2004 International Symposium on
Distributed Computing and Applications to
Business, Engineering and Science

DCABES 2004

PROCEEDINGS

Volume II

Editor in Chief Guo Qingping

Hubei Science and Technology Press, Wuhan, China

2004 International Symposium on Distributed Computing and Applications to Business, Engineering and Science

DCABES 2004

PROCEEDINGS

Volume II

Editor in Chief Guo Qingping

Wuhan, China
September 13-16, 2004

Hubei Science and Technology Press, Wuhan, China

图书在版编目(CIP)数据

2004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/郭庆平主编.—武汉：湖北科学技术出版社，2004.9
ISBN 7-5352-3269-8

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(2004)第093693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
ISTCA International Science and Technology Cooperation Association of Hubei Province
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
SUN Microsystems (Hong Kong Headquarter)

DCABES 2004 PROCEEDINGS © Editor in Chief Guo Qingping

Editorial Production by Wu Ruilin, Li Zhiming 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 September 2004 First Impression September 2004

ISBN 7-5352-3269-8/ TP·79 Price RMB 150

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

CONTENTS

Volume II

PREFACE

High-performance computing is increasingly being used in all aspects of modern society. It is well known that the distributed parallel computing plays a main role in the HPC. 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.

It is the second time for the DCABES international conference to be held in Wuhan China. We are gratified that this time nearly 400 papers submitted 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 referee's. All papers contained in this Proceedings give us a glimpse of what future technology and applications are being researched 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 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 Tsui, Mr Y M Thomas, Chinese University of Hong Kong, for his enthusiastically taking part in and supporting the DCABES conference.

I am also grateful to Prof Peter Jimack, University of Leeds, UK; Prof Peter Kacsuk, LPDS, MTA SZTAKI, Hungary; Prof Simon Cox, University of Southampton, UK and Dr. Simon See, Global Science and Technology Center, Sun Microsystems Inc, Singapore, 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 2004 could not be held in Wuhan 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. It should also be mentioned that the SUN Microsystem (Hong Kong Headquarter) made a contribution to 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. Zhou Cheng for the conference website design, Mr. Zhou Cheng and Yang Hao for their efforts in organizing activities. It also should be mentioned that my graduate students, Mr. Ouyang Lin, Tang Guosheng, Yang Hao, Zhang Feng, Chen Jun, Shen Dingcai, Li Xiaoxin, Mao Liming, Zhang Feng and Ms Rao Jing of the grade 2002; Mr. Zhou Cheng, Wu Yanmao, Wang Qingsong, Liu Feng, Sun Hao, Cheng Haifeng, Han Guangming, Jiang Weijian Wu Weiwei of the grade 2003, and graduate student Guo Yucheng spent a lot of time and efforts typesetting the proceedings. Without their help the proceedings could not looks so good.

Enjoy your stay in Wuhan. Hope to meet you again at the DCABES 2005.

Guo, Professor Qingping
Chair of the DCABES2004
Dept. of Computer Science
Wuhan University of Technology
Wuhan, China

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

Scientific Committee (in alphabetical order)

Cai, Professor X.-C.	University of Colorado, Boulder, U.S.A.
Cao, Professor J.W.	Research and Development 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
Kang, Professor L.S.	Wuhan University, Wuhan, China
Keyes, Professor D.E.	Columbia University, New York, USA
Lai, Dr. C.-H.	University of Greenwich, London, UK
Lee, Dr. John.	Hong Kong Polytechnic, Hong Kong, China
Liddell,Professor H. M.	Queen Mary College, University of London, 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, Dr. Michael	University of Hong Kong, Hong Kong, China
Sun, Professor J.	Academia Sinica, Beijing, China
Tsui,Thomas	Chinese University of Hong Kong, Hong Kong, China
Xu, Professor W.	Southern Yangtze University, Wuxi, China

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, Processor L. (Co-Chair)	Wuhan University of Technology, Wuhan, China
Liu, Professor Z.Y.	Wuhan University of Technology, Wuhan, China

Chen, Professor J.S.	Wuhan University of Technology, Wuhan, China
Kang, Professor L.S.	Wuhan University, Wuhan, China
Jin, Professor Hai	Hua Zhong University of Science and Technology, Wuhan, China
Liu, Professor Q.	Wuhan University of Technology, Wuhan, China
Lu, Professor J.G.	South Central China Nationality University, Wuhan, China
Tan, Professor L.S.	Central China Normal University, Wuhan, China
Xu, Prof. H.Z.	Wuhan University of Technology, Wuhan, China
Zeng, Professor C.N.	Wuhan University of Technology, Wuhan, China

REFEREES

Douglas, Professor Craig, Yale University, USA
Guo, Professor Q. P., Wuhan University of Technology, Wuhan, China
Ho, Dr. P. T., University of Hong Kong, Hong Kong, China
Jesshope Professor Chris R., Hull University, Hull, UK; Director of NZEdSoft, New Zealand
Kwan, Mr. W. K., University of Hong Kong, Hong Kong, China
Lai, Dr. Choi-Hong, University of Greenwich, London, UK
Lee, Dr. John, Hong Kong Polytechnic University, Hong Kong, China
Liddell, Professor Heather, Queen Mary and Westfield College, University of London, London, UK
Lin, Dr. Ping, National University of Singapore, Singapore
Loo, Dr. Alfred, Lingnan University, Hong Kong, China
Lu, Professor Zhengding, Huazhong University of Science and Technology, Wuhan, China
Ng, Dr. Michael, University of Hong Kong, Hong Kong China
Paker, Professor Yakup, Computer Science Department, QMW, University of London, London, UK
Tan, Professor L.S., Central China Normal University, Wuhan, China
Jin, Professor Hai, Hua Zhong University of Science and Technology, Wuhan, China
Xu, Professor W, Southern Yangtze University, Wuxi, China
Ng, Dr. Michael, University of Hong Kong, Hong Kong, China
Cai, Professor X.-C., University of Colorado, Boulder, U.S.A
Cao, Professor J.W., Research and Development Centre for Parallel Algorithms and Software, Beijing, China
Keyes, Professor D.E., Columbia University, New York, USA
Lin, Dr. H.X., Delft University of Technology, Delft, Netherlands
Dr. Rüdiger Reischuk, Universität Lübeck, Germany
Prof. Dr. Joerg Rothe, Institut fuer Informatik, Universitaet Duesseldorf
A/Prof. Dr.Nayyer Masood, COMSATS Institute of Infromation Technology, Wah Cantt, Pakistan
Dr. Mike Brayshaw, QMW, University of London, London, UK
Dr. Ajay K Katangur, Department of Computer Science, Georgia State University, Atlanta, USA
Dr Bing Wang, Computer Science Department, University of Hull, Hull, UK
Associate Prof. Yuh-Shyan Chen, National Chung Cheng University Taiwan, China
Miss Srilaxmi Malladi, Georgia State University, Atlanta, USA
Prof. Yi Pan, Georgia State University, Atlanta, USA
Dr. HE, Lifeng, Faculty of Information and Computer Sicience, Aichi Prefectural University, Aichi, Japan
Heng Pheng Ann, Professor, the Chinese University of Hong Kong, Hong Kong, China
Dr. Shiduan Cheng, BUPT, Beijing, China
Dr Tang Ming Xi, School of Design, the Hong Kong Polytechnic University, Hong Kong, China
Dr. Christian Sohler, Computer Science Dept., UPB, Germany
Mr. Praveen Madiraju, Georgia State University, Atlanta, USA
Professor Madhusudhan Govindaraju, Binghamton, UK

Dr. Ze Dang, zdang@eecs.wsu.edu,
Prof. Henry Wolkowicz, Dept of Comb and Opt, University of Waterloo, Waterloo, Ont. CANADA
Dr. Wong Tien Tsin, Chinese University of Hong Kong, Hong Kong, China

ISBN 7-5352-3269-8

9 787535 232694 >

ISBN 7-5352-3269-8/TP · 79

Price RMB 150/YUAN

Deadlock Detection and Resolution in A Dike Safety Detection Management Information System

Wu Jie, Liu Xiangsheng, Wu Wei
Engineer，Yangtze River Scientific Research Institute
Senior Engineer，Yangtze River Scientific Research Institute
Assistant，Wuhan Polytechnic University

ABSTRACT

In this paper, we discussed the case of deadlock in a distributed database management system. An emphasis is how to detect or resolve deadlock in a dike safety detection management information system.

Keywords: Distributed database, Deadlock detection, Deadlock resolution.

1. INTRODUCTION

A Dike Safety Detection Management Information System (DSDMIS) is a distributed system which is designed to manage detection information for dike safety. In this system, according to some parametric and corresponding mathematical computational models about the dike, we can speculate the safety state of the dike.

