fraction multiplication and division. On doing fraction multiplication and division, two fraction signs must be known, so and so only result sign can show correctly. Fetch the sign bit of two numbers and make nonequivalence operation between them. The 31st bit expresses fraction sign in terms of Floating-Point structure. Because all input numbers merge together and REGVALUE(65) expresses all digit bits , the sign of the first number is expressed as REGVALUE(32) and the sign of the second number is expressed as REGVALUE(65). If result of two numbers nonequivalence operation is 1, it expresses that the sign bit of two numbers is opposite and the final output result is negative number. If result of two numbers nonequivalence operation is 0, it expresses that the sign bit of two numbers is sameness and the final output result is positive number. There is one output pin named OUTCODE in the chip, which shows final sign bit and final exponent combine and output together. Detailed chip design shows as Fig. 6.

3.5 Multiplication Arithmetic Module(MULT)

This module charges of multiplication of two input numbers. MULT function unit is called in the design. Its two input pins are dataa[23......0] and datab[23......0]. They are both 24 bits, so the input two numbers are asked to add abbreviatory "1" to the fraction of them. The two numbers which come into the unit is sign magnitude form. Through arithmetic its final result is 48 bits which is [47......0]. The defined output pin is result[47......0]. Detailed chip shows as Fig. 6.

3.6 Division Arithmetic Module (DIVIDE)

This module charges of two input numbers division arithmetic. DIVIDE function unit is called in the design. Two input numbers are also asked to be sign magnitude form in the function unit. There are two input pins in the chip defined as number[23......0] and denom[23......0]. The result through the arithmetic is that quotient[23......0] shows quotient and remain[23......0] shows remainder. Because remainder problem is not considered in the design, only one output pin quotient[23......0] is defined to express final arithmetic result. Detailed chip structure shows as Fig.6.

3.7 Resister 4(FLOATERG4)

This register charges of judging input data before output. Considering that there is 32-bits data bus in the design, the result enters the register with sequence. It must choose different data bit as final output result by judging INOP in the register. Data input into the register is sorted from lowest bit to highest bit. For example, the first data is from 0th bit to 23rd bit and other data sorted on this way. There are some pins in the register, detailed chip design shows as Fig. 7.

- Input pin INOP , judges what arithmetic type is;
- Input pin CLK, clock signal;
- Input pin INCODE[8......0]，shows sign and exponent of final result;
- Input pin INDATA3[47......0]，shows sign result of multiplication arithmetic;
- Input pin INDATA4[23......0]，shows final result of division arithmetic;
- Output pin OUTDATA[23......0], shows data value on getting correct digit bit.

3.8 Normalization and Rounding Module (OUTRESULT)

This module charges for normalization and rounding of final fraction result. Result can not output directly, because there is a problem with data which enter the register: no matter how result is came from multiplication or division arithmetic, it must assure the fraction is 1.m form, which can accord with IEEE 754 form. So fraction that is not 1.m form must be left-shift normalized and the final result is made 1.m form.

Because the problem exists, data can not output directly as result. The problem must be solved in the register: as for fraction is 0.1m form, fraction should be left-shifted and lowest bit fill the emptied with 0; As fraction is 0.01m form, fraction should be left-shifted normalized and lowest bit fill the emptied with 00. Because range of Floating-Point number fraction is between 1/2 and 1, every multiplication result only left-shift normalized two bit at the most. On fraction multiplication the low word length part of the double word length and on fraction division because of indivisibility more bit of quotient will define word length. On normalizing the arithmetic result, the kept bits can be left-shifted into the defined word length of fraction. Round toward + infinity is adopted in the design, so no guard digit need kept.

Put the final result into the output pin OUTDATA[31......0]. This result is the final result of multiplication or division arithmetic. Detailed chip design shows as Fig. 7.

4. CONCLUSIONS

Final hardware in RISC microprocessor based on FPGA is tested on GW48 EDA system, which is produced by Hangzhou Kang-Xin Corp. Through QuartusII4.1 simulation、integration and synthesis assemble, result expresses that the Floating-Point number multiplication and division ALU according with IEEE 754 standard accomplishes expectant function. Pipelining structure implements each sub-mission parallelism operation and enhance system performance of computer.

REFERENCES

[1] Mo-JianKun, Gao-JianSheng,Computer Organization, Huazhong University of Science and Technology Press,1996;

[2] Zheng-WeiMin, *Computer System Structure*,Tsinghua University Press,Oct 2004.

[3] Bai-ZhongYing,*Computer Organization,*Science Press,Nov 2000.

[4] Pan-Song,Huang-JiYe,*EDA Technology Utility Tutorial* [M]．BeiJing：Science Press,2002;

[5] Zhang-XiuJuan,Chen-XinHua,*EDA Design and emulation Practice[M]*,BeiJingLEngine Industry Press, 2003, WeiPu Information.

[6] "IEEE Standard of Binary Floating-Point Arithmetic" *IEEE Standard754, IEEE Computer Society*, 1985.

[7] Jan M. Rabaey,Digital Integrated Circuits- A Design Perspective Prentice—H al l I nternational,I nc,Tsinghua University Press,Feb 1999.

[8] John L. Hennessy David A,“Patterson Computer Organization&Design,”*The Hardware/SoftwareInterface*, Engine Industry Press 1999.9;

[9] *Aldec Active-HDL the Design Verification Company Online Help;*

Offline Handwriting Digital Recognition System Based on Information Granules*

Jianfeng Xu [1] , Leiyue Yao [2], Weijian Jiang [2]
[1] School of Software, NanChang University
NanChang, Jiangxi, 330029, China
[2] Computer Technology Department, Jiang Xi Blue Sky University
NanChang,Jiangxi, 330098 ,China,
Email: [1] Jianfeng.nc@gmail.com , [2] ylyyly2001@163.com

ABSTRACT

Offline handwriting digital recognition is a classical problem in pattern recognition. In this paper, we present an approach to recognize offline handwriting digits which is based on information granules. Moreover a matching algorithm for decision rules by using information Granule template is also proposed. The effect of the algorithm has been demonstrated by experiment.

Keywords: Information Granule, Pattern recognition, Decision Rule, Handwriting Digit

1. INTRODUCTION

Information granules and Granular Computing is a very important subject which has been developing dramatically in recent years, and this subject is also playing an important in a lot of fields. Granular Computing is geared to representing and processing basic chunks (i.e. granules), Granules are viewed as collections of entities, which are arranged together due to their similarity, functional relativity, indiscernibility, etc. The procedure of generating granules is called granulating [1, 2, and 4]. Granulating of the universe is the fundamental issue in granular computing. Granularity and granules play a crucial role in many areas of knowledge representation and problem solving. Granulation could help us to reduce an overall computing effort. Without granulation, large data sets are often computationally infeasible.

Basic concepts and principles of granular computing, though under different names, have in fact been appeared in many related fields, such as programming, artificial intelligence, divide and conquer, interval computing, quantization, data compression, chunking, cluster analysis, fuzzy and rough set theories, quotient space theory, belief functions, machine learning, databases, and many others. In the past few years, we have witnessed a renewed and fast growing interest in Granular Computing (GrC). Many applications of granular computing have appeared in a lot of fields, such as medicine, economics, finance, business, environment, computer and information science. So it is an important research direction that picking up special information granule's structure mode and realizing approximate reasoning .For the cause we proposed and discussed the system of offline handwriting digital recognition based on Information Granules in the paper.

2. OFFLINE HANDWRITINF DIGITAL RECOGINITION SYSTEM BASED ON INFORMATION GRANULES

The offline handwriting digital recognition is a classical problem in pattern recognition; it has a wide application prospect in many aspects, such as postal service, banking system, etc. But this problem is still not really perfectly solved in the field of Characters Discern. In 1929, Tauscheck had found a device that attempted to discern Arabic numerals by matching the characters of the optical picture elements (Fig.1 is the picture of the device) [3].The method of character matching is the basic of almost all techniques involved in Characters Discern. Therefore it has multiple meaning that we study the offline handwriting digital recognition that based on Information Granules.

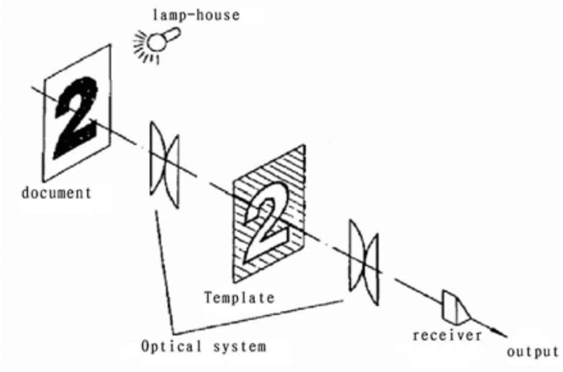


Fig.1. The picture of the device

2.1 Granulating of Information

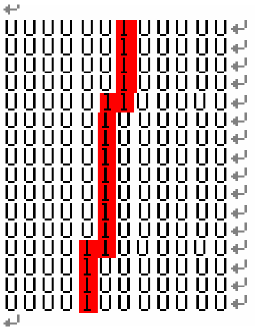
Definition 1 DT=（U, A∪d）is a decision table of information system,let U be the universe of the objects, A be the condition attribute set,d be the decision attribute set. Any granule of decision rule φ→ψ of DT is denoted by (φ,ψ) and m(φ,ψ),where(φ ,ψ) is formula defined , m(φ ,ψ) is the meaning set corresponding to (φ ,ψ).

Definition 2 A information Granule of a decision rule φ→ψ ≡ φ1∧ φ2∧ … φi ∧… ∧ φn →ψ ,where φi is the child Granules of condition attribute Granules, i=1,2, …n.

Eg: Firstly, the ten study samples of Arabic numerals have been equidistant partitioned as 192(16×12) grid points . Some instances showed as Fig. 2.Each grid is a condition attribute. If a grid be drew, the attribute value of the grid would be 1, else would be 0.The granulating showed as follows:

DT=(U,C∪D),
U:{0,1,2,3,4,5,6,7,8,9},
C:{16×12 grid points | the value of the grid point is 0 or 1 } ,
D:{0,1,2,3,4,5,6,7,8,9}.
Then we can pick up the decision rule Granules. Take 1 of study samples for example Fig.2.

* This work was Supported by the Natural Science Foundation of Jiangxi China(0311101)



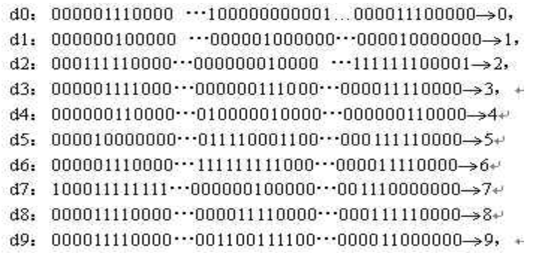
```
UUUUUU1UUUUU
UUUUUU1UUUUU
UUUUUU1UUUUU
UUUUUU1UUUUU
UUUUU11UUUUU
UUUUU1UUUUUU
UUUUU1UUUUUU
UUUUU1UUUUUU
UUUUU1UUUUUU
UUUUU1UUUUUU
UUUUU1UUUUUU
UUUUU1UUUUUU
UUUU11UUUUUU
UUUU1UUUUUUU
UUUU1UUUUUUU
UUUU1UUUUUUU
```

Fig.2. Study sample of 1

d1:（1,1）0 ∧（1,2）0 ∧···（I,j）1∧···∧ （16,12）1 ··· →1. where（I,j）1 be the grid attribute point of Ith row and Jth column,the value of the grid point is 1;（I,j）0 be the grid attribute point of Ith row and Jth column，the value of the grid point is 0.

d1 can be simplified as:
d1:000000100000 ···000001000000···000010000000→1.

Therefore we can make the decision table (Information Granules template) of the ten decision rule Granules. As Fig.3, each column be representation of the same grid attribute point, each row be representation of a decision rule Granules which was picked up from each study sample.



```
d0: 000001110000 ···100000000001…000011100000→0,
d1: 000000100000 ···000001000000···000010000000→1,
d2: 000111110000···000000010000 ···111111100001→2,
d3: 000001111000···000000111000···000011110000→3,
d4: 000000110000···010000010000···000000110000→4
d5: 000010000000···011110001100···000111110000→5
d6: 000001110000···111111111000···000011110000→6
d7: 100011111111···000000100000···001110000000→7
d8: 000011110000···000011110000···000111110000→8
d9: 000011110000···001100111100···000011000000→9,
```

Fig.3. Information Granules template

2.2 Neighborhood Extend of Condition Attribute Granules

Definition 3 Decision rule Granules :

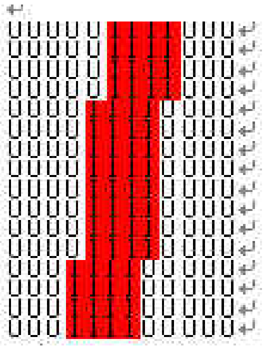
$\varphi 1 \wedge \varphi 2 \wedge \varphi i \wedge \cdots \wedge \varphi n \rightarrow \psi$;

Extended Granules:

$(X1) \wedge (X2) \wedge Xi \cdots \wedge (Xn) \rightarrow \psi$,

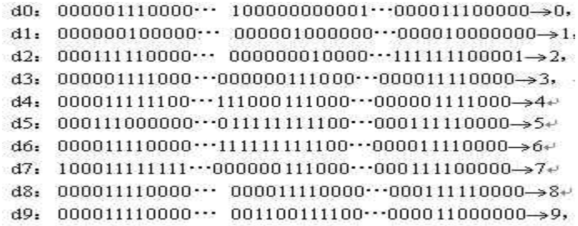
Where φi is the Ith condition attribute Granule, ψ be decision attribute Granule, Xi be Neighborhood set of φi. The extension of the Neighborhood can be decided by experience or statistics

Fig. 4 is the effect of the extension of the ten study samples. Fig.5 is the information Granules template of the extension of the ten study samples.



```
UUUUU1111UUU
UUUUU1111UUU
UUUUU1111UUU
UUUUU1111UUU
UUUU1111UUUU
UUUU1111UUUU
UUUU1111UUUU
UUUU1111UUUU
UUUU1111UUUU
UUUU1111UUUU
UUUU1111UUUU
UUUU1111UUUU
UUU1111UUUUU
UUU1111UUUUU
UUU1111UUUUU
UUU1111UUUUU
```

Fig.4. The effect of the extension



```
d0: 000001110000··· 100000000001···000011100000→0,
d1: 000000100000··· 000001000000···000010000000→1,
d2: 000111110000··· 000000010000···111111100001→2,
d3: 000001111000···000000111000···000011110000→3,
d4: 000011111100···111000111000···000001111000→4
d5: 000111000000···011111111100···000111110000→5
d6: 000011110000···111111111100···000011110000→6
d7: 100011111111···000000111000···000111100000→7
d8: 000011110000··· 000011110000···000111110000→8
d9: 000011110000··· 001100111100···000011000000→9,
```

Fig.5. Information Granules template of the extension study samples.