Due to the dike is so long, therefore, corresponding to the system’s several sites, the data of the dike are fragmentally imported. Then，all the data of the dike are processed in the central processor unit. The DBMS of the DSDMIS we used is SQL server. To keep the data consistency, we use data locking technique. And what is a locking technique? Let’s introduce it from the distributed database.

2. WHAT IS A DISTRIBUTED DATABASE SYSTEM

The advances in networking technology are increasingly making the deployment of distributed system architectures a popular, sometimes even an essential option. The main advantages of distributed system architectures include increased overall system availability through better fault tolerance, parallel execution of an application on multiple hosts and a simplication of scalability.

A distributed database can be defined as a collection of multiple, logically interrelated database distributed over a computer network. A DDBS is then defined as the software system that permits the management of the DDBS and makes the distribution transparent to the users.

To form a DDBS, files should not only be logically related, but also should be structuralized, and data access should be via a common interface.

3. LOCKING-BASED CONCURRENCY

In a distributed database, there must be some operations on data. To keep the data consistency, a control of locking-based concurrency is done.

The main idea of locking-based concurrency control is to ensure that the data is shared by one operation at a time. This is accomplished by associating a “lock” with each lock unit. This lock is set by a transaction before it is accessed and is reset at the end of its use. Obviously a lock unit cannot be accessed by an operation if it is already locked by another operation. Thus a lock request by a transaction is granted only if the associated lock is not being held by any other transaction.

4. DEADLOCK IN DDBS

Deadlock literature formally defines a deadlock as, "A set of processes is deadlocked if each process in the set is waiting for an event that only another process in the set can cause". A more informal description is that deadlocks can occur whenever two or more processes are competing for limited resources and the processes are allowed to acquire and hold a resource (obtain a lock). If a process waits for resources, any resources it holds are unavailable to other processes. If a process is waiting on a resource that is held by another process, which is in turn waiting on one of its held resources, we have a deadlock. When a system attains this state, it is effectively dead and must resolve the problem to continue operating. There are four conditions that are required for a deadlock:

1. Mutual exclusion: Each resource can only be assigned to exactly one resource.
2. Hold and wait: Processes can hold a resource and request more.
3. No preemption: Resources cannot be forcibly removed from a process.
4. Circular wait: There must be a circular chain of processes, each waiting for a resource held by the next member of the chain.

Any locking-based concurrency control algorithm may result in deadlocks, since there is mutual exclusion of access to shared data and transaction may wait on locks.

Deadlocks can arise in each database system that permits concurrent execution of transactions using pessimistic synchronization schemes, i.e., locking protocols, which is the case in most of today (distributed) database systems. In centralized database systems, deadlock detection and resolution has been thoroughly investigated. Deadlocks have also been studied in other areas, such as operating systems. It is a permanent phenomenon. If one exists in a system, it will not go away unless outside intervention takes place. This outside intervention may come from the user, the system operator, or the software system.

From what have been described, we can see that the DSDMIS don’t avoid the deadlock. How can deadlock to prevent, detection or resolution?

5. DEADLOCK PREVENTION

There aren't some good methods to prevent deadlock in database environments.

Commonly, a transaction is checked by the transaction manager when it is first initiated and is not permitted to proceed if it may cause a deadlock. To perform the check, it is required that all of the data items that will be accessed by a transaction be predeclared. The transaction manager then permits a transaction to proceed if all the data items that will be accessed are available. Otherwise, the transaction is not permitted to proceed. The transaction manager reserves all the data items that are predeclared by a transaction that it allows to keep on.

Because of that it is usually difficult to know precisely which data items will be accessed by a transaction. Access to certain data items may depend on conditions that may not be resolved until run time. So the systems are not very suitable for database environments.

6. DEADLOCK AVOIDANCE

No deadlocks occur if there are no cycles in the waits for Graph

– Order all of the locks in the systems and all transactions acquire locks in this order

– Transactions are either ahead or behind of other transactions, but no wait cycles occur

There are also avoidance techniques based on a predeclaration of all locks that a transaction will use. Why is this not practical?

– Need an on-line, dynamic solution

– Not a useful interface, few applications know or can express their needs ahead of time

7. DEADLOCK DETECTION AND RESOLUTION

There are four techniques commonly employed to deal with deadlocks: ignore the problem, deadlock detection, deadlock prevention and deadlock avoidance. Ignoring deadlocks is the easiest scheme to implement. Deadlock detection attempts to locate and resolve deadlocks. Deadlock avoidance describes techniques that attempt to determine if a deadlock will occur at the time a resource is requested and react to the request in a manner that avoids the deadlock. Deadlock prevention is the structuring of a system in such a manner that one of the four necessary conditions for deadlock cannot occur. Each solution category is suited to a specific type of environment and has advantages and disadvantages.

Blocking conditions between transactions can be represented through a transaction wait-for graph. A WFG is a directed graph in which nodes correspond to transactions and a directed edge from Ti to Tj expresses that Ti waits for a resource currently held by Tj. A deadlock can be detected by examining the structure of the WFG. Which graph structures indicate a deadlock depends on which deadlock model applies, as described in the next section.

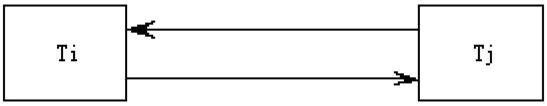


Figure 1: Wait-for Graph

In distributed systems, it is not sufficient that each local distributed DBMS form a local wait-for graph (LWFG) at each site, it is also necessary to form a global wait-for graph (GWFG) which is the union of all LWFGs.

There are three detection methods of deadlock methods which are commonly called centralized, distributed and hierarchical deadlock detection. In the centralized deadlock detection approach, one site is designated as the deadlock detector for the entire system. Periodically, each lock manager transmits its LWFG to the deadlock detector, which then forms the GWFG and looks for cycles in it.

An alternative to centralized deadlock detection is the building of a hierarchy of deadlock detection. Deadlock are local to a single site would be detected at that site using the local WFG. Each site also sends its local WFG to the deadlock detector at the next level. Thus, distributed deadlocks involving tow or more sites would be detected by a deadlock detector in the next lowest level that has control over these sites.

Distributed deadlock detection algorithms delegate the responsibility of detecting deadlocks to individual sites. Thus, as in the hierarchical deadlock detection, there are local deadlock detectors at each site which communicate their local WFGs with one another.

8. DEADLOCK IN DSDMIS

The occurrence of deadlocks in a database application is unpleasant. Even if it only happens every now and then, successful working with the application is no longer possible! They diminish acceptance of the application, since deadlocks are often an application problem.

The basic reason of the all deadlock is resource rival.
There are always two deadlocks:
The first one is :
– Two transactions: Ti and Tj
– Two resources: A and B
– T1 holds A wants B
– T2 holds B wants A
– Neither transaction is will to release its current holding

Because resource B is hold by Tj, so Ti has to wait for resource B is released by Tj. In the same time, Tj is waiting for the resource A which is hold by Ti. So, the deadlock is setup.
This deadlock caused by bug in the program, so there is the only way to resolve the deadlock is to modify the program. To analysis the logic of the program the time:

- Two resources must not to be hold at the same time
- If there are to avoid hold two resources at the same time, it must ensure the resources be hold in the same order at any

time

The second deadlock is:

- Transaction Ti read resource A, and then modifies it.
- Here transaction Tj modifies resource A.

It can be found that, the lock in the Ti move up form shared lock to exclusive lock. Whereas the exclusive lock on resource locked by Tj must wait for Ti to release the shared lock on resource A. In the same time, the shared lock on resource A locked by Ti wait for Tj to release the exclusive lock ,and the shared lock will move up to exclusive lock. In this case, the two transactions Ti and Tj will wait indefinitely for each other to release their respective locks.

So the deadlock is took place. The second deadlock always occurs in some large project.

The resolution of the deadlock is:

The transaction A (read-then-write operate) use “Update lock” when it select.

Grammar as follows :

```
select* from table1 with(updlock)
  where ....
```

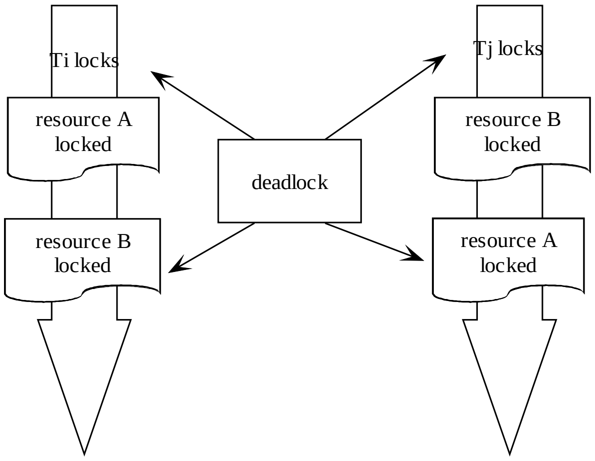


Figure 2: Deadlock

9. TIME-OUT LAW

Because the deadlocks always occur in the large project, so use the Time-out law. Time-out law is the simplest method in the deadlock resolution methods. Its principle is that when transactions don’t gain a lock in a long time after they apply for some right to lock some data items, it will be recognized that the system be in deadlock state. So the system must process deadlock, it will abort this transaction，give off the resource appropriative by it .

The most important in the method is that how long we can wait for the lock. The time will not too long, otherwise there will waste resource after a long time to process transactions. And the time will not too short similarly. Because the short time will cause the transactions abort frequently, even bring into abortion all transactions.

10. THE SCHEME TO RESOLVE THE DEADLOCK IN DSDMIS

According to the complexity of procession the data item in the DSDMIS, we design the plan as follow: the system use the time-out law to resolve the deadlock, and the wait-for time is two minutes. If there is any deadlock, the system will use the centralized deadlock detection approach to find the transaction which will be aborted.

11. CONCLUSION

Sophisticated and reliable distributed DBMSs are now available in the market, but there is also a number of issues need to be solved satisfactorily.

The distributed deadlocks prevent DDBS from completing successfully used. Only to successfully detect prevent and resolve distributed deadlocks can make DDBS available.

12. REFERENCES

[1] M.Tamer Özsu ,Patrick Valduriez,(2002),Principles of distributed database systems. Prentice-hall,Inc.

[2] Doreen L.Galli.(2003),distributed operating systems concepts & practice. Prentice-hall,Inc.

[3] Hector Garcia-Molina, Jeffrey D.Ullman, Jenninfer Widom. (2003), Database systems :the complete book. Pearson educaton ,Inc.

[4] Tanenbaum, A: Modern Operating Systems. Prentice Hall Inc., Englewood Cliffs, (1992)

[5] Coulouris, G., Dollimore, J. and Kindberg, T.: Distributed Systems Concepts and Designs. 2nd ed. Addison-Wesley, Don Mills, Ontario, (1994)

[6] Randal Burns, Distributed Database Systems Scheduling and Deadlock, Computer Science 600.416 Johns Hopkins University

Mining Fuzzy Associate Rules for Anomaly Detection

Xiong Ping, Zhu Tianqing, Huang Tianshu
School of Electron and Information, Wuhan University, WuHan 430079.China
Email: zhutq@126.com **Tel.**: 027-62034306

ABSTRACT

In this paper, we describe the technology of mining fuzzy associate rules. An approach is presented that the fuzzy sets of each transaction’s attributes is divided and calculated as separate attributes in mining fuzzy associate rules. The process of applying the approach for anomaly detection is discussed in detail. Using experiments on network traffic analysis, the feasibility of applying the mining fuzzy associate rules in intrusion detection is validated. Finally, we establish response mechanism according to the similarity of rule sets.

Keywords: Anomaly Detection, Data Mining, Fuzzy Associate Rules.

1. INTRODUCTION

As network-based computer systems play increasingly vital roles in modern society, security of network systems has become more important than ever before. Intrusion detection System (IDS) has thus become a critical technology to help protect these systems. As a key role in security architecture of network, IDS attempts to identify existing attack patterns and recognize new intrusion, employing methods from fields such as mathematics, statistics and machine learning.

An intrusion can be defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource[1]. IDS approaches can be divided into two main categories: misuse or anomaly detection[2]. However in practice, IDS is still unsatisfactory for its primary problems of false negatives and false positives, which give a misleading sense of security. Network administrators have to analyze the reports of IDS artificially, intuition and experience are comparatively relied upon in making right decision. Thus applying soft computing methods has become a research direction in intrusion detection.

Recently, there has been an increased interest in applying data mining approaches to build detection models for IDSs[3][4]. However, data mining technology is generally more effective in handling boolean attributes. For quantitative attributes, the attribute domain should be divided into several discrete intervals in preprocess of data mining, which causes the problem known as “sharp boundary”. To solve the problem, fuzzy theory has been applied and fuzzy data mining technology was proposed in some papers. In [5], the problem “sharp boundary” and fuzzy data mining technology have been introduced.

In this paper, we use fuzzy association rules mining in anomaly detection by improved algorithm of Apriori (an classic algorithm of mining association rules).

The rest of this paper is organized as follows. In Section 2, we propose the approach that applying fuzzy data mining in IDS. In Section 3, an experiment of mining association rules of network traffic attributes is described in detail. According the experimental results, the response mechanism is built in section 4. Section 5 offers conclusive remarks and future work.

2. ANOMALY DETECTION WITH FUZZY ASSOCIATION RULES

Mining fuzzy association rules in anomaly detection includes steps as follows:

A. Mining fuzzy association rules set S with system in normal state;
B. Mining fuzzy association rules set S_1 with system in given transient state;
C. Calculating the similarity of S and S_1 to evaluate the real-time system state.

2.1 Anomaly Detection

Anomaly detection works on the assumption that many attackers behave differently from normal users, or that a system or process behaves differently during an attack. Thus anomaly detection can be viewed as finding non-permitted deviations of the characteristic properties in the monitored system[6]. The normal state profile should be defined by some characteristic parameters and threshold of network or system at first. Comparing the normal profile with real-time transient profile and evaluating the deviation degree, anomalies or intrusions in system can be detected.

The characteristic parameters monitored come from multiple levels including user level, system level, packet level and etc. In [7], the parameters and the Commands used to obtain data of the parameters were described. However, the selection of these parameters is not final and may vary (based on their usefulness) in future implementation.

2.2 Mining Fuzzy Association Rules

Let $D=\{d_1,d_2,\ldots,d_n\}$ be the database and $d_i(i>0)$ represents the i^{th} column in D. A database sample is shown in table 1.

Table 1 A Sample Database

TCP	UDP	ICMP
56.3	72.5	65
45.5	58	23
62	36.2	54.5
68.2	54.1	21.4

According to table 1, $D=\{d_1,d_2,d_3,d_4\}$, $d_i(1\leq i\leq 4)$ is a transaction of D, for example, $d_3=\{62,36.2,54.5\}$.

Let $W=\{w_1,w_2,\ldots,w_m\}$ be the attribute set of database D. According to the fuzzy theory, each attribute $w_k\,(1\leq k\leq m)$ is divided into several fuzzy set: $w_k=\{w_{k1},w_{k2},...,w_{kt}\}(t>0)$. For example, the attributes set of database D shown in table 1 is $W=\{TCP, UDP,ICMP\}$ where $W_1=TCP$, $W_2=UDP$ and $W_3=ICMP$. And each attribute is composed of several fuzzy sets. For example, the UDP is composed of three fuzzy sets: *high*, *medium* and *low*, namely, $w_{UDP}=\{high,medium,low\}$. The membership function of each fuzzy set will be given by domain experts.

Similar to classic data mining, we hope to get the association rules set in the following form:

$$<X,A>\rightarrow<Y,B>\ [c,s],$$

where $X=\{x_1,x_2,...,x_p\}$ and $Y=\{y_1,y_2,...,y_q\}$ are subsets of attributes set W and $X\cap Y=\varnothing$. $A=\{w_{x1},w_{x2},...,w_{xP}\}$ and $B=\{w_{y1},w_{y2},...,w_{yq}\}$ are the fuzzy sets of X and Y respectively. s(support degree) and c(confidence degree) should be less than the predefined threshold, i.e. the minimum support degree (*minsupport*) and minimum confidence degree(*minconfidence*). To generate fuzzy association rules, we have first to calculate the support degree to find out all the frequent sets. Let $<X,A>$ represent the itemset-fuzzy set pair, where X is a set of attributes x_j and A is the set of fuzzy sets a_j. A transaction satisfies the pair $<X,A>$ means that the vote of the transaction is greater than zero. A sample of attribute-fuzzy set pair and corresponding membership degree of the attributes in table 1 is shown as table 2. In this case, X={TCP,UDP},A={high, low}.

Table 2 A Samples of Attribute-Fuzzy Set Pair

<TCP, high>	<UDP, low>
0.6	0.3
0.2	0.7
0.8	0.5
0.8	0.3

The vote of each transaction is calculated by the membership function of x_j. We use $d_i[x_j]$ to represent the value of x_j in the i^{th} transaction, and $m_{aj\in A}(d_i[x_j])$ is the membership degree of d_i. After getting all the membership degree in a transaction, we can get the final vote of the transaction to $<X,A>$: $\Pi_{xj\in X}\{\alpha_{aj}(d_i[x_j])$. Summing up the votes of all transactions and dividing the value by the total number of transactions, we can get the support degree s finally. Thus the support degree is calculated with the following formula:

$$s_{<X,A>}=\frac{\sum_{\mathrm{d_i}\in D}\prod_{x_j\in \mathrm{X}}\{\alpha_{aj}(d_i[x_j]\}}{total(D)} \qquad (1)$$

where $\alpha_{aj}(d_i[x_j])=m_{aj\in A}(d_i[x_j])$.