2.3 Reduction

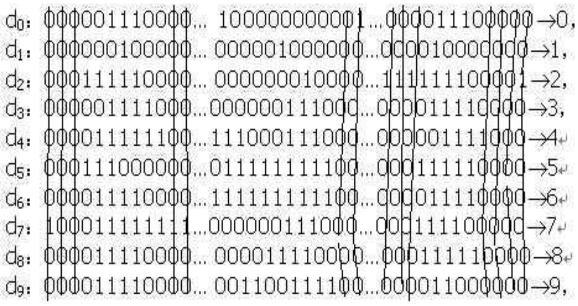
Definition 4 Determinacy factor:

$$cer(\varphi,\psi)=|m(\varphi)|/|U|$$

where $|m(\varphi)|$ be the individual number of set which has the same condition attribute value as φ→ψ in the information Granules template. |U| be the number of decision rule Granules in the information Granules template.

Proposition 1 Reduction of the information Granules template: If the Determinacy factor of condition attribute (which value is 0) ≥ λ, where λ can be got by experience . The condition attribute of the decision table would be reduced.

After reduction the information Granules template of Fig5 the template remains 149 condition attributes. The process of the reduction as Fig.6(set λ＝0.9).



```
d0: 000001110000... 100000000001...000011100000→0,
d1: 000000100000... 000001000000...000010000000→1,
d2: 000111110000... 000000010000...111111100001→2,
d3: 000001111000...000000111000...000011110000→3,
d4: 000011111100...111000111000...000001111000→4
d5: 000111000000...011111111100...000111110000→5
d6: 000011110000...111111111100...000011110000→6
d7: 100011111111...000000111000...000111100000→7
d8: 000011110000... 000011110000...000111110000→8
d9: 000011110000... 001100111100...000011000000→9,
```

Fig.6. Process of the reduction

2.4 Binary Information Granules Template Matching Algorithm

Definition 5 Similarity degree of two binary information Granules:

Clp(G, G’)=Card˜ (G Δ G’)/Card˜ (G), where G≠∅ , Card(G)˜ be the sum of 1 in the G , Δ be iff operation according to bit.

eg：given G: 100101, G’: 111000,
Card˜（100101 Δ 111000)/Card˜ (100101)
＝Card˜（100010)/ 3＝2/3.

Proposition 2 Let set G and G’ are two binary decision rule information Granules in an information system. If the condition Granules of the two Granules have same similar degree P, The decision Granules of the two Granules are P too.

Definition 6 Algorithm for binary information Granules template matching:

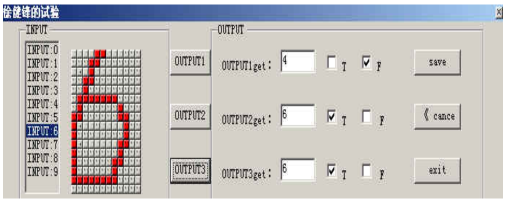
Step1: Pick up the decision rule Granules and construct the information Granules template.

Step2: Pick up the condition attribute Granules of the new numeral samples which would be identified

Step3: Compute the similarity degree of the condition attribute Granules and every decision rule Granule which belongs to the information Granules template. The decision value of the decision rule Granules which has the maximal similarity degree with the new numeral sample is the result class.

3. EXPERIMENT

Run the algorithm for binary decision rule Granules matching by the three information Granules template (Fig.3, Fig.5, Fig.6) respectively. Finally we got three output mode (output1, output2, output3). Fig.7 is the cut chart of our experiment .



OUTPUT1get: 4 □ T ☑ F

OUTPUT2get: 6 ☑ T □ F

OUTPUT3get: 6 ☑ T □ F

Fig.7 experiment system cut chart

The experiment were carried out 5000(500×10) times, the correct rate of output1 is 58.4%, the correct rate of output2 is 85%, the correct rate of output3 is 80.6%. Fig.9 is the block diagram of the each numeric correct rate in the experiment.

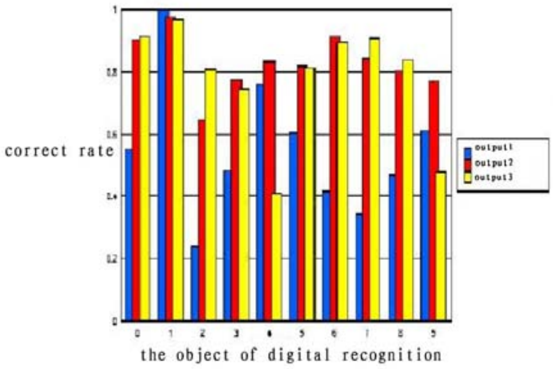


Fig.8 block diagram of every numeric correct rate

4. CONCLUTIONS

In this paper the offline handwriting digital recognition system were studied by using information Granules template based on the original character matching methods. In recent years a lot of new methods were invented in the field of digital recognition. Obviously, the integrated approaches with information Granules should be a future research work.

REFERENCRES

[1] Qing Liu, *Rough Sets and Rough Reasoning,* Science Press., Beijing, 2003 [M],Second,.(In Chinese)

[2] Qing Liu, Granules and Reasoning Based on Granular Computing, Lecture Notes in Artificial Intelligence 2718, 16th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE 2003, Uk, June 2003, Proceedings,516-526.

[3] Zhang honglin, *Digital picture mode recognition technology and project practicing post & telecommunications.* Press., Beijing, 2003 [M],.(In Chinese)

[4] Y.Yao and N.Zhong, "Granular Computing Using Information Tables," In: *Data Mining, Rough_ Sets and Granular computing,* Physica-verlag, Berlin,2002, 102-124.

Jianfeng Xu master is a Docent. Of Computer Science & Technology Department in School of Software, The NanChang University He graduated from NanChang University in 2006 with specialty of Computer Science; His research interests are in Granule computing, AI and e-commence.

Leiyue Yao master is a Docent. of Computer Science & Technology Department in Jiang Xi Blue Sky University. He graduated from NanChang University in 2006 with specialty of Computer Science; His research interests are in RIA and e-commence.

An Efficient Solution for the Scoped Memory in RTSJ

Yang Li, Wenbo Xu
School of Information Technology, Southern Yangtze University
Wuxi, Jiangsu, 214122, China
Email: crazy55600@hotmail.com

ABSTRACT

This paper proposes the approach of how to improving the performance of memory management for RTSJ. The RTSJ imposes strict assignment rules from or to memory areas preventing the creation of dangling pointers and thus it can maintain the pointer safe in the real-time Java system and the distributed real-time Java system. The implementation requires for some cases ensuring the checking of these rules at run-time, which adversely affects both the performance and predictability of the RTSJ and DRTSJ application. This paper presents an efficient pattern for managing scoped areas which can avoid violating the single parent rule and check the illegal reference. Accordingly to our implementation we use a tree and each thread a stack to support the design pattern.

Keywords: RTSJ, DRTSJ, Single Parent Rule, Display, Scoped Memory, Illegal References, Dangling Pointers, Write Barrier

1. INTRODUCTION

Garbage collection is terrible for real-time systems. Most garbage collectors cause the system to stop and collect garage at hard-to-predict intervals. Especially in the distributed real-time Java system, it is more complex to manage the real-time garbage collector remotely. Region-based memory introduced in RTSJ[1] can solve this problem efficiently. But the scoped areas can be reclaimed at any time, so objects within an area with a longer lifetime are not allowed to create a reference to an object within another area with a potentially shorter lifetime. Therefore the checking of the illegal references during the run-time is necessary. But the distributed real-time Java system introduced in DRTSJ[2] requires more constantly and predictably checking and creation of scoped memory time, so the current algorithm for managing the scoped memory is not efficiently enough. In this paper we will introduced a more efficient model for managing scoped memory.

The run-time checking requires the introduction of write barriers[5] which has been the subject of[6]and[7]. The idea of using both write barrier and a stack of scoped areas ordered by life-times to detect illegal inter-area assignments was first introduced in[8] But the efficiently using of the illegal references checking based on write barrier requires a data structure that can contain the memory areas, so we will propose a data structure and a algorithm that can avoid violating the single parent rule and check illegal references during the creation of the data structure.

The rest of the paper is organized as follows. Section 2 describes the single parent rule described in the current RTSJ and then proposes the problems of it. Section 3 describes the data structure that can maintain the scoped memory and convenience the checking of the illegal reference. Section 4 shows experiment results. Section 5 draws the conclusion.

2. SINGLE PARENT RULE

2.1 The Scoped Memory Model in RTSJ

RTSJ assumes the Java has its traditional threads, and adds two new real-time thread types: RealtimeThread and NoHeapRealtimeThread, they have some access rules. A traditional thread can allocate memory only on the traditional heap. Real-time threads may allocate memory from a memory area other than the heap by making that area the current allocation context. A scoped memory allocation context is entered calling the MemoryArea.enter() method or by starting a real-time thread whose constructor was given a reference to an instance of MemoryArea. Once a scope is entered, all subsequent uses of the new keyword, within the program logic, will allocate the memory from the current scope. When the scope is exited by returning from the enter() method, all subsequent uses of the new operation will allocate memory from the memory area associated with the enclosing scope. A real-time thread is associated with a scope stack containing all the memory areas that the thread has entered but not yet exited.

To use the nested memory safely, it is necessary to consider in more detail the nesting of memory areas. The real-time JVM will need to keep track of the currently active memory areas of each schedulable object. Every time a thread enters a memory area, the identity of that area is pushed onto the stack. When it leaves the memory area, the identity is popped off the stack. The single parent rule guarantees that a parent scope will have a lifetime that is not shorter than of any of its child scopes, which makes safe references from objects in a given scope to objects in an ancestor scope, and forces each scoped area to be at most once in the tree containing all area stacks associated with the tasks that have entered the areas supported by the tree. The single-parent rule also enforces every task that uses an area to have exactly the same scoped area parentage. The implementation of the single-parent rule as suggested by the current RTSJ edition makes the behavior of the application non-deterministic. In the guidelines given to implement the algorithms affecting the scope stack, the single parent rule guarantees that once a thread has entered a set of scoped areas in a given order, any other thread is enforced to enter the set of areas in the same order. Notice that determinism is an important requirement for real-time applications.

2.2 The Situations that will violate the Single Parent Rule

Let us consider four scoped areas: A, B, C and D, and two threads t1 and t2. We supposed that the thread t1 enters these four scoped areas in the following sequence: first A again B again C again D and the entering sequence of t2 is: first A again B again D again C. We further consider that during the execution time thread t1 enters A, B, C advanced, and then thread t2 enters A, B, D and C. In this situation, the scoped memory C will have two different parents: B and D (see Fig.1), and this will violate the single parent rule. Note that if the thread t1 has excited from all the scoped memory and t2 has excited from the area C, the scoped memory areas C will be reclaimed while the area D is still alive, and then pointers from objects allocated in D to objects allocated in C are dangling pointers.

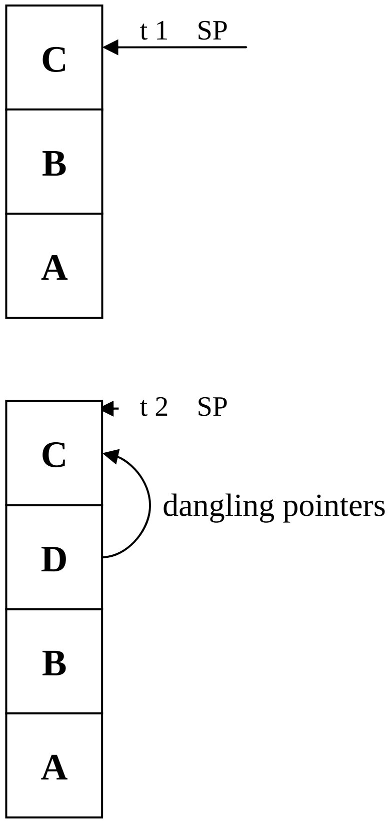


Fig.1. C's two different parents

3. THE IMPLEMENTATION OF THE DATA STRUCTURE

3.1 Checking Illegal References

As stated in the RTSJ the ancestor relation among scoped memory areas is defined by the nesting areas, and this parentage is supported by the area tree. We apply the technique based on displays that has been presents in[3] (see Fig.2). In this tree, direct edge from node B to node A, means that A is the parent of B. The advantage of this formulation is that subtype-testing algorithms[4] can be applied to the parenthood tree to determine legal references. Algorithm 1 contains the pseudo-code shows how to uses this tree to check the illegal references.

```
Algorithm 1:
Input: MemoryArea A, MemoryArea B (reference from
A to B)
start
      ValidReference = false;
      if   A.depth >= B.depth then
            if   B.depth = -1 then
                  ValidReference = true;
            else
                  if   A.display[B.depth] = B then
                        ValidReference = true;
end
```

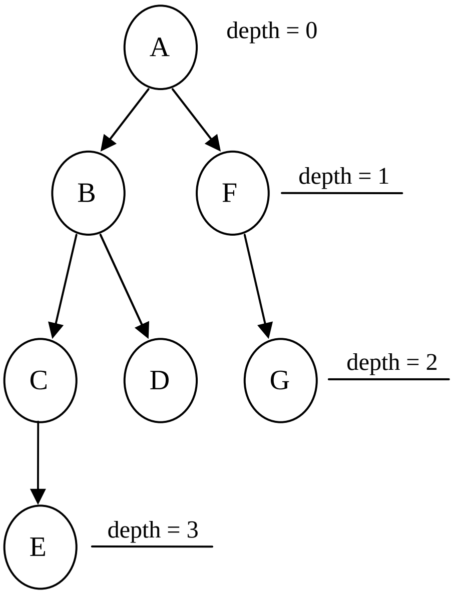


Fig.2. Tree 1

3.2 Maintaining the Tree Data Structure

We next show how to construct area tree data structure which can maintain the single parent rule and check the illegal references.