For example, calculating the support degree and confidence degree according to the table 1 and table 2, we can get:
$s_{<X,A>}$=(0.18+0.14+0.40+0.24)/4=0.24, and
$c_{<X,A>}$=(0.18+0.14+0.40+0.24)/(0.6+0.2+0.8+0.8) =0.4.

Otherwise, because that some attributes are divided into several fuzzy sets, the calculating cost will be much greater applying the classic Apriori algorithm. Thus we improve the algorithm that the elements in fuzzy sets of transaction's attributes are handled as separate attributes of database. The item sets in which some attributes belong to the same fuzzy set will be deleted in the pruning process after the join step. Application of the new algorithm will be described in detail in section 3.

2.3 Similarity of Rule Sets[8]

Given two association rules $R1:X\rightarrow Y,c,s$ and $R2:X'\rightarrow Y',c',s'$, when $X=X'$ and $Y=Y'$, the similarity of the two rules can be calculated by the formula:

$$similarity(R1,R2) = max(0,1-max(\ |c-c'|/c,\ |s-s'|/s)) \qquad (2)$$

When $X\neq X'$ and $Y\neq Y'$, the similarity of the rules will be zero.

For two given rule sets S_1 and S2, the similarity of the sets is:

$$similarity(S_1,S_2)=\frac{s^2}{|S_1|*|S_2|} \qquad (3)$$

where $s=\sum_{\substack{\forall R_1\in S_1\\ \forall R2\in S_2}} similarity(R1,R2)$, $|S_1|$ and $|S_2|$ are respectively the number of rules in S_1 and S_2.

3. EXPERIMENTATION

To validate the feasibility of the method presented above, we analyze the network flow of an LAN in experimental environment. By mining the databases of network flow in normal state and attack state, the profile of system in each state are defined with fuzzy association rule set.

3.1 Mining Association Rules

In the experiment, four characteristic parameters related to network flow are selected to analyze the network communication in different states, including P_{tcp}(the proportion between number of *TCP* packets and the number of total packets), P_{udp}(the proportion between number of *UDP* packets and the number of total packets), Avg.packet/sec(the average number of packets in one second) and Avg.Mbit/sec(the average number of bits in one second). We capture packets every hour and calculate the value of the parameters for ten times, thus we get ten groups of data as shown in table 3.

Table 3 Experimental Data

P_{tcp}(%)	P_{udp}(%)	Avg.packet/sec	Avg.Mbit/sec
96.0	0.2	169.541	0.530
95.9	0.5	171.836	0.531
93.8	0.4	183.936	0.596
96.2	1.0	171.477	0.523
95.7	0.7	133.544	0.407
92.2	0.8	168.651	0.531
97.2	0.7	177.258	0.565
85.0	0.6	193.379	0.547
89.0	0.3	141.256	0.425
84.6	0.3	190.285	0.524

According to the value in table 3, we divide each attribute into tow fuzzy sets (*high* and *low*) and present the membership functions respectively. For example, the membership functions of P_{tcp} are presented as follows:

$$low：y=\begin{cases} 1 & (x\leq 86.0) \\ -0.12x+11.32 & (86.0<x<94.3) \\ 0 & (x\geq 94.3) \end{cases}$$

$$high：y=\begin{cases} 0 & (x\leq 90.0) \\ 0.15x-13.5 & (90.0<x<96.7) \\ 0 & (x\geq 96.7) \end{cases}$$

The function curve is shown in figure 1.

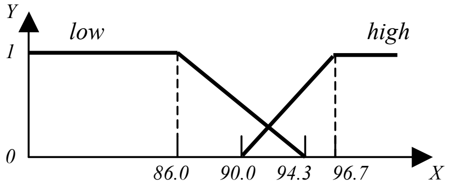
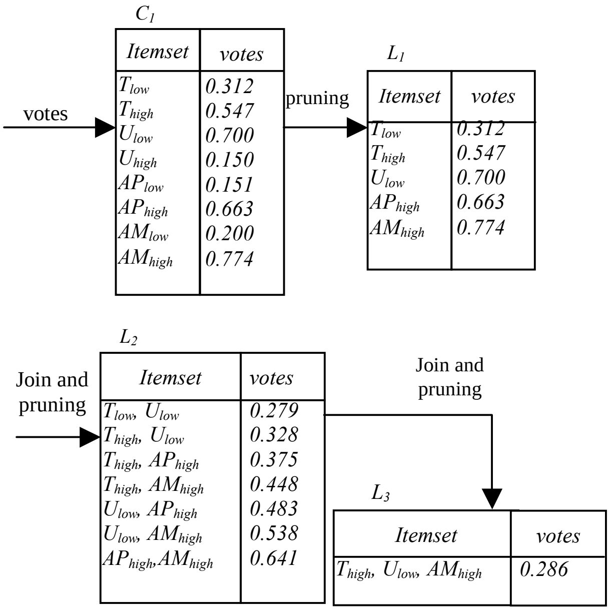


Fig.1: Membership Function

Thus each attribute is replaced by two fuzzy sets(for example, P_{udp} is replaced by U_{low} and U_{high}) and table 3 in converted to a table with 8 separate attributes including T_{low}, T_{high}, U_{low}, U_{high}, AP_{low}, AP_{high} ,AM_{low} and AM_{high}.

Now we set the values of minimum support degree and minimum confidence degree: *minsupport=0.25 and minconfidence=60%.*

Calculating the votes of each attribute and handling the data according to the improved Apriori algorithm with join and pruning process, we can get an frequent set which contains the most of attributes. A key step is that the item sets in which some attributes belong to the same fuzzy set should be deleted in the pruning process. For example, the 2-itemsets contains some itemsets such as {T_{high},T_{low}} after the join step to reduce the calculating cost. The itemset should be deleted because the items T_{high} and T_{low} belong to the same attibute P_{tcp}. The whole data mining process is shown in figure.2.



C_1

Itemset	votes
T_{low}	0.312
T_{high}	0.547
U_{low}	0.700
U_{high}	0.150
AP_{low}	0.151
AP_{high}	0.663
AM_{low}	0.200
AM_{high}	0.774

L_1

Itemset	votes
T_{low}	0.312
T_{high}	0.547
U_{low}	0.700
AP_{high}	0.663
AM_{high}	0.774

L_2

Itemset	votes
T_{low}, U_{low}	0.279
T_{high}, U_{low}	0.328
T_{high}, AP_{high}	0.375
T_{high}, AM_{high}	0.448
U_{low}, AP_{high}	0.483
U_{low}, AM_{high}	0.538
AP_{high},AM_{high}	0.641

L_3

Itemset	votes
T_{high}, U_{low}, AM_{high}	0.286

Fig.2 Generation of Frequent Set

Based on the frequent set L_3 obtained in fig.2, we extract the association rules as follows:

$T_{high}\wedge U_{low}\rightarrow AM_{high}$	*s=0.286，c=87.2%*
$T_{high}\wedge AM_{high}\rightarrow U_{low}$	*s=0.286，c=63.8%*
$U_{low}\wedge AM_{high}\rightarrow T_{high}$	*s=0.286，c=53.2%*
$T_{high}\rightarrow U_{low}\wedge AM_{high}$	*s=0.286，c=52.3%*
$U_{low}\rightarrow T_{high}\wedge AM_{high}$	*s=0.286，c=40.1%*
$AM_{high}\rightarrow T_{high}\wedge U_{low}$	*s=0.286，c=36.9%*

Restricted by *minsupport* and *minconfidence*, the final strong association rules set *S* is:

R_1：$T_{high}\wedge U_{low}\rightarrow AM_{high}$ s=0.286，c=87.2%
R_2：$T_{high}\wedge AM_{high}\rightarrow U_{low}$ s=0.286，c=63.8%.

3.2 Realization of Anomaly Detection

To compare similarity between the association rules in normal state and abnormal state of network system, we collect two groups of data respectively with system in normal state and abnormal state(intruding one host computer of system by DoS attack). Mining association rules in the groups of data with the approach described in section 3.1, two rules sets($S_{normal}$and $S_{abnormal}$, represent respectively normal state and abnormal state of system) are output. Thus we use the formula (2) and (3) to calculate the similarity between S_{normal} and $S_{abnormal}$ with the rules set *S* output in section 3.1 and the experimental result is shown in table 4.