To understand the main idea about our technique for implementing the tree structure we consider a generic scope areas tree. In the tree each node represents a scoped memory area and then each node has a reference-count that account the number of the threads that has entered into this memory area. When the thread enters this memory region the reference count of this memory's node add one and when the thread exit from this memory region the reference count of this memory's node reduce one. When the reference count of the memory area goes to zero this memory area can be collected. The tree is uses for to maintain the ancestor relations of the memory regions. Once the memory region's ancestor relation is determined, other threads will not be able to change them but only to follow this relation. After this relation is recycled and the old relation is broken other thread will be able to establish the new relation. In addition each thread also must have a stack which can record the memory regions it has entered. In order to maintain single parent rule we will provide an algorithm for creating the tree data structure.

```
Algorithm 2: thread t1 enters scoped area A
If A does not exist in the tree it means that there is no thread
enters into A, so we must join A into the tree first and then
create its ancestor relation by traversing all ancestors of the
current memory area of t1:
Join the A to the tree as the child of the current memory area
of t1;
      for( traverse all ancestors of A)
      {
            if   ( the ancestor of the current memory area of t1
does nor exist in the stack of t1 )
            {
                  Add the reference count of the ancestor one;
                  Add the reference count of A one;
                  Push A and its ancestor into the stack of t1;
            }
            else
```

```
        {
              Add the reference count of A one;
              Push A into the stack of t1;
        }
  }
```

If A has existed in the tree it means that some threads has entered into A, so we should maintain its ancestor relation in the tree by traversing all its ancestors.

```
  for( traverse all ancestors of A)
  {
        if    (the ancestor of A does nor exist in the stack of t1)
        {
              Add the reference count of the ancestor one;
              Add the reference count of A one;
              Push A and its ancestor into the stack of t1;
        }
        else
        {
              Add the reference count of A one;
              Push A into the stack of t1;
        }
  }
```

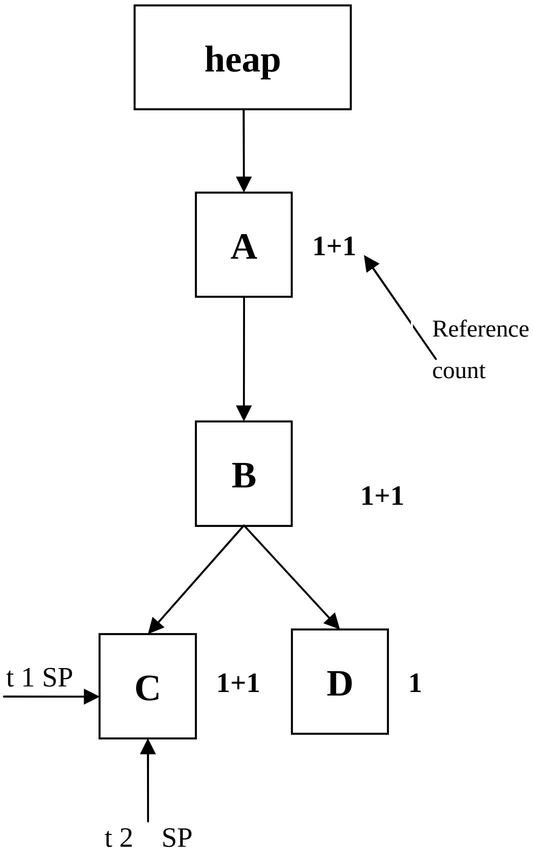


Fig.3. Tree 2

Now let us review the example in the chapter 2.2. According to our algorithm we can get the tree (see Fig.3). In this tree the memory area C has only one parent, so the pointers from objects allocated in D to objects allocated in C is not allowed, it will be checked out and the system can throw an exception.

4. EXPERIMENT RESULTS

To test our implementation algorithm generic scope area tree, we modify the SableVM interpreter base on three algorithms: our algorithm generic scope area tree, jRate's scope-stack implementation algorithm, and the proposed implementation algorithm each thread a stack.

Table 1 shows that the generic scope area tree algorithm can avoid violating single parent rule much more efficiently.

	Executed Bytecodes	Object Accesses	Violating single parent rule %
Generic area tree	2,867,000	608,000	41.60
jRate's algorithm	3,158,000	1,240,000	46.82
RTSJ algoritm	2,907,000	862,000	52.64

Table 2 shows the average check time.

	Parent Traversal	push new scope area	Check illegal references
Generic area tree	68.539ns	22.886ns	36.901ns
jRate's algorithm	56.840ns	14.924ns	50.762ns
RTSJ algoritm	42.802ns	16.954ns	78.580ns

The parent traversal and pushing of generic area tree do not execute as efficiently as other tow algorithms because our algorithm must maintain a generic tree. The execution of checking illegal references is much more efficiently. But in the Real-time system especially in the distributed Real-time system the execution of the checking illegal references are more frequently. So the generic area tree algorithm will work better

5. CONCLUSIONS

In this paper we have introduced an implementation model of the data structure that can avoid violating the single parent rule and support an efficiently illegal reference checking. This algorithm can make the real-time Java and distributed real-time Java programming simple and safe. Even if the programmer violate the single parent rule the generic scope areas tree can maintain the ancestor relation of the scoped memories instead of throwing an exception during the run time. Because we apply the displays so the checking of illegal reference can be more efficiently than the proposed implementation in the current RTSJ. Because this algorithm can provide more constantly and efficiently memory management it can make the using of scoped memory in distributed real-time Java system more efficiently.

REFERENCES

[1] The Real-time for Java Expert Group, "Real-Time Specification for Java," RTJEG 2006. http://jcp.org.

[2] The JSR-50 Home Page, http://jcp.org/en/jsr/detail?id=50.

[3] A.Corsaro and P.K Cytron, "Efficient Reference Check for Real-time Java," ACM SINGPLAN Conference on Languages, Compilers, and Tools for Embedded Systems, LCTES 2003.

[4] Norman H. Cohen, “Type-Extension Type Tests Can Be Performed In Constant Time,” ACM Transactions on Programming Languages and Systems (TOPLAS), 1991.

[5] D.J Cannarozzi, M. P. Plezbert, and R. K. Cytron, “Contained Garbage Collection,” in *Proc. Of the Conference of programming Languages Design and Impementation (PLDI)*, ACM SIGPLAN, May 2000.

[6] M. T. Higuera, V. Issarny, M. Banatre, G. Cabillic, J. P. Lestot, and F. Parain, “Memory Management for Real-time Java: an Efficient Solution using Hardware Support,” *Real-time Systems journal*, Kluber Academic Publishers, to be published.

[7] M. T. Higuera, V. Issarny, M. Banatre, G. Cabilic, J. P. Lesot, and F. Parain, “Region-based Management for Real-time Java,” in *Proc of the 4th International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC)*, IEEE 2001.

[8] M. T. Higuera, V. Issarny, M. Banatre, G. Cabillic, J. P. LESOT, and F. Parain, “Area-based Memory Management for Real-time Java,” in *proc. of the 4th International Symposium an Object-Oriented Real-Time Distributed Computing (ISORC)*, IEEE 2001.

Yang Li, male, master graduate student, his research interests are in distributed real-time Java system;

Wenbo Xu, male, Full Professor, his research interests are in the artificial intelligence, the computer control, inserts the type operating system, the parallel computation, the pattern recognition.

Application of AMCCS5933 Controller in PCI BUS

Kui Yi, Yuehua Ding
Department of Computer Science and Information Engineer, WuHan Polytechnic University,
Wuhan, HuBei Province 430023, China
Email: ykll1903@126.com

ABSTACT

With the development of computer technology, PCI peripheral bus is the most popular bus in the Pentium computer because of its excellent performance. But PCI bus criterion is so complicated that it is difficult to do interface decode. PCI data and address time division multiplex and their operation is synchronized with 33MHZ bus clock. Firstly, this paper introduces PCI bus fundamental theory and interface implementation schemes simply. In allusion to PCI interface circuit implement's complication, the paper introduces AMCCS5933 interface chip scheme, and expatiates AMCCS5933 internal structure, working theory, data transfer mode, function characteristics, interface driver and its application in PCI card design with details.

Keywords: PCI Bus, AMCCS5933, Controller Driver

1. INTRODUCTION

PCI bus is the most popular bus in computer. PCI bus is a kind of 32/64 bit address and data multiplexer bus which supports for burst transfer. Its highest speed can reach 528MB/s and configure automatically. For PCI bus protocol's complication it is difficult to implicate its interface circuits, so adopting the universal PCI interface chip is very usual. Universal PCI interface circuits supporting well for PCI protocol and providing interface for card designer, which reduce card design's work time. There are AMCCS59XX serials of AMCC(Applied Micro Corporation) Corp. and PLLLX serials of PLXTEGH Corp. interface circuits. This paper introduces PCI bus AMCCS59XX serials of AMCC Corp. circuits theory and application.

2. PCI BUS INTRODUCTION

PCI（Peripheral Component Interconnect Special Interest Group）acts as a kind of synchronous 32 or 64 peripheral component interconnection independent on processor. PCI highest working frequency is 33MHZ. PCI is not controlled by processor. It provides a bridge for interaction between processor and high speed peripheral equipment, and can be a traffic administrator among bus. PCI highest throughput is 132Mb/s on 32bit bus, so it is very suit for internet adapter, hard disk driver, dynamic video card, graphic card and high speed peripheral equipment. Fig.1shows a computer system frame based on PCI.

The connection to PCI bus device has master device and target device. PCI support multi threaded bus master device. PCI adopts address and data multiplex bus structure which reduces signals greatly. Target device needs at least 47 signals and master device needs at least 49 signals. PCI bus's important feature is configuration space. Configuration space provides a configuration association. The association system is suitable for present or future configuration mechanism. The association can realize parameter auto configuration and make device compatible with PCI plug-and-play really.

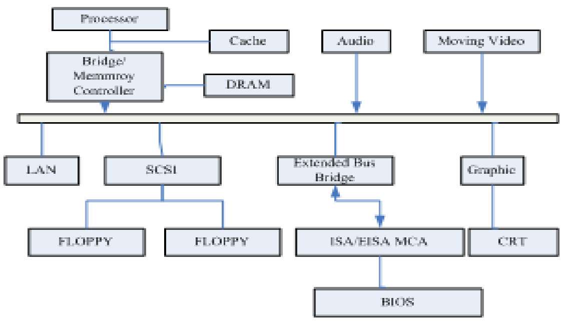


Fig.1. PCI System Structure Diagram

3. IMPLEMENT OF PCI BUS INTERFACE

PCI bus has burst and concurrency data transfer mode. It has special configuration space supporting for Plug-and-Play. PCI bus address/data and controller signals are generated by microprocessor. PCI bus, as synchronization bus, is independent on CPU and supports burst transfer. Burst transfer is composed with address segment and afterwards segments. For support Plug-and-Play function, PCI bus defines 256 bytes configuration space. Some space is used for read. Some space is used for read and write. PCI bus space configuration is generally composed by un-volatility RAM. Furthermore, PCI bus defines district characteristic. Developing application of PCI bus is difficult, so it is very urgent to develop and apply PCI interface validly in Pentium machine. For these reasons, AMCC Corp. produces a kind of PCI interface controller S5933. It implements PCI bus standards. Its structure shows as Fig.2. S5933 support three kinds of physical bus interfaces: PCI bus interface, extended bus interface and un-volatility RAM bus interface. Un-volatility RAM bus interface is used for mapping PCI configuration space and initiating program of device; Extended bus can connect to peripheral equipment. Data transfer between PCI bus and extended bus can be implemented by three methods. [1]

4. AMCCS5933 WORKING THEORY

S5933 provides three kinds of data transfer mode: MAILBOX mode, FIFO mode and PASS THRU mode. S5933 can act as not only PCI target device, but also PCI master device. PCI configuration space can be configured by serial or parallel EEPROM. EEPROM controls bus operation and data transfer by means of driver program setting up

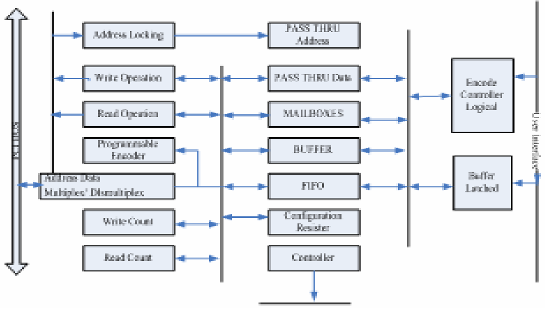


Fig.2. S5933 Function

bus controller register. There are two kinds of bus controller register which can be set: PCI bus controller register and ADD-ON bus controller register. PCI bus is used for controlling PCI bus. ADD-ON bus is used for ADD-ON bus operation. [2]

PCI bus controller register is only accessed by host system. Host system access these register by means of PCI bus I/O writing and reading operation. ADD-ON bus controller register is accessed only by CPU, its pins shows as below:

- ADR[6：2], BE[3：0]#, SELECT#, WR#, RD# and data address bus DQ[31：0]. S5933 pin architecture shows as Fig. 3.
- ADR[6：2]:provides address of accessing register
- BE[3：0]#: defines access bytes which is in double bytes
- WR#: writing enabled
- RD#: reading enabled
- SELECT#: operating enabled
- #: low level enabled
- MODE pin: defines add-on data bus is 16 bit or 32bit. If data bus is 32bit, it is low level. If data bus is 16 bit, it is high level. When ADD-ON bus is 16 bit, BE[3：0]# is used for ADR1 address bus.