Table 4: Similarity of Rules Set

rules set	*rules*	*similarity of rules*	*similarity of rules set*
S_{normal}	$R_{normal1}$:$T_{high}\wedge U_{low}\rightarrow AM_{high}$ *s=0.307，c=83.2%*	*0.954*	*0.845*
	$R_{normal2}$:$T_{high}\wedge AM_{high}\rightarrow U_{low}$ *s=0.307，c=69.3%*	*0.885*	
$S_{abnormal}$	$R_{abnormal1}$：$T_{high}\wedge U_{low}\rightarrow AM_{high}$ *s=0.269，c=78.4%*	*0.899*	*0.404*

The value of similarity in table 4 implies that the rules sets mined with network in normal state and abnormal state are comparatively different. Thereby we can define the threshold value of similarity to ascertain the state of the monitored network system and realize anomaly detection.

4. RESPONSE TO INTRUSION

Response to intrusion is an indispensability part in IDS. The similarity of rules sets quantitates the deviation degree of current system state from the normal state and an less similarity generally means an more dangerous state of monitored system. Thus we divide the domain of similarity into several intervals, which map to responses in different strict degree. The response mechanism of IDS is established as follows:

- *similarity>80%* → Take no action
- *72%<similarity ≤ 80%* → Inform the system administrator via e-mail or messaging system
- *65<similarity≤72%* → Change the priority of user processes
- *57<similarity≤65%* → Block a particular IP address
- *48<similarity≤57%* → Refuse a remote connection request
- *42<similarity≤48%* →Terminate all existing network

connection

- *similarity≤42%* → Restart machine

5. CONCLUSIONS AND FUTURE WORK

Application of mining fuzzy association rules in anomaly detection is a hot spot in research of IDS. In this paper, an approach is presented that the elements in fuzzy sets of attributes are handled as separate attributes of database and the experimentation of network flow analysis validate availability of the approach.

However, anomaly detection is comparatively based on limited statistic data and knowledge of domain experts. Many intelligent methods should be applied to improve the approach in future work, such as tuning the parameters in membership function with genetic algorithm to reduce the similarity when monitored system is in abnormal state, using neural networks to identify intrusive behavior within the analyzed data stream and improve self-learning ability of IDS.

6. REFERENCES

[1] R. Heady,G.Luger,A.Maccabe, and M.Sevilla. The Architecture of a Network-level Intrusion Detection System, Technical report, CS90-20. Dept. of ComputerScience, University of New Mexico, Albuquerque, NM87131.

[2] Stefan Axelsson. Intrusion detection systems: Asurvey and taxonomy. Technical Report No 99-15, Dept.of Computer Engineering, Chalmers University of Technology, Sweden, March 2000.

[3] Wenke Lee at all. Mining audit data to build intrusion detection models. Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'98). pages 66-72, 1998.

[4] Lee W,Stolfo S,ChanPK,Eskin E,Fan W,Miller M, Hershkop S,Zhang J. Real Time Data Mining-based Intrusion Detection. Proceedings of DISCEX II, 2001.

[5] Kuok, C., A. Fu, and M. Wong. 1998. Mining fuzzy association rules in databases. *SIGMOD Record* 17(1): 41-6.

[6] Debar, H., Dacier, M., Wepspi, A. A Revised Taxonomy for Intrusion Detection Systems. Technical Report Computer Science/Mathematics (1999)

[7] Dipankar Dasgupta and Fabio A. Gonzalez. An Intelligent Decision Support System for Intrusion Detection and Response MMM-ACNS, May 21-23, 2001, St. Petersburg, Russia.

[8] WengdongWang. Denetic Algorithm Optimization of Membership Functions for Mining Fuzzy Association Rules. International Joint Conference on Information Systems, Fuzzy Theory and Technology Conference, March 2, 2000

A Pocket Spatial Database Prototype and its Query Language*

Guobao Yu[1]**, Husheng Liao**[2]**, Yuming Zheng**[3]
College of Computer Science and Technology, Beijing University of Technology
Beijing 100022, P.R. of China
Email 1: NLP@bjut.edu.cn **Tel.:** (8610) 6739.1745
Email 2: Liaohs@bjut.edu.cn **Tel.:** (8610) 6739.2987
Email 3: Zhengyuming@bjut.edu.cn **Tel.:** (8610) 6739.1745

ABSTRACT

This paper describes software architecture of a spatial database for pocket computers, its spatial query language named GSQL and some techniques to improve its performance. Some GSQL query examples are also included.. Based on this architecture and its query language, a spatial database prototype has been built to verify its implementation feasibility in lightweight Pocket PC. In the last section of this paper, conclusions and further research outlook for a distributed pocket spatial database are discussed.

Keywords: Geographic Information Systems, Mobile Spatial Database, Spatial Query Language.

1. INTRODUCTION

With the wide spreading of mobile computing platform, mobile Geographic Information Systems (GIS) is bringing fundamental changes to GIS community and general public with the ability of mobile geographic computing. Mobile GIS integrates GIS and lightweight hardware that carries wireless communication functions. Therefore, Mobile GIS has certain different characteristics in database access, map rendering and GIS integration to users from traditional GIS systems.

In this paper, a pocket spatial database prototype called PSDB is presented. We focus on the software architecture, its spatial query language and some features for mobile GIS.

The spatial query is a significant part of Geographic Information Systems (GIS). It can be defined as follows[4]:

Retrieve all entities with non-spatial constrains and spatial constrains.

There are three fundamental categories of queries in Geographic Information Systems[2]:

- Queries exclusively about spatial properties, e.g., "Retrieve all cities that are crossed by the Yangtze River."
- Queries about non-spatial properties, e.g., "Retrieve the population of a city."
- Queries about both spatial and non-spatial properties, e.g., "Retrieve all neighbors of the parcel located at a street."

The third kind of queries is the most common in spatial applications.

GSQL-the Spatial Query Language is based upon the relational database query language SQL, thus relational database users can leverage their experiences on SQL as well as retrieve both spatial and non-spatial properties in the uniform query syntax. In order to handle spatial geometry objects in the expression of the query language, some extensions on geometry data types, spatial operations and relationships have been incorporated in the GSQL. These include both Point, LineString, Polygon, GeometryCollection as GSQL built-in data types and spatial operations/relationships conformed to the OpenGIS Simple Features Specification for SQL[3] such as *contains*, *cross*, *ConvexHull*, *disjoin*, *distance*, *Difference*, *Equal*, *intersect*, *overlap*, *touch*, *within*, *Buffer*, *Union*. To date, only subsets of the above spatial predicates have been developed.

The remainder of this paper is organized as follows: Section II illustrates the architecture of a pocket spatial database (PSDB). Section III discusses PSDB spatial query language and its geometry object model. Some GSQL query examples are also given in this section. Finally, in section IV we draw our conclusions and give some further outlook of our research.

2. POCKET SPATIAL DATABASE ARCHITECTURE

Geometry Object Model

The whole PSDB prototype is composed of five components as briefly described bellow:

1) Client Application
A demonstration client application has been developed to provide a GUI interface for users to easily operate the PSDB prototype. Its major functions are map querying, rendering, zooming, panning and layers managing as well as constructing Geographic query statement.

2) PSDB Interface
PSDB provides a simple interface for client application to parse GSQL statements, and then call low level interface to access spatial database. Finally it constructs the querying result obtained from RDBMS and returns it to the client application.

3) GSQL2SQL
GSQL2SQL converts GSQL statement to standard SQL statement if GSQL statement contains spatial operations or relationships, and send the SQL statement to low level interface to be executed.

4) DB Access Interface ODBC/OKAPI
To simplify spatial database access, we use ODBC as the spatial database access interface. Moreover, ODBC has been a standard access interface for relation database, therefore has better transplanting. Database vendor specific interface such as Oracle Lite access interface – OKAPI should be used to

* This research was partially funded by grants from the Science & Technology Development Program of Beijing Municipal Education Committee.

access Oracle Lite database for performance consideration.

5) Oracle Lite DBMS

Oracle Lite is a light weight DBMS suitable for pocket computers. Therefore, it is chosen as the spatial data repository in the pocket spatial database prototype.

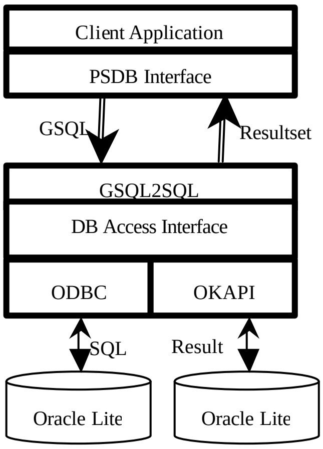


Figure 1 PSDB architecture

Features for Mobile GIS

Mobile devices are resource limited, such as limited memory, storage, display area, etc. All these factors dampen the performance of mobile applications, especially of mobile GIS system. To deal with this problem, we take following measures in the PSDB prototype.

1) Multithreaded Programming

The map queried may consist of multiple layers. We use one thread to query layers and put part of the whole query result to a buffer while using another thread to render map concurrently.

2) Geometric Tolerance

Tolerance is a numerical value defining the acceptable error range that a feature will have from its actual point found on earth. Tolerance can improve performance by reducing amount of spatial feature data in specific tolerance of measures, especially on Pocket PC that has small size screen.

For example, given a geometry feature consisted of points P_1, P_2, ..., P_n where P_1 is the start point and P_n is the end point of the feature, as shown in figure 2.

In our prototype, Geometry Tolerancee is an experience threshold value based on map scale. Given a Geometry Tolerancee, the following pseudo-code algorithm which is based on a algorithm described in natural language in reference[6] can be applied to compress the feature of figure 2. It uses “ divide-and-conquer” strategy and recursive technique to process a curve (LineString).

```
Algorithm to Compact Curve(LineString)
Start = $P_1$;
End = $P_{n;}$;
for k =1 to n do
      $d_k$ = distance($P_k$, Line(Start, End));
$d_i$ = max($d_1$,$d_2$,..,$d_n$) where i $\in$ [1, n];
if ($d_i$   <= e)
       {Replace the curve with Line(Start, End);}
else {
      Divide the curve into curve(Start, $d_i$) and curve($d_i$, End);
      Recursively apply this algorithm to curve(Start, $d_i$);
      Recursively apply this algorithm to curve($d_i$, End);
      };
```

The best time complexity of this algorithm is O(n) whereas the worst time complexity is O(n2). The average worst time complexity is O(n log n).

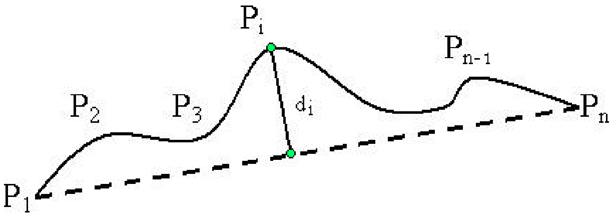


Figure 2 Applying Geometry Tolerance to compact feature

Because a polygon can be divided into two LineString features, the algorithm above could be adjusted to process polygon feature. However, its performance is not good enough since the GIS data is huge. An alternative algorithm without recursion is as follow.

A polygon of n Points could be denoted as polygon (P_1, P_2, ..., P_n), n>=3. A link List of n nodes is used to store the data as shown in figure 3. A node contains a concrete *geometry* object and a pointer to next node.

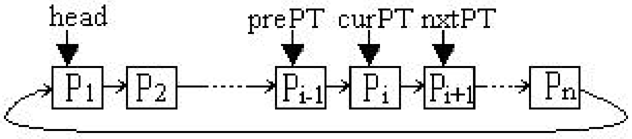


Figure 3 A Link List to Store a Polygon

```
Algorithm to Compact Polygon
//Initialize head pointer and three pointers
//to three consecutive nodes of the link list.
Head = prePT = $P_1$;
curPT = $P_2$;
nxtPT = $P_3$;
while(curPT!=head)
{
      $d_i$ = distance (curPT->Geometry,
        Line(prePT->Geometry, nxtPT->Geometry));
      if ($d_i$ <= $\varepsilon$ )
       {
       //Delete current point from the list:
        RemoveNode(curPT);
      curPT = nxtPT;
      nxtPT = nxtPT->Next;
      }
      else {
            prePT = curPT;
            curPT = nxtPT;
            nxtPT = nxtPT->Next;
      }
}
```

This concise algorithm reduces the calculation of maximum distance and eliminates recursion of divide-and-conquer algorithm. Its time complexity is O(n) compared to precious “divide-and-conquer” algorithm. This is meaningful in GIS, which has massive data.

Even more, this algorithm can also be applied to compact LineString objects.

Another improvement in our prototype is that, Geometry Tolerance **ε** is not only dependent on the map scale, but also dependent on the size of map view range.

3) Partition Query Result
In order to enhance the response time to user requests, PSDB partitions the query result and renders one partition at a time.

4) Replication
PSDB utilize replication tool to replicate main site spatial data to Pocket PC.

3. GSQL AND ITS GEOMETRY OBJECT MODEL

This section introduces the spatial query language and its geometry object model, which is the fundamental of the GSQL.

Geometry Object Model
The object model for geometry is organized into a class hierarchy based on the Open GIS Geometry Model[3], and is shown using UML class diagram in the Figure 4.

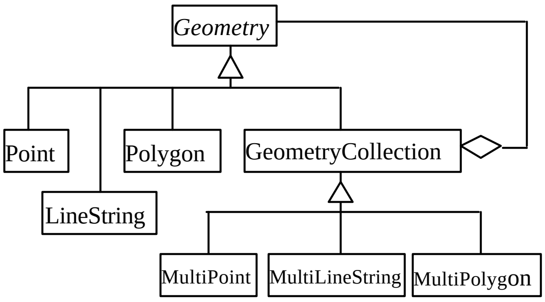


Figure 4? Geometry Object Model

The base Geometry class has subclasses for Point, LineString, Polygon and GeometryCollection. The class of GeometryCollection also has three subclasses for MultiPoint, MultiLineString and MultiPolygon. A GeometryCollection object could be composite of one or many concrete geometry objects. This object model effectively harnesses the polymorphism and design patterns[7] of object-oriented technology and leads to succinct code without losing functionalities, and system expandability has also been improved.

Examples of GSQL
GSQL has the same syntax as SQL in relational database. Besides that, you can also use geometry data type and object such as POINT, 'POINT (10,10)', ‘LINESTRING (10 10, 20 20, 30 40)', etc., in GSQL statement. GSQL can be used as Data Definition Language (DDL) and Data Manipulation Language (DML), which will be described in detail in the following subsections.

Data Definition Language (DDL): You can use DDL to create/drop table, index, and view in your spatial database (SDB). For example, if you want to create a hotel layer in your map, you can create a corresponding feature table in your SDB using following GSQL statement.

CREATE TALBE Hotel (id INTEGER PRIMARY KEY, name CHAR (32), position POINT);

Where position is of the geometry type POINT that is the built-in data type of GSQL.

The following statement can be used to remove Hotel table created by the above statement.

DROP TABLE Hotel;

Data Manipulation Language (DML): DML can be used to insert, update, and delete data in your SDB. For example, you can use the following statement to insert hotel information in the Hotel table.

INSERT INTO Hotel
VALUES (1, "Friendship Hotel", "POINT (24562.5 456221)")

Following statement can be used to retrieve all cities that are crossed by the Yangtze River.

```
SELECT city.*
FROM city, river
WHERE INTERSECTS(city.location,river.location)
        AND river.name = 'Yangtze River';
```

In the statement above, INTERSECTS is a spatial predicate, and both city.location and river.location are of the geometry type POLYGON.

Assume railway is a table for storing national railways information including railway ID, name, and route that is geometry type LineString. You can retrieve every railway name and its length by following GSQL statement.

SELECT railway.name,length(railway.route)
FROM railway

4. CONCLUSIONS AND FUTURE WORK

By introducing spatial data types, geometry object model, and spatial operations and relationships predicates conformed to the OpenGIS Simple Features Specification for SQL, a pocket spatial database prototype has been built on Compaq iPAQ H3760 with Windows CE using Microsoft eMbedded Visual C++ 3.0 as a development tool. This prototype incorporates some techniques that benefit the performance in a lightweight Pocket PC platform. These techniques can also be used to build distributed GIS applications.

Since we take ODBC as the spatial database access interface, this prototype in Pocket PC could be reconstructed to access distributed spatial database on demand.

Considering the hardware limitation of Pocket PC and the performance of ODBC, further research work will focus on high performance database access interface for specific databases such as Microsoft SQL Server CE and Oracle Lite for Windows CE. We also found out that a thick distributed heterogeneous spatial data source integration server could provide better performance for the thin client pocket computers in access of distributed spatial data sources including RDBMS and file data (i.e. shapefile). This spatial data integration technology will base on XML/GML/Web Services technology, which can pass through firewall protection, whereas thin client applications in Pocket PC only provide map request in XML/SOAP and map rendering capability based on the results returned from the spatial data sources integration server. Data caching, pre-fetching and compression techniques could be used to improve GIS performance in the thin client of Pocket PC.

5. REFERENCES

[1] Edward P.F. Chan and Rupert Zhu, " QL/G: A Query Language for Geometric Databases", University of Waterloo. Proceeding of the First International Conference on GIS in Urban Regional and Environmental Planning, Samos, Greece, April 1996, pp. 271-286.

[2] Max J. Egenhofer, Spatial SQL: A Query and Presentation Language. IEEE Transactions on Knowledge and Data Engineering, 1994, Vol.6 No.1, pp.86-95.