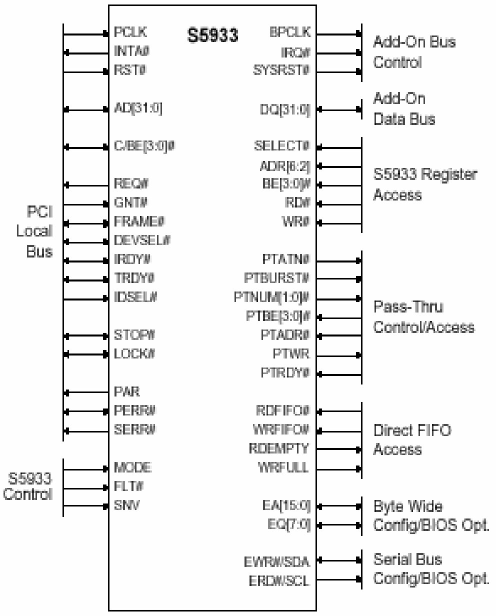


Fig.3. S5933 Pin Architecture

5. AMCCS5933 DATA TRANSFER MODE FEATURES

5.1 MAILBOX Data Transfer Mode

S5933 has eight 32-bit mailbox registers. Mailbox is used for transferring instruction and controlling information between ADD-ON and PCI bus. ADD-ON and PCI both have four input mailboxes and four output mailboxes. Input mailbox of PCI bus corresponds to output mailbox of ADD-ON bus. Output mailbox of PCI bus corresponds to input mailbox of ADD-ON bus.

There are two methods which can monitor mailbox state. PCI bus and ADD-ON bus both have one mailbox which shows empty/full state of mailbox's bytes. Programming for mailbox is another method, which can generate interrupt for PCI interface or ADD-ON interface.

5.2 FIFO Data Transfer Mode

AMCC S5933 inner has two unilateral FIFO: output FIFO and input FIFO. S5933 can use its FIFO interface for DMA(bus master controller) transfer on PCI. It has two independent FIFO. Each depth of FIFO is 8. Each width of FIFO is 32 bit. One transfers data from PCI bus to ADD-on. Another FIFO transfers data from ADD-on to PCI bus. Host system and CPU can access FIFO by means of accessing register.

5.2.1 Implementation of DMA

It can implement DMA transfer in the condition of FIFO. Step of DMA transfer implementation is:

(1) Define DMA transfer controller, which is host system or local logic, by way of software configuration. It is supposed that host system controller mode.
(2) Configure relative PCI bus controller register.

5.2.2 Initiation of DMA Transfer

Initiate DMA transfer by means of configure MCSR register.

Step of DMA transfer controlled by local logic is only to change controller mode to local logic mode. Configure relative register by way of local logic.

5.3 PASS THRU DATA Transfer Mode

PASS THRU mode provides a direct thoroughfare of address and data from PCI to ADD-ON. This mode can implement interface card such as VGA display function.

Configuration space has six base address registers. Base address register 1-4 is relative to PASS THRU mode. Base address register 0 is defined for assign I/O resources in S5933. Base address register 6 is kept.

Pins of PASS THRU mode which are relative of data transfer are PTATN#, PTRDY#, PTNUM[1：0], PTBE[3：0], BE[3：0]#, PTADR#, PTWR, PTBURST#, DQ[31：0].

PTNUM[1：0]: shows which base address register allocate memory space or I/O space.

Data transfer of PASS THRU mode is completed mainly by the pins interaction and handshake between S5933 chip and user circuits. Time sequence of PASS THRU handshake can be implemented with synchronous state machine mode by means of programmable logic and also can be implemented with VHDL hardware program language easily.

5.4 Configuration and Generation of Interrupt

AMCC S5933 has two interrupt pin: INTA# and IRQ#. INTA# generates system PCI interruption for PCI bus signals. IRQ# generates local logical interruption for local bus signals. Otherwise, Capacity changes of MAILBOX, DMA writing and reading transfer can also generate interrupt. Methods of generating interrupt shows below:

A. PCI device interrupt INTA#

(1)Fill interrupt pin INTA# in interrupt pin register of configuration space;
(2)Get interrupt vector from vector register of configuration space;
(3)Set condition of generating interrupt in PCI bus controller, such as MAILBOX getting full or DMA writing over;
(4)Read value of PCI bus controller register INTCSR and MBEF, judge interrupt source and do deal with interrupt in interrupt handling program;
(5)Clear interrupt and abort interrupt handling program. Clearing interrupt is done by means of filling "1" in INTCSR corresponding interrupt flag.

B. Local device interrupt IRQ#

(1) Set condition of generating interrupt in local bus controller register AINT, such as MAILBOX getting full or DMA writing over;
(2) Read value of local bus controller register AINT and

AMBEF. Judge interrupt source and do deal with interrupt in interrupt handling program;

(3) Clear interrupt and abort interrupt handling program. Clearing interrupt is done by means of filling “1” in AINT corresponding interrupt flag.

6. CONFIGURATION AND INITIATION of S5933

We must know and configure PCI device configuration space correctly for developing PCI device. Each PCI device has a configuration space other than I/O space and memory address space. Configuration space is a 256 bytes address space, which is in the form of record structure. The aim of configuration space is to provide a suit configuration method which can satisfy present and future foresaw system configuration architecture. These configurations have some functions: such as installation, configuration and guide without user interaction; Relocate device freely; Map system address with software not relative of device.

Configuration space has head area and relative device area. One device’s configuration space is accessed not only on system booting but also at anytime. Length of head area is 64 bytes, each device must support allocation register of the area. Each flag of the area identify device exclusively, and can control device with normal method. 192 bytes of relative device area is different in terms of device difference.

Head area of configuration space splits into two parts also. Definition of forwards 16 bytes is same in all kinds of devices, but others 48 bytes can be assigned differently by basic function of device supporting.

7. DRIVER

7.1 Initiation of Device

PCI device driver program should implement distinguishing PCI component, addressing PCI component resource and service for PCI interrupt. PCI system BIOS provides BIOS accessing and controlling features. All software(such as device driver program, extended ROM code) access special component by way of regular interrupt number 1AH calling BIOS function.

7.2 Port Operation

I/O port address space is different from memory address space on PC, so their process method is different. I/O space is a 64K bytes addressing space. I/O port address is not like memory having real mode and protected mode or sameness in all kinds of addressing mode. Because PCI bus is 32 bits standard bus, It always operates with DWORD when it writes or reads. There is no DWORD function in most C/C++ compilers, so we want to construct DWORD operation function inpd/outpd which can read or write.

7.3 Writing and Reading of Memory

Windows works at 32 bits protected mode. The entire difference between protected mode and real mode is CPU address mode, which is also Windows driver program design to be resolved. Windows adopt mechanism of subsection and pagination. Every virtual address is composed by 16 bits segment address and 32 bits segment offset. In subsection mechanism system generates linear address by virtual address, and generates physical address by linear address in pagination mechanism. Linear address splits into page directory, page table and page offset three parts. When a new Win32 thread is created, operating system will allocate a mass of memory, and build its page directory and page table. Page directory’s address is recorded in present information at the same time. When an address is calculated, system read address of page directory from CPU controller CR3 firstly. Secondly, get address of page table from page directory and get page frame of actual code/date page according to page table. Finally access appointed unit by page offset. Hardware device writes or read is physical memory, but application program writs or read is virtual address, so mapping physical address to application program linear address is very important.

The work of conversion from physical address to linear address is also completed by driver program. VMMCall__MapPhysToLinear in DDK can be used for address mapping in Window95. Memory mapping of driver program is mainly means of VxD system service MapPhysToLinear.

7.4 Configuration, Response and Invocation of Interrupt

The work of assigning interrupt, response and call should be processed in driver program. Call of interrupt is can be completed by Exec_Int in DDK, just like BIOS of 1AH interrupt reading configuration space does.

PCI device driver program should be get interrupt information from INTLN of PCI configuration register and interrupt pin register. Also DDK provides service for corresponding to interrupt event. For example, VPICD service is used for manage all hardware interrupt event in Window 95. PC hardware must distinguish hardware interrupt’s IRQ. VPICD provides default interrupt handling function or allowes other VxD overloads interrupt handling function for a given IRQ interrupt source. To process hardware interrupt VtoolsD should generate a class inheritting VHardwareInt class. VtoolsD provides function to code interrupt response program by the class.

7.5 Driver Invocation

Coding device driver is not ultimate aim. The ultimate aim is to need user application call the driver and implement some functions. Normal calling device driver should get a file handle to open device file by means of calling CreateFile function. For example, In our device dricer we use below language opening file:

```
hVxD=CreateFile("\\\\.\\PCIBIOS.VXD",0,0,0,)
CREATE-NEW,FILE-FLAG-DELETE-ON-CLOSE,0);
```

After program open device file, device driver program can exchange data freely by way of calling DeviceIoControl function.

After program completed hardware operation, program close device driver by means of calling CloseHandle(hVxD) function.

As for VxD, there are other calling method, such as DPMI mode. Method of DeviceIoControl can assure compatibility between Windows NTand Windows 9X. CreateFile statement is only one difference between two operation systems.

7.6 Encapsulation of Driver

We complete basic design of driver program. But consider using DeviceIoControl function is still complicated when calling device driver, program is not very universal. Furthermore, some development tools don’t include direct writing or reading I/O port language, such as Visual Basic, so we can consider encapsulate driver program according to different software. At present, we realize encapsulation with DLL, ActiveX, VCL and C++. DLL which can be called in mostly software environment. ActiveX can be used, such as Visual Basic etc, in visual program environment. VCL can be used in Delphi and C++ Builder. Consider many users use Visual C++, we provide C++ class libraries.

8. CONCLUSIONS

S5933 is applied in PCI interface circuits. It provides a flexible and easy periphery circuits environment to make design work easy. Its work frequency is 33MHZ, that is to say clock cycle is 30ns, so it establishes foundation for high speed transfer. In real test, its DMA transfer speed is so fast that transfer 16 bytes with only 500ns. Its high speed provides superiority condition for periphery circuits. The whole communication system includes data obtain independently, transfer and controller of data, real-time and controllability. The system can not only apply in normal data communication, but also apply in data transfer and process in radar system or communication message.

REFERENCES

[1] Li GuiShan, Qi DeHu. *PCI Bus Developer Guide*. XiAn Electronic Technology University Press.1997

[2] Wu AnHe, Zhou LiLi. *Windows Device Driver (VxD and WDM) Development and Implementation*. China Publishing House of Electronics Industry,2001.

[3] “Interfacing the TMS320C6201 to a PCI Bus Using the AMCC S5933 PCI Controller.” *Brian G. Carlson DNA Enterprises*，Inc, 1999.

[4] *AMCC S5933PCIcontroller Data Book*,1996

[5] Ma WeiGuo, He PeiKun,“Universal High Speed PCI Bus Target Module Design,” *Application of Electronics Technology*, 1999, 25(1).

[6] *Art Baker.Windows NT Device Driver Design Guide*, BeiJing,China Machine Press,1997.

[7] *AMCC S5933 PCI Controller Data Book*,Applied Micro Circuits Corporation, 1996.

The Study on Low-side Embedded Unit Designing In Distributed System*

Dong Chen
Wuhan Technical College of Communications
Wuhan,Hubei,China
Email:dongchen67@163.com

ABSTRACT

An embedded unit is a measuring-control cell integrating data acquisition and controlling into a whole and a terminal in distributed network system. Too many controlling points on the site as well as the duty of data acquisition and processing (DAP) will make it real time and controlling capability decrease rapidly. Therefore, the unit is designed into a host-slave structure including double single-chip-microcomputers to perform parallel communication. The host-processor is for data acquisition and controlling of the executing elements and the slave processor is for processing data. They synchronously operate and transmit data with the interrupt mode so that real time feature and controlling capability of the embedded unit can be guaranteed. In addition, the unit is connected to the main control room by means of the serial ports of the SCM to receive the control orders or enquiry of the host-computer, which can easily and economically establish the distributed network system integrating management and control. As a result, the development objective of "light, thin, short, multi-purpose and low cost" of the building-in system is realized to the utmost extent. Therefore it is specially suitable for the application site of the low-side embedded system.

Keywords: Embedded System, Serial Communication, Parallel Communication, Distributed Network

1. USING SERIAL INTERFACE TO PERFORM SYSTEM EXTENDING

In the site of the industry, because more points need to be monitored and the distances between them are greater and more industrial interfering exists in the site, the small-sizing SCM with strong anti-interfering capability and the host computer are used to establish a distributed data acquisition and control system. Each measuring and control unit separately operates to measure and process datum. In the main control room, host-computer monitors the operating conditions of measure orders to each slave computer as well as uniformly displays and stores the send out command to apiece salve computer, carry through data display and memory datum produced during production. The structure schematic diagram of its communication network is as follows:

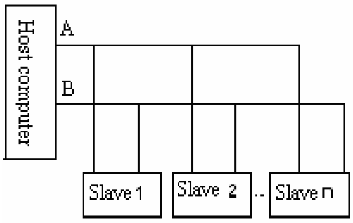


Fig.1 Schematic Diagram of Distributed Network Structure

1.1 Serial Interface of MCS-51 Single-chip Microprocessor

1.1.1 Description of Serial Interface of MCS-51 Single-Chip Microprocessor

In the pins of MCS-51, P3.0 (Pin 10) is for RXD and P3.1 (Pin 11) is for TXD. Serial interface of MCS-51 is a full-duplex one, i.e. it can synchronously transmit and receive signals. Because transmitting and receiving can't synchronously be performed, two serial interface registers are accessed through the special function registers SBUF and two buffers sue the same address (99H).

1.1.2 Operation Mode of Serial Interface of MCS-51

Four operation modes of serial interface of MCS-51 single-chip microprocessor are shown in **Table 1**:

Table 1. Four operation modes of serial interface of MCS-51

SM0 SM1	Operation Mode	Description of Function	Baud Rate
00	Mode 0	8-bit synchro-shift register	fosc/12
01	Mode 1	10UARS	Variable
10	Mode 2	11UARS	fosc/32 or fosc/64
11	Mode 3	11UARS	Variable

1.2 Principle of Multi-computer Communication

In multi-computer communication, in order to guarantee reliable communication between the host computer and the selected salve computer is realized, the communication interface must be guaranteed to have the identification function so that multi-computer communication can be realized by means of controlling SM2 bit in SCON of 8051.

1.2.1 Work Principle

Host-slave multi-computer communication can be realized with Mode 3 of 8051 serial interface and controlling matching of SM2 bit and RB8 in SCON. When the serial interface receives data with Mode 3, if SM2=1, data can be installed into SBUF only the ninth bit of datum received by the salve computer (in RB8) is 1. In addition, RI=1 must be set to apply for interruption to CPU. If the interrupt flag RI isn't set, information will be lost. If SM2=0, regardless of the ninth bit of the data is 1 or 0, RI is set to 1 and the received datum are installed in SBUF.

1.2.2 Communication Protocol

Host computer:

(1) The host computer calls the interruption setting program and sets initialization and parameters before transmission.

(2) The host computer calls each slave computer to transmit the control orders to them.

(3) The host computer transmits a frame of address information to the slave computers, including 8-bit address. If the programmable ninth bit is 1 (TB8=1), indicate what is transmitted is address, so all salve computers are interrupted.

(4) After receiving the response information of the salve

*Eleventh five-year educational science program of Hubei province(accession number:2006B237)

computer, the host computer compares it with the transmitted address information. If they are consistent, TB8 is reset 0 and the information will formally be transmitted. Otherwise error information will be transmitted.

(5) The host computer checks the status of the set flag bit to determine whether datum is completely transmitted.

(6) Communication between different computers is performed with the same frame format and baud rate.

Salve computer:

(1) SM2 of all slave computers = 1 and they are in monitoring status of only receiving the address frame.

(2) After receiving the address, the slave computers with determine the received address from the host computer is consistent with the local address.

(3) If it is consistent with the local address, SM2 bit is cleared 0 and the local address is transmitted to the host computer as the response signal.

(4) The salve computer enters the formal communication status and starts to receive data and command information from the host computer.

(5) Because of fault address, SM2 of the other slave computers are still 1 and return from interruption.

The host computer usually operates its own main program and transmits data to the slave computers by means of interruption. The slave computers usually acquire and process the data of the monitored objects. When serial interruption occurs, the serial interruption sub-program is activated and the corresponding operation is executed.

1.3 Realization

1.3.1 Bus Mode Serial Communication with RS-485 bus

Principle diagram of its data interface is as follows:

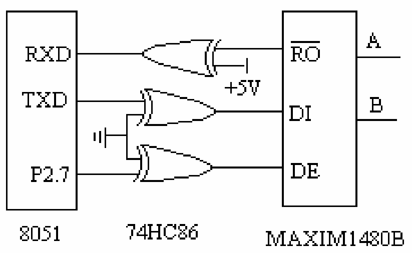


Fig.2. Principle Diagram of Data Interface

RS-485 serial bus interface transmits signal at the differential balance mode. The Drivers transform TTL level signal into the differential signal at the transmission end to output and the receivers revert the differential signal to TTL signal at the receiving end. Thus it has the strong common-mode interference resistance capability and high receiving sensitivity. When data transmission rate reaches 100kb/s, the communication distance can reach 1200M. It allows a transmitter drives several load devices at a twisted-pair is characterized by simple equipments, low cost and easy maintenance.

MAXIM1480B is a RS-485/RS-422 single power supply isolated data interface chip that enclosures a photoelectric coupler, a isolating DC-DC converter and RS-485/RS-422 driver into a whole, which can make the system realize isolation between physical physics interface chip and communication controller without additional power supply and decrease impact of the interference on communication reliability.

When DE=1, driver A and B are enabled (transmission status):
If DI=0, then A=0 and B=1.
If DI=1, then A=1 and B=0.
When DE=0, the driver is in condition of high-resistance (receiving status)
If electric potential of A is higher 200MV than that of B, RO=0.
If electric potential of $\overline{A}$ is lower 200MV than that of $\overline{B}$, RO=1.
74HC8 is a “OR and SAME” gate circuit, indicating the opposite-phase relation between MAX1480B and Pin 8051.

In the system, only a computer is used as the host computer, mutual communication between the host computer and salve computers can be performed with the software protocol. While the host computer operates main program, it needs to continuously perform cyclic query of the salve computers to monitor the status of the slave computers, receive the requests of the slave computers and transmit the order to the slave computers.

2. PARALLEL COMMUNICATION

When there are too many duty of DAP in a system, it is very difficult to guarantee real time of data acquisition and effectiveness of data processing. Therefore two single-chip computers must be used, data communication between which is performed through mutual connection. Their serial interface can not be used as the communication interface between two single chip computers in the system. According to the internal structure of MCS-51, they aren’t equipped with the additional hardware devices and are directly connected with the parallel ports. Therefore, different parallel connection methods are used according to different service requirements.

2.1 Realization of One-way Direction Parallel Communication

If only one SCM is needed to transmit data to another in application, the one-direction parallel communication mode can be adopted because this mode is simpler. What is shown in Fig. 3 is their configuration method.

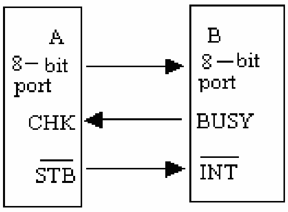


Fig.3. Configuration Method of One-direction Parallel Communication Interface

In Fig. 3, SCM A transmits data and SCM B receives data. 8-bit port can be anyone of P0～P3. Data transmitting flow is as follows:

(1) SCM A transmits data to the port.

(2) $\overline{STB}$ is set to “1” to interrupt SCM B.

(3) SCM B enters interruption mode and read data from the port.

(4) After reading data, BUSY is set to “1”.

(5) SCM A transmits next data after detecting response signal at port cable CHK.

This method can make full use of MCS-51 resource to extend the components of the complete system such as serial communication interface, parallel communication interface and timer so on.

2.2 Realization of Host-Slave Mode Parallel Communication Interface

The host-slave parallel communication interface features that two SCM can mutually transmit data through parallel communication interfaces. However, for this method, one SCM must be in host computer status and the other is in slave computer status. Fig. 4 is the principle theory diagram of the host-slave parallel communication interface. SCM A is the host computer and SCM B is the slave-computer. These interfaces use an 8-bit port (for example P0 or P1) and 4 control signal cables.

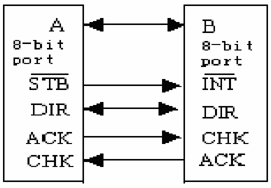


Fig.4. Principle Diagram of Host-slave Parallel Communication Interface

In host-slave operating mode, the operation mode of the interface includes two: the host computer transmits data and the slave computer receives them, or the host one receives data and the slave one transmits them.

2.2.1 The Host Computer Work Flow in The Mode of The Host One Transmitting and The Slave Receiving

Work flow of the host computer is as follows in the mode of the host one transmitting and the slave one receiving:

(1) The host computer is set to control bit DIR=0 to indicate the host computer will transmit data to the slave one;

(2) The host is set to $\overline{STB}$ =1, the slave one enters interruption and is ready for receiving data.

(3) After the host computer transmit data to 8-bit port, set ACK=1;

(4) CHK=1 in the host computer indicates the data transmission is completed.

2.2.2 Work Flow of the Host Computer in the Mode of the Host One Receiving and the Slave One Transmitting

Work flow of the host computer in the mode of the host one receiving and the slave one transmitting is as follows:

(1) When the host computer is set to DIR=1, indicates the host computer will read data from the slave one;

(2) The host computer is set to $\overline{STB}$ =1 to make the slave one enters interruption and be ready for transmitting data.

(3) CHK=1 in the master indicates the data are being read from 8-bit port.

(4) After data receiving of the host computer is completed, ACK is set to 1.

2.2.3 Work Flow of The Slave Computer

In work mode of the host-slave parallel communication interface, the slave receives or transmits data in interruption (inquiry) mode. Work flow of the slave is as follows:

(1) The slave computer enters ISR.

(2) If DIR=0 and CHK＝1, read the data. After reading the data, ACK is set to 1;

(3) If DIR=1, the slave computer transmits data to 8-bit port and ACK is set to 1;

(4) Exit ISR.

2.3 Data Communication Response Time of Two Parallel Port Modes

If work clock frequency of MCS-51 is 12MHZ, select P0 as 8-bit port. There is the only interruption source in every MCS-51 in system. For a single interruption source, interruption response time is 3～8μs.

In one-direction parallel communication mode, data is only transmitted from SCM A to SCM B and SCM B receives at the interrupting mode. Transmittal process of each byte includes three parts: data is transmitted to the port, the receiving computer is notified and response of the receiving computer is waited for, the program of which is generally written into:

```
MOV   P0,#DATA
CLR   STB
JB   CHK,$
```

If executing time of each instruction is respectively 1μs, 1μs and 3～8μs, transmitting time of the next byte of this mode will be about 10μs and the shortest time is 5μs. So data transmission efficiency is rather high.

In host-slave parallel communication interface mode, data transmission includes two directions: transmitting and receiving. Because data are transmitted under control of the host computer, and the time of transmitting is basically equal to receiving time, data transmitting program of the host computer is as follows according to the data transmission:

```
  CLR   DIR
  CLR   STB
  MOV   P0, #DATA
  SETB   ACK
  JB   CHK, $
```

Because executing time of the first four instructions is all 1μs, the executing time of the last instruction is the time of waiting for responder of the slave computer, which is 3～8μs. In this mode, transmission time of next byte is about 12μs and the shortest transmission time can reach 7μs. The transmission efficiency is still higher.

3. CONCLUSIONS

An embedded unit is a monitoring unit integrating data acquisition and control and a terminal of the distributed network system. If the control points are more in the site with heavier duty of data acquisition and processing (DAP), real time and control capability of this embedded unit will greatly lower. Therefore, this unit is designed into the host-slave structure including two single-chip microcomputers parallel communication. The host computer is for data acquisition and controlling the executing elements and the salve computer is for data processing, shown as Figure 5:

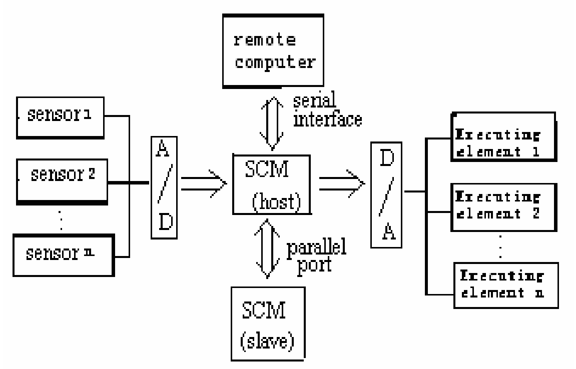


Fig.5. Structure Diagram of Embedded Unit

When two SCM synchronously operate and data is transmitted at the interruption mode, real time and control capability of the embedded unit is guaranteed. In addition, if they are connected to the main control room with the serial interface of the SCM and control order or enquiry of the host compute is received, the distributed system integrating management and control into a whole can easily and economically be established. If an embedded unit is combined with internet at the same mode and information is directly transmitted to Internet with the existing TCP/IP without needing the special line to establish the distributed network system with remote accessing capability, which realizes the development objective of “light, thin, short, multi-purpose” to the utmost extent. Therefore it is especially suitable for lower-ended embedded application site.

REFERENCES

[1] Ian Foster,“Ian foster on recent changes in the grid community”, *IEEE Distributed System Online*,Vol.5,PP2-3,Feb.2004

[2] Ian Foster,“The globus toolkit for grid computing”. *1st International symposium on cluster computing and the grid*,Feb.2001

Research on Failure Diagnosis for Military Electronic Equipment Based on Fuzzy Theory

Linchun Xing, Renbin Zhou, Qingbin Cui
Self-propelled Gun Department of WuHan Ordnance Sergearnt Technology School
WuHan, HuBei province, 430075 PRC
Email:xlc86619759@163.com

ABSTRACT

In conformity to the maintain system of military electronic equipments, the main repair means to military electronic equipment is to replace the failure module by a new one. This paper discusses an information fusion method of failure diagnosis for military electronic equipment, a fuzzy arithmetic model is presented for a analysis of the fuzzy relationship between the failure characteristics and the failure mode. The arithmetic model can effectively eliminate the uncertainty of the failure diagnosis for electronic equipment by combine the weight of the objective statistical data of the failure sample and the subjective evaluation of the experts; also, it can enhance the scientific degree and reliability of the failure diagnosis. The use of the method is discussed in the paper. At last, an instance is given to certify that the method is feasible.

Keywords: Fuzzy Theory, Information Fusion, Electronic Equipment, Failure Diagnosis, Expert Estimation.

1. FUZZY INFORMATION FUSION THEORY

On the assumption that *A* is the possible decision-making policy set of failure diagnosis system, such as failure electronic modules set of the diagnosis objects; *B* is the failure characteristics information set. The element μ_{ij} in the connection matrix $R_{A\times B}$ expresses the probability for failure characteristics information *j* to deduce failure characteristics information *i*, X denotes the reliability of the failure characteristics information, Y gained from the fuzzy commutation figures the probability of decision-making policy after fusion. To tell it in detail, In case of there are *m* failure characteristics information which will be tested, and the sum of decision-making policy may be *n* ,then, A={y_1(policy 1),y_2(policy 2),…y_n(policy n)}, B={x_1(failure characteristic 1),x_2(failure characteristic 2),…x_m(failure characteristic m)}.

The estimation from failure characteristics information to each probability decision-making policy is showed by the subjection degree of *A*, suppose the estimate results from failure characteristic *j* to the diagnosed system are: [μ_{j1}/policy 1,μ_{j2}/policy 2,…μ_{jn}/policy n] in it, $0\le\mu_{ji}\le 1$.