[3] Open GIS Consortium, Inc., " OpenGIS. Implementation Specification #99-049, OpenGIS. Simple Features Specification For SQL, Revision 1.1", 2001, http://www.opengis.org/techno/specs/99-049.pdf.

[4] WANG Feng, SHA Jichang, CHEN Huowang, YANG Shuqiang, GeoSQL: A Spatial Query Language of Object-oriented GIS. Proceedings of the 2nd International Workshop on Computer and Information Technologies CSIT'2000.

[5] Shashi Shekhar, Sanjay Chawla, Spatial Databases: A Tour, Prentice Hall, Inc. 2003.

[6] Lixing Wu, Wenzhong Shi, Principles and Algorithm of GIS, Science Press, Oct. 2003.

[7] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides, Design Patterns: Elements of Reusable Object-oriented Software. Addison Wesley Longman, Inc. 1995

[8] Len Bass, Paul Clements, Rick Kazman, Software Architecture in Practice, Second Edition (English reprint edition), Tsinghua University Press, Aug. 2003.

Guobao Yu is a Senior Engineer and Ph. D candidate in the Beijing University of Technology. He got his bachelor's degree from Huazhong University of Science and Technology in 1988 and master's degree from Chinese Academy of Sciences in 1991. He has published over 10 papers and books. Current research interests are Object-oriented technology, Database and GIS.

Husheng Liao is a Professor and doctoral supervisor in the Beijing University of Technology. He graduated from Tsinghua University and got his master's degree in 1981. He has published over 40 papers. His research interests are in Object-oriented technology, GIS and software automation

Research and Implementation of Distributed Data Dissemination

Jing Feng Kong Yi Chunhui Fan Weijun Ma
Meteorology Institute, PLA university of Sci. & Tech,
Nanjing, Jiangsu 211101, China
Email: jfeng@seu.edu.cn **Tel:** +86-25-52644114

ABSTRACT

In this paper we analyzed the classes and actions of data dissemination, and proposed a distributed data disseminating mechanism that is oriented distributed database applications on networks. This approach provided an integrated data transport between dissemination brokers based on distributed control and management, and its components can be composed to meet different usage environments. As an example, we described the prototype system based on above principle.

Keywords: dissemination, distributed system, broker

1. INTRODUCTION

With the deepen development and prevalence of the application on computer networks, large computing prefer distributed environment to centralized computer. Therefore, pervasive computing becomes a hot spot, which can provide a universal network computing plane and abroad transaction processing capability. The data distribution with a certain constraint is the basic condition to realize distributed processing.

Existing data dissemination can be classified by the direction of data flow, architecture layer and communication mode. According to the direction of data flow, one is pushing technology by which the data was pushed from servers to subscribers [1], and the other is collecting technology by which the data was collected by servers from designated clients [2]. According to the architecture layer, one is based on layer 3 and 4 of OSI/RM[3,4], and the other makes use of the distributed application plane above layer 4 [5,6]. According to the communication mode, there are two modes of data dissemination such as synchronous and asynchronous.

In this paper we focused on how to make use of these dissemination technologies to construct a distributed data dissemination system that can meet the networked and layered transact processing. Meanwhile we proposed a management mechanism for distributed data dissemination, and give its implementation of relevant prototype system.

The rest parts of the paper was arranged as follows: section two, type and action of data dissemination, analyzed and induced the main data dissemination technologies and applications; section three, design of distributed data dissemination system, analyzed the requirements of data dissemination oriented networked database applications, designed a distributed data dissemination system which can integrate multiple dissemination means and provide different combinations depending on different authorization, and proposed relevant management mechanism; section four, implementation of the prototype system, described the construction and run time environment of the prototype. Finally, we concluded the whole article in section five.

2. TYPE AND ACTION OF DATA ISSEMINATION

From the development of computer network itself and its application, the early automatic data delivery indicated the communication between network protocols in most of instance, say, messages or notifications, whereas it would rather be used in content-oriented delivery in the present, which refers to the up layers of computer networks from layer 4 to layer 7. While we discuss the technologies of lower layers (e.g. from layer 1 to layer 3), we pay attention to data framing, switching, forwarding and routing. Now we have to move our eyes upside to think about end-to-end transportation spanning long distance and carrying mass data with quality of service such as image, audio and video. Some persons considered the networks that serve for content transportation as content-network [7]. It means that an additional layer, content layer, is over the top of OSI protocol stacks, which makes traditional networks with the 7 layers architecture can provider more abundant service. These novel services are developed based on the protocols such as HTTP, RTSP (Real-Time Streaming Protocol) etc. but not traditional applications depending upon FTP.

From the development of distributed computing, it originally was defined that a big computing problem can be divided into several small parts and distributed to multiple processors at a limed environment (e.g., within a computing center) and then composed the results to a solution. The present distributed computing already try to deploy on the Internet, which can make use of all available computing resources in the world to process the big computing problems with millions of raw data[8]. Therefore, distributed computing environment is developing towards supporting grid computing, which must provide data protecting, access controlling, resource discovery and locating, directory service and scaleable user organizing model. The key technologies include the current most important computing areas: security, WWW and distributed object management.

Almost all of the above transactions need data dissemination. Generally the actions of data dissemination include following three aspects:

(1) Expanded data replication capability, e.g., universal file transportation, e-mail, downloads etc;
(2) Data collection and distribution, e.g., weather forecast, stock exchange etc;
(3) Data consistency management, e.g., distance education, resource data management etc.

Different data delivery technologies are for different goals. For data consistency management of distributed transaction, the best way is synchronous data dissemination due to real time data update. If the applications are time sensible, they demand a tight coupling data dissemination to realize safe and reliable data exchanging in time.

3. THE DESIGN OF DATA DISSEMINATION

With the expanding of networking area, the connected machines increase in geometric lever, which makes the centralized database become not suit to the huge amounts of data. In order to not only protect the former investment but also adapt to new requirements, taking distributed technology is a good solution. For data accessing, is distributed itself within enterprises, which indicates the different data orientation. For instance, the leaderships pay attention to statistic and abstract data mostly, and the common persons are interest in classified and detailed data. According to this characteristic, database can be adjusted localized, standardized and layered.

3.1 The Thought of Design

The data dissemination system designed by us can provide either synchronous transport service or asynchronous transport service. It consists of three classes of components, i.e., dissemination processing, system maintenance, and log query.
The system maintenance provides a set of tools for creating and maintaining data dictionaries and temporary information library.

The data dictionaries include three types, i.e., enterprises' attribute, relative applications attribute, and user authorization. The temporary library can be used to store task log, file description, and so on.

The log query can let user inquire about the states of every dissemination task neither sending out nor receiving in.
Dissemination processing is the core of the whole system, which includes the following modules.

(1) data extracting
(2) data loading
(3) data processing (e.g., associating, encrypting/ decrypting, compressing/ decompressing)
(4) data transfer
(5) data auditing
(6) dissemination defining
(7) rule interpreting

There is a dissemination broker server (DBS) at every network administration area. The DBS is responsible for dealing with delivery transaction, system maintenance within the area, and every client must configure the designated machine as its dissemination broker.

This kind of design has three advantages. Firstly, the whole system is scaleable, because local DBS may manage all IP addresses of clients in one area and the broker knows other DBS' IP address. Secondly, the system is reliable, because delivery transaction may independent on applications on clients, and its fault will not effect applications demanding dissemination tasks. Finally, it can be expanded to Web-based distributed query service of dissemination tasks if all DBS could provide web services. All DBSs construct a logical full-connected network shown in fig.1.

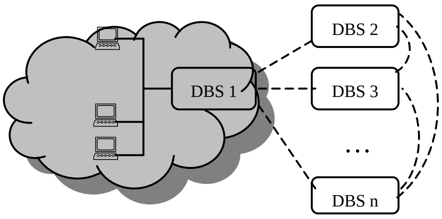


Fig.1 dissemination logical network

3.2 Definition of Metadata

In order to give a unified dissemination description and explanation, it is necessary to manage metadata that is responsible for defining the rules of dissemination operation and data extraction. The rules are described in XML, which is transparent to user and for submitting instructions of transaction between components.

The data extraction may be defined as: <Extraction>::= <indata><outdata> where, <indata> describes the type of database, base name, table name, SQL for extraction and other operation for loading, and <outdata> describes the storing place of output file. The rule of <indata> is illustrated in figure 2.