The probability of policy i will be regard as μ_{ji}, marked as vector (μ_{j1},μ_{j2},…μ_{jn}), the A×B matrix is composed of *m* failure characteristics, it is denoted below:

$$R_{A\times B}=\begin{bmatrix}\mu_{11},\mu_{12},\ldots,\mu_{1n}\\ \mu_{21},\mu_{22},\ldots,\mu_{2n}\\ \ldots\ldots\\ \mu_{m1},\mu_{m2},\ldots,\mu_{mn}\end{bmatrix} \tag{1}$$

To show the reliability of each failure characteristic information by the subjection degree of *B*. X={x_1(failure characteristic 1),x_2(failure characteristic 2),…x_m(failure characteristic m)}, then making fuzzy transform by the formula:

$$Y=X\cdot R_{A\times B} \tag{2}$$

and the diagnosis result Y=(y_1,y_2,…y_n) will be gathered after fusion, in other words, it also can be regard ad probability set of failure decision-making policy after fusion.

2. THE ESTABLISHMENT OF THE SUBJECTION DGREE MATRIX

In the process of failure diagnosis based on Fuzzy information fusion theory, the most important thing is to establish the subjection degree, in this paper the subjection degree is determined by synthesizing the weight of the objective statistical data of the failure sample and the subjective evaluation of the experts. Tell it in detail, when some failure characteristic was observed, fist ,some experts which skilled in the failure diagnosis field for electronic equipment was chosen to give the probability evaluation of possibility of each failure mode, then mark the average of the estimation result from the experts as R_1, (R_1 is described as the fuzzy relation matrix obtained by experts evaluation between the failure characteristics information and the failure decision-making policy); then 10 failure samples which have the same failure characteristic was collected to make a statistic of the sum of each failure mode, the sum divided by 10 should be marked as R_2, (R_2 is described as the fuzzy relation matrix obtained by the sum statistic of failure samples between the failure characteristics information and the failure decision-making policy), according to the experience, the right distribute for experts discursion result is 0.4, and the right distributed for the statistic result is 0.6. With the failure samples increasing, the right distributed for experts discursion should be decreased, and the right distributed for the statistic result should be increased. When the sum of the failure samples will be more than 50, the probability of the statistic result already can replace the subjection of the failure, then the right distribute for the statistic result will be 1, experts discursion will not be care about. In conclusion, the final subjection degree Y can be calculated by the formula:

$$R=\lambda_1R_1+\lambda_2R_2 \tag{3}$$

The weight distributed of the statistic of failure samples and the experts evaluation is illustrated as Figure 1 and Figure 2, in the figure, *n* is described as the sum of the failure samples. In this paper, the symmetrical trigonometric function was took as the subjection degree function of the fuzzy set to partition the subjection degree of the electronic equipment modules. First, according to the observed Failure characteristics the fuzzy set of the failure possibility of each electronic equipment was compartmentalized, and the fuzzy set is partitioned as: H={f_1,f_2,f_3,f_4,f_5,f_6,f_7} ={VVB,VB,B,G,S,VS,VVS}, in it , fuzzy set partitioned by 7 subset as { f_1 , the failure possibility is very very big, marked as VVB }, { f_2 , the failure possibility is very big, marked as VB }, { f_3 , the failure possibility is big, marked as B }, { f_4 , the failure possibility is general, marked as G }, { f_5 , the failure possibility is small, marked as S }, { f_6 , the failure possibility is very small, marked as VS }, and{ f_7 , the failure possibility is very very small, marked as VVS }, and it is illustrated as Figure 3.

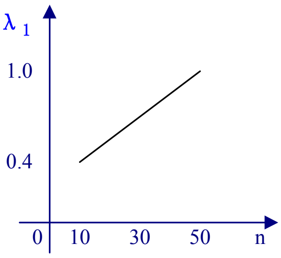


Fig.1. The weight distribution of the statistic of failure samples

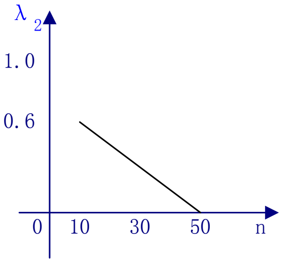


Fig.2. The weight distribution of the experts evaluation

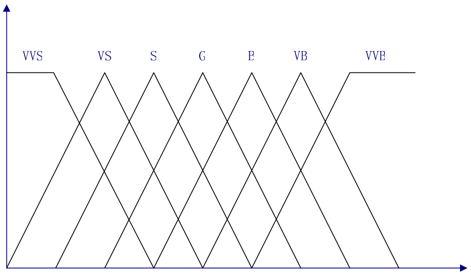


Fig.3. The distribution of the fuzzy set

The subjection degree X=(x_1, x_2, …, x_m)of matrix B was used to described the reliability of each failure characteristics, the fusion diagnosis result Y=(y_1,y_2,…y_n) is gained by the fuzzy transform formula Y=X·R, Y described as the possibility set of the failure decision-making policy.

3. THE DETERMINANT RULES FOR FAILURE MODULES

We shall do failure decision-making by the determinant rules, On the assumption that y_c denoted the failure element, then the basic rules are described as:

① The determinant failure element should have the biggest subjection degree value, $y_c = \max_i (y_i)$ ·

② The subjection degree value of the determinant element should be bigger than some threshold, and the threshold should be chosen in conformity the fact. $y_c > \alpha$, $\alpha \in R$且$\alpha \rangle 0$

③ The dispersion of the subjection degree value between failure element and other normal elements should be bigger than some threshold, and the threshold should be chosen in conformity the fact.
$y_c - y_i > \beta$, $\beta \in R$且 $\beta \rangle 0$

4. THE APPLICATION ON FAILURE DIAGNOSIS FOR MILITARY ELECTRONIC EQUIPMENT BASED ON FUZZY THEORY

A certain kind of military embedded computer is composed of 7 kinds of electronic module as mainboard, keyboard , power supply board, data adapter , LCD screen , power transformer board and GPS board etc. If we don't care about the software failure and wires failure , when the computer failure occurs , there will be only 7 kinds of failure modes as mainboard failure , keyboard failure , power supply board failure , data adapter failure , LCD screen failure , power transformer board failure and GPS board failure etc. In some failure diagnosis course , three failure characteristics as black screen , keyboard Lamp is extinct , the GPS cann't make location had been found from the report of the computer, μ_{i1}, μ_{i2}, μ_{i3}, μ_{i4}, μ_{i5}, μ_{i6}, μ_{i7} was used to describe the failure subjection degree of each electronic module for only the failure characteristic i was occurred. Then the failure characteristics and the fuzzy fusion arithmetic were used to found the failure electronic module, and the failure samples statistic information and the experts estimation information was listed in table 1 and table 2.

Table 1. The experts estimation information

Characteristic number	the subjection degree of the experts （μ_{i1},μ_{i2},μ_{i3},μ_{i4},μ_{i5},μ_{i6},μ_{i7}）
1	（0.05,0,0.11,0,0.24,0.6,0）
2	（0,0.2,0.18,0,0,0.62,0）
3	（0.15,0,0.1,0,0,0.35,0.4）

Frome table 1, table 2 and the formula R=λ_1R$_1$+λ_2R$_2$ mentioned above, it can be conclude:

$$R = \lambda_1 R_1 + \lambda_2 R_2 = \begin{bmatrix} 0.094 & 0 & 0.125 & 0 & 0.156 & 0.625 & 0 \\ 0 & 0.311 & 0.164 & 0 & 0 & 0.525 & 0 \\ 0.155 & 0 & 0.065 & 0 & 0 & 0.364 & 0.416 \end{bmatrix}$$

In this instance，we marked X=[1,1,0.8]，then
Y=X·R=[0.288,0.311,0.341,0,0.156,1.441,0.333]
Further, the unitary result is:

$$\bar{Y} = [0.1 \quad 0.108 \quad 0.119 \quad 0 \quad 0.054 \quad 0.503 \quad 0.116]$$

Table 2. The failure samples statistic information of the electronic module

Characteristic number	Failure Characteristics	Sample number	Subjection degree of the failure samples number （μ_{i1},μ_{i2},μ_{i3},μ_{i4},μ_{i5},μ_{i6},μ_{i7}）	λ_1	λ_2
1	black screen	64	（0.094,0,0.125,0,0.156,0.625,0）	0	1
2	keyboard Lamp is extinct	37	（0,0.324,0.162,0,0,0.514,0）	0.104	0.896
3	GPS cann't make location	19	（0.158,0,0.053,0,0,0.368,0.421）	0.248	0.752

at last, we marked α=0.5，β=0.2, then it can be concluded that the failure electronic module is the power supply board. Which is consistent with the examination result, so the failure recognition rate is 100%.

5. CONCLUSIONS

In the electronic modules failure diagnosis course, if the failure mode was concluded only by either of the experts estimation or the failure samples statistic according the failure characteristics, the result will be uncertain. As introduced in the fuzzy information fusion method, the infection weight of the failure characteristics is brought into the subjection function of set theory, the fusion subjection function and the fuzzy respiratory matrix is used to solve the uncertain relationship between the failure mode and the failure characteristics, so as to realize the failure diagnosis. The uncertain factor are cared in this method, and the method also has some advantages as small calculation and easy to carry out etc. It can be testified by instance, the fuzzy fusion method mentioned in the paper is efficiency, and it also provide a way to solve the similar problem.

REFERENCES

[1] Zhu Daqi, *The principle and practice of the failure diagnosis of electronic equipment*, Electronic Industry Press,pp 121-180,2004

[2] Li Shengyi, Wu Xuezhong, Fan Dapeng. “ The application of multi-sensor fusion theory and intelligent manufacturing system,” the University of Science and Technology of national defence press,pp 198-219,1998

[3] Liu Tongming,Xia Zhuxun, Xie Hongcheng. “Data Fusion Technology and Its Application,” National defence Industry Press,pp 56-61,2002

[4] Quan Taifan, *Information Fusion Theory and Application Based on NN-FR Technology*, National defence Industry Pres,pp 87-91,2002

Linchun Xing is an instructor and a teacher of Self-propelled Gun Department of WuHan Ordnance Sergearnt Technology School. He graduated from BeiJing Academy of Armored Force Engineering in 1998; He has published two books, over 15 Journal papers, 5 of them have been indexed by ISTP. He was born in Huhehaote city in Inner Mongolia province in 1975, specializes in information fusion,failure diagnosis of mechanism and electronic equipment.

32 bit Floating-Point Addition and Subtraction ALU Design

Kui Yi[1], Pin Xiong [2], Yuehua Ding[1],
[1]Department of Computer Science and Information Engineer, WuHan Polytechnic University,
Wuhan, HuBei Province 430023, China
[2]Information School, Zhongnan University of Economics and Law
Wuhan, HuBei Province 43074, China
Email: [1]ykll1903@126.com

ABSTRACT

Floating-Point arithmetic unit is always key factor of restricting microprocessor performance. This paper brings forward a scenario about Floating-Point addition and subtraction ALU which supports IEEE-754 standard. The scenario adopts 4-Level pipelining structure: 0 operation numbers check, match exponent, fraction arithmetic, result normalization and rounding. Each step can be act as a single module. Among these modules, there are some registers which can prepare correct data for next operation.

Keywords: Pipelining, IEEE-754 Standard, VHDL

1. INTRODUCTION

With the development of SOC technology, IP technology and integrate circuit technology, RISC soft-kernel processor research and development design is thought much of by people now. RISC soft-kernel processor based on FPGA is applied on many kinds of trade widely, especially in embedded type system.

Data type is data expression in the computer realization. There are basic type, flag type, stack type, vector type etc in common data type. Basic type is more widely applied in advanced programming language. Floating-Point takes up important status in the basic data type. Therefore, Floating-Point arithmetic ability becomes one guideline of data arithmetic. Floating-Point provides big-range and high-precision data, so lots of microprocessor adopt hardware to act as Floating-Point arithmetic unit directly. IEEE-754 Floating-Point standard is just the most popular Floating-Point number standard, which defines format, precision and arithmetic operation of Floating-Point accorded with IEEE-754.

Through the object’s characteristic and its feasibility research, QuartusII produced by Altera Corp. is chosen as environment in which we design and simulate. Top-down design method is adopted. The method of designing and testing enhances at the same time method enhance design reliability greatly.

2. FLOATING-POINT ADDITION AND SUBTRACTION WORKING THEORY AND STRUCTURE

Floating-Point addition ALU has biasing alignment, biasing normalization, exponent arithmetic , LZA(Leading Zero Anticipation) and rounding process besides fraction addition and subtraction arithmetic, so its arithmetic speed have a big gap compared to integer adder. High-speed Floating-Point addition ALU is very important for Floating-Point unit performance. Therefore, technology of reducing Floating-Point addition arithmetic time is regarded as very important research. In this research, pipelining technology realize four steps operations: 0 operation number checkexponent match, fraction arithmetic, result normalization and rounding; Floating-Point number choose IEEE754 standard; Fraction is expressed by implement and exponent is expressed by biasing. This addition and subtraction ALU can enhance the CPU arithmetic speed and the whole system performance.

2.1 Key Question to be Resolved in the Research Design

(1) Check the 0 operation number at first. If addend or summand is 0, result is summand or addend.

(2) Because of adopting IEEE754 standard Floating-Point number in the design and exponent expressed by biasing in computer, match exponent before arithmetic. The method to make exponent of two numbers equal is aligning the point of two numbers. Adopts smaller-fraction right-shift method. Once right-shift 1 bit exponent adds 1 until exponent of two numbers equals. Fraction right-shift can result in the digit bit missing and affecting precision.

(3) On fraction addition and subtraction operation, because fraction structure is 1.m actually, the lost “1” should be filled in the actual arithmetic. The 1.m format should be kept for the result after arithmetic, so fraction must be left-shifted or right-shifted normalization when is not “1.m” form in the final result, and final result should be rounded with different rounding process.

2.2 Structure Design

Through analysis of two Floating-Point numbers addition and subtraction arithmetic process, the whole arithmetic process is divided into 4 steps in the whole arithmetic process: 0 operation numbers check,, match exponent--make smaller exponent equal bigger exponent, fraction addition and subtraction, result normalization and rounding. Between each module there are registers designed to store previous pipelining segment result. The four modules can act as four different sub-missions, as long as input two numbers to addition and subtraction arithmetic into the pipelining uninterruptedly. So the missions can implement each operation parallel and time parallelism is realized too. Abstract structure graph designed shows as Fig. 1:

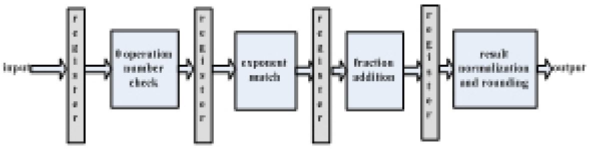


Fig.1. Arithmetic Design Structure

2.3 Floating-Point Number Structure and Design

2.3.1 Floating-Point Structure

Because addition and subtraction ALU is for 32 bit Floating-Point number and Floating- Point number adopts IEEE745 standard, the Floating-Point number structure that the design adopts shows as Fig.2:

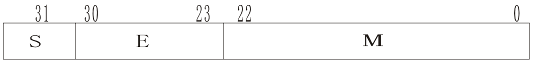


Fig.2. Floating-Point Number Structure

(Caption: In the Floating-Point number structure, M is value of fraction, which is expressed by sign magnitude adopting “hidden bit” expression; E is value of exponent, which is expressed by biasing; S is fraction sign.)

2.3.2 Guard Digit and Rounding Method

Guard Digit: for addition or subtraction arithmetic, if two normalized number with same sign add or two normalized number with opposite sign subtract, the sum of it and fraction of bigger-exponent Floating-Point number need not be left-normalized definitely on matching exponent. No matter how many bits the fraction of smaller-exponent Floating-Point right-shift. Supposing mA is the fraction of bigger-exponent Floating-Point and mB is the fraction of smaller-exponent Floating-Point after exponent match. On

$$\frac{1}{r_m} \leq \left| m_A + \mathrm{m}_B \right| < 2$$

fraction adopting sign magnitude, fractional part must be：

Sum of the two fraction is :

$$\frac{1}{r_m} \leq \left| m_A \right| < 1, \quad 0 \leq \left| \mathrm{m}_B \right| < 1$$

Sum of mA+mB must not be left-normalized, but perhaps 1 bit should be right-normalized.

If two normalized number with same sign subtract or two normalized number with opposite sign add, there are three conditions coming with different exponent remainder of two Floating-Point numbers:

(1) When exponent remainder of two Floating-Point normalized numbers equals 0, sum of the two Floating-Point numbers’ fraction is left-normalized p-1 bits at the most (p is word length of fraction value). But there are no any digit bit shifted from lowest bit of fraction on exponent matching, that is guard digit is need but has no resource, so guard digit should not be configured.

(2) When exponent remainder of two Floating-Point normalized numbers equals 1, sum of the two Floating-Point numbers’ fraction is left-normalized p bits at the most. But there is one digit bit shifted from lowest bit of fraction of smaller-exponent Floating-Point number on exponent matching, so only 1 guard digit should be configured.

(3) When exponent remainder of two Floating-Point normalized numbers is equal or greater than 2, lots of bits shift from lowest bit of fraction of smaller-exponent Floating-Point number on exponent matching, but sum of two Floating-Point numbers left-normalize 1 bit at the most, so only 1 guard digit should be configured for the left-normalization.

In one word, on any condition two Floating-Point normalized numbers addition, the bit for left-normalization Floating-Point is only 1 bit.

Rounding Method: On exponent matching or right-normalization time, fraction is left-shifted, so the lowest bit of fraction right-shifted is lost and error is made. Therefore, rounding should be done. Common rounding method are truncation, round toward + infinity , round to nearest, R* rounding, ROM rounding. Round toward + infinity is widely applied in all kinds of computer system at present. This system adopts round toward + infinity. Implement difficulty of round toward + infinity only inferiors to truncation.

Rounding rule of round toward + infinity is: set up the lowest bit of fraction valid word length p bit normalized as r/2 (r is basic value of fraction), no matter what g digit bits are. The g digit bits are super valid word length. When basic value of fraction is 2, configure the lowest bit of fraction’s valid bit as 1. When basic value of fraction is 16, configure the lowest bit of fraction’s valid bit as 8.

If fraction normalized mx has p+g bits before rounding:

mx＝±0.xxx...x0|000...00～±0.xxx...x0|111...11,

Value of fraction in the range of valid word length is left to Signal “｜”, having p bits. Value of fraction beyond the range of valid word length is right to Signal “｜”, having g bits. Following this, fraction normalized m has p bits after rounding and its value is：

m＝±(0.xxx...x0＋2-p)＝±0.xxx...x1，

The g bits beyond the range of valid word length is all rounding and value in the range of valid word length adds 2-p.

Before rounding if fraction normalized mx is:

mx＝±0.xxx...x1|000...0～±0.xxx...x1|111...1,

Then, after rounding fraction normalized m is：m＝±0.xxx...x1,

The g bit beyond the range of valid word length is all rounding, which is same as truncation.

The range of round toward + infinity ’s error in the positive area is :

δ＝m－mx＝－2-p(1－2-g)～＋2-p，

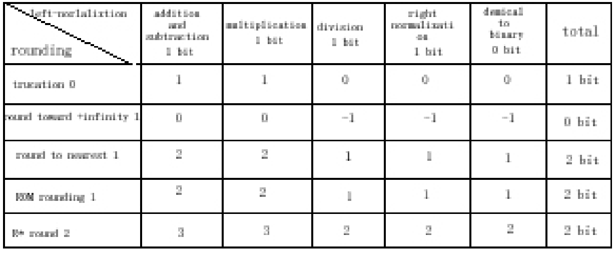
The range of round toward + infinity ’s error in the negative area is :

＝m－mx＝－2-p～＋2-p(1－2-g)，

Main shortcoming of round toward + infinity is digit expression precision is low, because the lowest bit of fraction is configured r/2 stably, so 1 bit precision lost. The main advantage of round toward + infinity is that implement is easy in some way and the error in positive area or negative area is a small, so it can achieve balance.

The first line in Table I shows that guard digit needed on any condition in left-normalization. The first column in Table I shows guard digit needed on any rounding method. The number on cross-point shows the total guard digit need on any condition (including rounding and left-normalization). The last column in Table I shows guard digit which is configured on adopting some rounding method. In addition, it should be pointed out that the lowest bit for rounding should be configured one binary bit when basic value of Floating-Point number fraction is greater than 2. [4]

Table I Guard Digit Needed on any Condition



left-norlalixtion rounding	addition and subtraction 1 bit	multiplication 1 bit	division 1 bit	right normalizati-on 1 bit	demical to binary 0 bit	total
trucation 0	1	1	0	0	0	1 bit
round toward +infinity 1	0	0	−1	−1	−1	0 bit
round to nearest 1	2	2	1	1	1	2 bit
ROM rounding 1	2	2	1	1	1	2 bit
R* round 2	3	3	2	2	2	2 bit

2.3.3 Floating-Point Addition and Subtraction Arithmetic Process

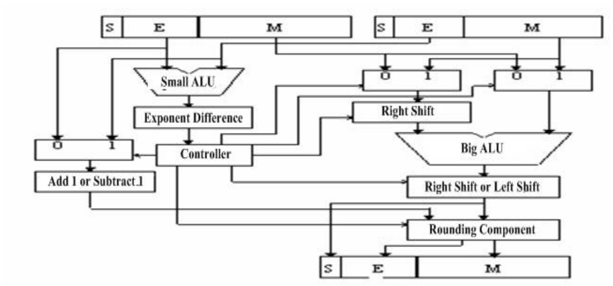


Fig.3. Floating-Point Addition and Subtraction Arithmetic Flow Chart

3. MODULE DESIGN

3.1 Operation Number Check Module

Floating-Point addition and subtraction arithmetic is more complicated than fixed-point arithmetic. If one of the two operation number is checked equal 0, the arithmetic result is obviously and the next operation does not need any more. 0 operation number check module just accomplishes function of checking two input number are 0 or not. On doing addition arithmetic, if addend/ summand is 0, the result must be summand /addend.

3.2 Exponent Match Module

This module mainly charges of the exponent align operation of two numbers. When two Floating-Point numbers add or subtract, check the exponent of two numbers same or not. Only if point is aligned the fraction addition and subtraction arithmetic can be done. On the contrary, if the exponent of two numbers is different, expressing the point not aligned, make the exponent of two numbers same. Make smaller-exponent accorded with bigger-exponent. Supposing there are two Floating-Point numbers x and y.E1 is exponent of x and E2 is exponent of y. They are both expressed by biasing, that is $X=M1*2^{E1}$， $Y=M2*2^{E2}$ (M1 and M2 is fraction). At first, obtaining exponent difference E'=|E1-E2| and kept bigger-exponent E'=max(E1,E2). If E'=0, right-shift fraction of smaller-exponent number and add E' to its exponent. Because exponent is expressed by biasing, data bit of exponent and implement is same entirely but their sign is on the contrary. Converting exponent of y to its complement form can realize exponent addition and subtraction arithmetic. Considering digit expression efficiency, standard fraction of every Floating-Point number is 1.m default form in IEEE 754, but input number dose not show the "1". Ensure correctness of next fraction addition and subtraction arithmetic, so output data in this module must show the "1". The "1" is filled in front part of output number's fraction and 1.m form is showed. In this way, fraction takes part in arithmetic with 24 bits on next addition and subtraction arithmetic.

3.3 Fraction Addition and Subtraction Arithmetic Module

This module charges of addition arithmetic of two input numbers. Because fraction is adopted hidden bit expression method, fraction must be added omitted "1" before arithmetic. Its method is the same as Fixed-Point addition and subtraction arithmetic completely.

3.4 Result Normalization Module

Because fraction word length of any Floating-Point expression is limited, the rounding of Floating-Point is considered. There are two questions encountered on using Floating-Point numbers: one is general decimal digit Floating-Point converting to Floating-Point in computer, which valid length surpass the given fraction word length. The surpassed part must be rounding. The other is addition, subtraction, multiplication, division result of two Floating-Point numbers normalization, whose fraction word length surpass given fraction word length. For example, the fraction word length of product is double the given Floating-Point word length when two Floating-Point numbers multiply; In the same way, when two Floating-Point numbers divides indivisibility can result in quotient need more bit and fraction word length of quotient surpass given word length probably. The 01.ø…ø form can be appeared in the result of fraction summation on Floating-Point addition and subtraction arithmetic. Two sign bit are not same, which is called overflow in the fixed-point addition and subtraction arithmetic and not permitted. But in Floating-Point arithmetic, it shows absolute magnitude of fraction summation result is greater than 1 and destroys left-normalization. Right-shift fraction result is to realize normalization expression, which is called right-normalization. Fraction right-shift 1 bit and exponent adds 1. Left-normalization must be done when fraction is not the 1.M form.

4. CONCLUSIONS

This paper mainly discusses one function unit in CPU design. The Floating-Point number addition and subtraction ALU is simulated successfully on the QuartusII 4.1. Test proves that this ALU is efficient. The most excellent scenario could be chosen in terms of actual condition, as long as only do a little change in the scenario.

REFERENCES

[1] Wang-YanFang,"RISC Technology Development and Research,"*Sci/Tech Information Development & Economy*,May 2005.

[2] Zhao-Xin,*VHDL and Digital Circuit Design*,Engine Industry Press,Jun 2005.

[3] Mo-JianKun,Gao-JianSheng,*Computer Organization*, Huazhong University of Science and Technology Press,1996.

[4] Zheng-WeiMin, *Computer System Structure*，Tsinghua University Press,Oct 2004.

[5] Bai-ZhongYing,*Computer Organization*, Science Press, Nov 2000.

[6] Chen-Lei,Gao-DeYuan,Pan-XiaoYa etc,"A embedded rise Micro-Process Integer Component Design[J]," *Micro-Electronics and Computer*,2004.

[7] Pan-Song,Huang-JiYe,*EDA Technology Utility Tutorial* [M]．BeiJing：Science Press, 2002.

[8] Zhang-XiuJuan,Chen-XinHua,*EDA Design and emulation Practice[M]*,BeiJing：Engine Industry Press, 2003, WeiPu Information.

[9] Jan M.Rabaey,Digital Integrated Circuits- A Design Perspective Prentice—H al l I nternational,I nc. Tsinghua University Press 1999.2;

[10] Behzad Razavi, *RF Microelectronics Prentice Hall PTR.* Jul 1997.

[11] John L. Hennessy David A. "Patterson Computer Organization&Design" — — *The Hardware/Software Interface*,Engine Industry Press English version Second Edit ion Sep 1999.

[12] Aldec Active-HDL the Design Verification Company Online Help.

The Specification of the Embedded System of Real-time IR

Yong Zhu [1,2]
[1]**College of Computer Science, Wuhan University of Science and Engineering**
Wuhan, Hubei 430073, China
[2]**School of Computer Science & Technology, Huazhong University of Science & Technology**
Wuhan, Hubei 430074, China
Email: zhudz_1964@163.com, zy@wuse.edu.cn

ABSTRACT

The specification of real-time embedded system is about design methodologies, which involves capture-and-simulate, describe-and-synthesize and specify-explore-renement. The architecture, schedule and implementation of the real-time IR video processing are described with HDL (Hardware Description Language), and the validity of the system has been proved by logic simulation. It has advantage for real-time applications and overhead-saving.

Keywords: Specification, Embedded System, Real-time, IR (Infrared)

1. INTRODUCTION

A conceptual view depends on application is very important to system specification. It is composed of computation conceptualized as a program and controller conceptualized as a state-machine. The design methodology of embedded system focused on lower levels, but need tools for system level in the future, because paradigm shift to higher levels can increase productivity. It deals with functionality specification, system design, component implementation, allocation, partitioning and renement.

The current all-purpose computer system aims at the goals of maximal throughput and minimal response time, but it is usually complicated, expensive and not optimized to the pertinence of applications [1],[2]. On the other hand, the embedded system focuses on the application to satisfy the functions and performances of the system by amending software and hardware. In fact, the pertinence is obvious in embedded applications. For example, the data of I/O port is sampled and sent out under the synchronous signals in IR applications. The design of all-purpose computer system is the same for above applications.

The IR video processing [3],[4] instruments are mostly integrative for convenience now. The most key problem, also a hot potato, is video processing in real time. So the DSP has been introduced because of its capability of high speed data processing. But the DSP has some disadvantages: There are not such interfaces as Ethernet, USB, IDE, TFT and so on built-in its architecture. The compatibility of I/O driver does not fit the software of PC sometimes, and there is not OS to manage the system resources. So it only suit simple computational task, i.e. single task system. It is not easy to maintain such system when the system software is updated because of the tightness between DSP and application program.

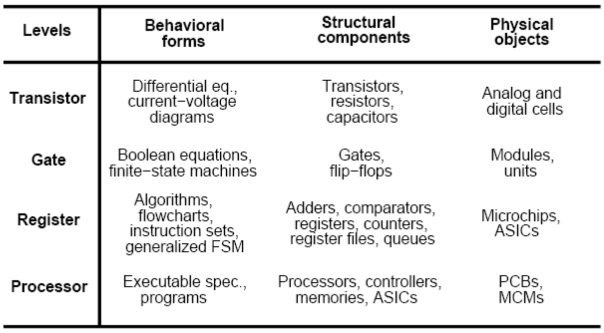
The other recommendatory architecture is embedded computer system which is introduced in the paper. The most superiority is standards compatibility and abundant resource. The mainstream CPU could be selected from ARM family, MIPS family, SC12XX x86 architecture of NS, Power PC and Dragonball of Motorola. OS could be free open source Linux and high performance commercial Linux from MontVista or VxWorks of Windriver. C language is used to development language because of its cross-platform to make the system port and update easy. We can get the best design, and the continuing of product is available.

There is an important topic which mentioned above is about real time. Although there is not special data processing unit like DSP in the embedded CPU, its computation capability is the same as advanced DSP along with the progress of IC technology and computer architecture. The embedded CPU has already been special hardware operation units to operate mathematics even DSP instructions. So the embedded technology with SoC will be competent for the real time images processing. The example put forward in the paper is the best evidence [5].

2. SPECIFICATION OF MODELS AND ARCHITECTURES OF EMBEDDED SYSTEMS

There are three design representations, which are behavioral, structural and physical that represent functionality, functionality and connectivity respectively. The levels of abstraction are illustrated as Tabel. 1[6].

Tabel 1. The levels of abstraction



Levels	Behavioral forms	Structural components	Physical objects
Transistor	Differential eq., current−voltage diagrams	Transistors, resistors, capacitors	Analog and digital cells
Gate	Boolean equations, finite−state machines	Gates, flip−flops	Modules, units
Register	Algorithms, flowcharts, instruction sets, generalized FSM	Adders, comparators, registers, counters, register files, queues	Microchips, ASICs
Processor	Executable spec., programs	Processors, controllers, memories, ASICs	PCBs, MCMs

Models are conceptual views of the system’s functionality, i.e. a set of functional objects and rules for composing these objects, and Architectures are abstract views of the system’s implementation, i.e. a set of implementation components and their connections.

The models could be described as:

- State-oriented models: Finite-state machine (FSM), Petri net, Hierarchical concurrent FSM
- Activity-oriented models: Dataow graph, Flowchart
- Structure-oriented models: Block diagram, RT netlist, Gate netlist
- Data-oriented models: Entity-relationship diagram, Jackson’s diagram
- Heterogeneous models: Control/dataow graph, Structure chart, Programming language paradigm,Object-oriented paradigm, Program-state machine, Queueing model

Also the architectures can be classified as:

- Application-specic architectures: Controller architecture, Datapath architecture, Finite-state machine with datapath (FSMD).
- General-purpose processors: Complex instruction set computer (CISC), Reduced instruction set computer (RISC), Vector machine, Very long instruction word computer (VLIW)
- Parallel processors

3. SCHEDULE OF HARDWARE TASK

The hardware logic tasks have been implemented by the hardware logic, so the overhead of operating system will be low. Furthermore they could be scheduled by the hardware logic. First some real-time schedule policies are summarized as follows[7].

Cyclic Executive. A cyclic executive consists of continuously repeated task sequences, known as major frames. Each major frame consists of a number of small slices of time, known as minor frames, and tasks are scheduled into specific minor frames.

Event-Driven System. An event-driven design uses real-time I/O completion or timer events to trigger schedulable tasks. Many real-time Linux systems follow this model.

Pipelined Systems. They use inter-task messages (preferably prioritized) in addition to I/O completion and timers to trigger tasks. The control flow for an event proceeds throughout the system from source to destinations.

Client-Server Systems. They also use inter-task messages in addition to an I/O completion and timers to trigger tasks. Sending tasks, or clients, block until they receive a response from receiving tasks, or servers.

State Machine System. In a state machine architecture, the system is broken down into a set of concurrent extended finite state machines. Each of the finite state machine is typically used to model the behavior of a reactive or active object.

Cyclic executives schedule system is illustrated as Fig.1.

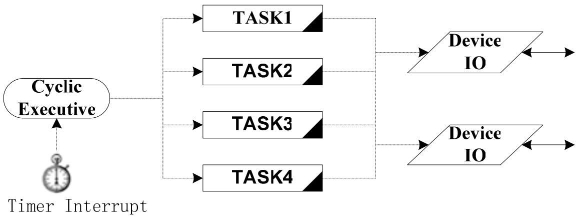


Fig.1. Cyclic executives schedule system

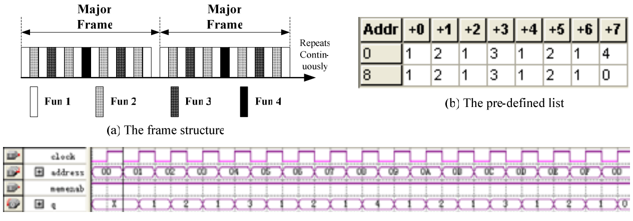
A timeline uses a timer to trigger a task every minor cycle. The operations are implemented as procedures, and are placed in a pre-defined list covering every minor cycle. When a minor cycle begins, the timer task calls each procedure in the list.

Cyclic executives schedule system is described by HDL as follows.

```
module Cyclic_Executives (Timer, address, q,
        clock, memenab );
        lpm_rom rom_list ( .address(address),
        .q(q),. inclock(clock), .outclock(clock),
        .memenab(memenab) );
        defparam rom_list.lpm_width = 3;
        defparam rom_list.lpm_widthad = 4;
        defparam rom_list.LPM_FILE =
        "rom_predef.mif";
always @( Timer )
begin
        TASKi( rom_list(procedure) );      --Pseudocode
end
endmodule
```

In fact, the event-driven systems is similar to the cyclic executives system in addition to the device I/O event of the sensitive list of above model. Because the state machine system is state machine itself, it could be interpreted easily by HDL.

The simulation of cyclic executives[8] system is shown in Fig.2.


Major Frame
Major Frame
Repeats Contin-uously
Fun 1
Fun 2
Fun 3
Fun 4
clock
address
memenab
q
00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 00
X 1 2 1 3 1 2 1 4 1 2 1 3 1 2 1 0

(a) The frame structure

Addr	+0	+1	+2	+3	+4	+5	+6	+7
0	1	2	1	3	1	2	1	4
8	1	2	1	3	1	2	1	0

(b) The pre-defined list

(c) The simulation of cyclic executives system

Fig.2. The simulation of cyclic executives system

Where the "q" in Fig 2(c) represents the scheduled task alone with the timeline indicated by the "address". Suppose that a minor frame is 10 ms (100 Hz) long. Consider 4 functions that must execute at a rate of 50 Hz, 25 Hz, 12.5 Hz, and 6.25 Hz respectively. Then, the frame of tasks and the relevant pre-defined list are showed respectively by Fig2(a) and Fig.2 (b).

4. THE STATE MACHINE OF REALTIME IR DATAPATH

(1) The Logic Description of Synchronization Signal

As discussed above, the synchronization signals of IR video processing system are HSync. and VSync.. The beginning horizontal position and vertical position could be confirmed by the synchronization logic, which carry out the horizontal counter and vertical counter based on the HSync and VSync respectively to locate the image pixel in 2-dimension.

The logic description in HDL lists as follows:

```
SIGNAL   H_CNT,V_CNT: INTEGER RANGE 0 TO
1023;
PROCESS( PCLK ) BEGIN --H Counter
        IF PCLK'EVENT AND PCLK='1' THEN
        IF HSn='0' THEN
                H_CNT <= 1023;
                ELSE
                        H_CNT <= H_CNT + 1;
                END IF;   END IF;
END PROCESS;
PROCESS( HSn )   BEGIN --V Counter
        IF HSn'EVENT AND HSn='1' THEN
        IF VSn='0' THEN
                V_CNT <= 1023;
        ELSE
                V_CNT <= V_CNT + 1;
        END IF;   END IF;
END PROCESS;
```

Where the variable PCLK, HSn and VSn in above HDL list are image pixel clock, HSync. and VSync. respectively, H_CNT and V_CNT are counter value which can express 1024X1024 resolution pixels maximally.

The simulations are illustrated as Fig.3:

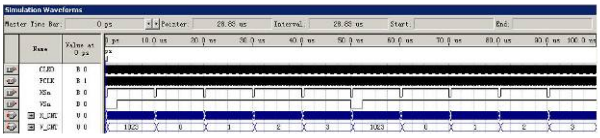


Fig.3. (a) The Simulation of Synchronization from VSync.

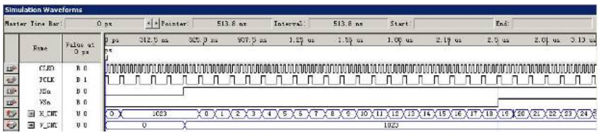


Fig.3. (b) The Simulation of Synchronization from HSync.

(2) The Logic Description of State Machine of Local Bus

The design of state machine of local bus is according to the signal time of PCI specifications to control and acknowledge, its state machine is described as Fig.4:

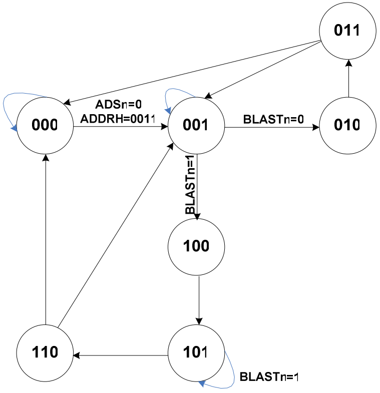


Fig.4. The State Machine of Local Bus

The simulations are illustrated as Fig.5:

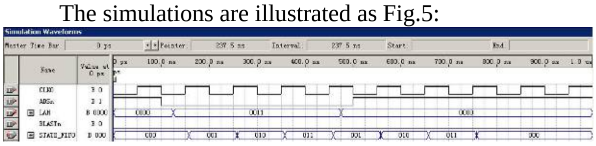


Fig.5.（a） The Simulation of State Machine of Local Bus in Non-Burst

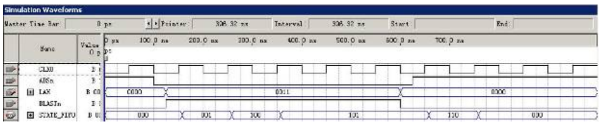


Fig.5.（b） The Simulation of State Machine of Local Bus in Burst

The condition of BLASTn=1 in Fig.5 (b) indicates the burst transfers way.

5. CONCLUSIONS

The concept of hardware logic task and its implementation by HDL are put forward in the paper. It has the characteristics of real-time and overhead saving to realize the tiny-core operating system. Furthermore, the system of hardware logic task has been discussed in two layers of hierarchy, i.e. the inner state transfer of hardware logic task and the schedule among them. And finally, the concepts of hardware logic task have proved to be correct through simulation results.

The conventional tasks must compete the unique CPU, so multi-tasks can not concur under operating system [9],[10]. Whereas the hardware logic tasks come true, because they not only run under CPU but also operate by their state machine to realize the concurrent I/O process.

The configuration of IR video processing system is as follows:

- CPU: GedoeTM SC1200 266MHz
- Memory: 32MB Flash，32MB SDRAM
- OS: General Linux with 2.4.18 kernel
- PCI chip: 32bit/33MHz PCI9054 made by PLX Co.
- Data path: Described in the paper
- Source video: Video in 640X480 resolution, 60fps

The speed of IR video processing can reach 60fps when the source video is operated in simple pseudo-color conversion, but it slow down to 30fps when added complicated FPA nonlinear correction and images processing.

REFERENCES

[1] Jean J L．*MicroC/OS-II The Real-Time Kernel*, Second Edition．Beijing: CMP Books, CMP Media LLC. 2002.

[2] ZOU Si yi． *Embedded Linux Design and Application*．Beijing: Tsinghua University Press. 2002.

[3] YONG ZHU. “The infrared video image pseudocolor processing system,” SPIE, *APPL OF DIGITAL IMAGE PROCESSING XXVI* Vol. 5203，ISSN 0277-786X，Aug. 2003.

[4] YONG ZHU. “The implementation of thermal image visualization by HDL based on pseudo color,” SPIE, *APPL OF DIGITAL IMAGE PROCESSING XXVII* Vol. 5558，ISSN 0277-786X，Aug. 2004.

[5] Zhu Yong, “Digital Process System for Infrared Thermal Image,” *Instrument Technique and Sensor*, 2001,9.

[6] Daniel D. Gajski *et al*, *Specification and Design of Embedded Systems*, 2005,7.

[7] “The Concise Handbook Of Real-Time Systems,” 2002 TimeSys Corporation Pittsburgh PA． www.timesys.com

[8] Sprunt H B, Sha L, Lehoczky J P．”A Periodic Task Scheduling for Hard Real-Time Systems,” *The Journal of Real-Time Systems*, 1989, 1: 27-60.

[9] YANG Ke feng, SHAO Shi. “Scheduler of Embedded Real Time System,” *Application Research Of Computers*. 2001, 8 : 31-33.

[10] HONG Ying, CHEN Xi. “EMBEDDED REAL-TIME MULTI-TASK PROGRAMMING,” Computer Applications. 2000, 7: 10-12.