The data delivery can be defined as: <distribution>::=<srcattri><destattri><datafile> where, <srcattri> describes the sender's information including file number, sending manner, sender address and demand of the dissemination, <destattri> describes the receiver's information including receiver addresses, application name, and so on, and <datafile> describes a list of files to be sent in one encapsulation.

```
<indata type="odbc">
   <odbc dsn=" " user="" pwd=""
           product=""version="">
         <sql> SQL </sql>
          <table>table name</table>
          <operation type＝{insert, update}/>
   </odbc>
</indata>
```

Fig.2 one of the rules for dada extraction

The rule of <srcattri> is illustrated in figure 3.

```
<srcattri>
   < distribution_id type="CHAR"length="50"/>
      <    sender    type={ip,email,manual}
unit=""value=""/>
      <  demand  encrypt={Y,N}  compress={Y,N}
reply={Y,N}/>
      < file_number type="INTEGER" length="2"/>
</srcattri>
```

Fig.3 one of rules for dada delivery

For all data files encapsulated in same task, we may define different receiving manners. The control file in XML will be sent to the receivers' brokers after processed by sender broker.

4. IMPLEMENTATION OF THE PROTOTYPE SYSTEM

According to above design thought, we developed a data dissemination system, which includes two parts such as dissemination server (or broker) and client. The broker had better run on the operating system of Windows 2000, and the client may run on operating systems upwards of Windows 98. In order to invoke e-mail function, an e-mail application software must be configured on dissemination server, e.g., Outlook Express.

The function configuration on server is as follows: rule explanation and execution, task scheduling, sending pretreatment, data transmission, receiving processing, dada direction maintenance, and log management.

The function configuration on client is as follows: data extracting and loading, dada replication, task definition and submission, and query command.

In loose couple mode, the receiving side broker will notify the receiving side client indicated by sender when a new data coming. The client can either receive the data immediately or later. In tight couple mode, the receiving side client may activate a demon to receive the message sent by its broker, and replicate the data to local directory and invoke designated procedure if need when it is conscious of the notification about new dada coming.

It is important for user to identify every application in the system so that the dissemination system can distribute the data to correct applications.

5. CONCLUSIONS

The objective of information dissemination is to transfer data from producers to interested consumers. Whereas applications impose various requirements, information dissemination has several fundamental objectives that are always present: making new content available, updating existing content, and revoking obsolete content. Information dissemination strategies fall into two broad categories, pull and push[9].

Many dissemination systems is Web-based by which a centralized server provides the data to be disseminated [10]. However, for most of distributed network-based applications, the data resources are dispersed, heterogeneous, and the information to be delivered includes structuralized data and non-structuralized data. So distributed dada dissemination is a good solution for software upgrade and daily business data transmission.

We chose aperiodic push for disseminating dynamic data in the system due to the following advantages:

- The sources can initiate communication instantly if and only if the data changes or new data is available.
- The infrastructure can apply multicast instead of point-to-point communication.
- The infrastructure can combine aperiodic push with aperiodic pull to allow new clients to gather the current or historical content if the DBS provide relevant services.

There are two versions dissemination software in our prototype system for loose couple and tight couple with applications respectively. How to use them depends on whether the exchanged data between applications has stable relation. We will improve error-toleration performance of the software in future work.

6. REFERENCES

[1] Bobby Bhattacharjee, Suman Banerjee, Scalable Application-Layer Multicast for Content Distribution, http://www.cs.umd.edu/projects/nice.

[2] J. Whitehead, J. Reschke, Ed.,Web Distributed Authoring and Versioning (WebDAV) Ordered Collections Protocol, Request for Comments 3648 December 2003,

[3] W Su Y Yi, S J Lee and M Gerla. On-Demand Multicast Routing Protocol (ODMRP). Internet-Draft ” draft-ietfmanet-odmrp-04.txt”, November 2002.

[4] H. Schulzrinne，A. Rao，R. Lanphier， Real Time Streaming Protocol (RTSP)，Request for Comments: 2326，April 1998

[5] LinkPro Technologies,Enterprise Data Distribution Using Real-Time File Replication. Web Site URL: http://www.linkpro.com

[6] Klaus Schmaranz, On Second Generation Distributed Component Systems, Journal of Universal Computer Science, vol. 8, no. 1 (2002), 97-116

[7] [M. Day, B. Cain, G. Tomlinson, P. Rzewski, A Model for Content Internetworking (CDI), Request for Comments: 3466, February 2003

[8] Morgan Kaufmann, The Grid: Blueprint for a New Computing Infrastructure, 1999

[9] [Gero Mühl, Andreas Ulbrich, Klaus Herrmann, and Torben Weis, Disseminating Information to Mobile Clients Using Publish–Subscribe, IEEE INTERNET COMPUTING, MAY • JUNE 2004

[10] United States Department of Agriculture, Soil Data Delivery And Distribution, Draft Requirements Statement October 30, 2001

Feng Jing is a full associate professor of the Meteorology Institute, PLA University of Science and Technology. She got her B.S.degree in computer technology at the Missile College of Airforce in 1984, and got the M.S and Ph.D degree in computer science at Southeast University in 1997 and 2000 respectively. She was a postdoctoral fellow of ARMOR project at IRISA/INRIA in France from April, 2001 to January, 2002, and then, she continued to cooperate with IRISA/INRIA and ENST-Bretagne until December, 2002. Now she is head of information network Lab in the Meteorology Institute, PLA University of Science and Technology. Her research interests include high performance network architecture, distributed computing, information integrated technologies, and network applications.

A Multi-granularity Locking Protocol Based on Ordered Sharing Locks in Engineering Databases that Supports Cooperative Design*

Chen Guoning [1] Li Taoshen [1,2] Liao Guoqiong [3]

[1] College of computer and information engineering of Guangxi University, Nanning, Guangxi, China, 530004
[2] College of Information science and Engineering of Central South University, Changsha, China, 410083
[3] College of Computer Sci. & Tech. of Huazhong University of Sci. & Tech., Wuhan, Hubei, China, 430074

ABSTRACT

The computer supported cooperative design (CSCD) is a useful technology in the field of integrity manufacture. Those traditional concurrency control mechanisms are incompetent to provide highly concurrency for cooperative design transactions. In this paper, a transaction model supporting cooperative design is proposed first. Then, in order to ensure the correctness and consistency of engineering database using the model, an advanced concurrency control mechanism combining multi-granularity locking with ordered sharing locks is presented. According to the characteristics of engineering data objects, a lock instance graph with multi-granularity for engineering database (EDB) is constructed. For supporting cooperative read and modification, the compatible matrix of multi-granularity locking based on ordered sharing locks is also given. In the end, we have proved that any concurrent execution is serializable if it obeys the proposed protocol and no node in the lock instance graph of EDB owns a conflicting lock explicitly or implicitly except ordered sharing locks. And an implemented algorithm of the protocol is also presented finally.

Keywords: computer supported cooperative design (CSCD), concurrency control, multi-granularity locking (MGL), ordered sharing locks

1. INTRODUCTION

The computer supported cooperative design (CSCD) has captured more and more attentions of many researchers, for it is really a useful technology in the field of integrity manufacture. It provides a practical way to resolve the problems confronted by most manufacturers during product design, e.g., *cooperative read* and *cooperative modification*. These problems are derived from the more and more demands to the quality and function of products from market.

To maintain the correction and reliability, an efficient CSCD system needs a practical and effective mechanism to coordinate the accesses to the cooperative objects during the execution of cooperative activities. Thus, a highly performed concurrency control mechanism is needed here. Those traditional concurrency control mechanisms are incompetent for the task above [1], so new mechanism should be developed. Last decades, there are many researches on the subject, and many achievements have been attained as well.
TRANSCOOP [2, 3] project, executed by ESPRIT-III LTR in Europe, is a most comprehensive research on cooperative transaction. It builds a CoAct transaction model and proposes

a concurrency control mechanism named *history merging*. The history merging can coordinate the operations performed by different cooperative transactions on the same cooperative object, merging them as they performed by a single transaction, and finally makes a single consistent version of the object. But the mechanism should be used in the CoAct model, which may limit its widely application.

The *ordered nested locking mechanism* [4] used in SE-EDBMSII developed by Southeast University (China) has been proposed to support cooperation. It considers the nested structure of transaction and combines it with the ordered sharing locks [5, 6], therefore it may render partly support to the cooperation. But it does not consider the *characteristics of engineering design objects*, so it still does not reach the goal of highly concurrency.

Other concurrency control mechanisms for EDB have also been proposed [7], including the improved optimistic method [8]. But all of them are ad-hoc, and inability to resolve the problems we meet. All of these drive us to make an advanced research on the subject, and propose a concurrency control mechanism called multi-granularity locking mechanism based on ordered sharing locks. It is developed by combining multi-granularity locking with ordered sharing locks, so that it can provide the support to cooperation design and has a highly concurrency as well.

2. TRANSACTION MODEL SUPPORTING COOPERATIVE DESIGN

A distributed engineering database system (DEDBS) based on extended Client/Server(C/S) structure is proposed in [9]. The system consists of three levels: the shared server level, the project level and the designer level, which store public DB, project DB and private DB respectively. Only the data objects in the public database can be shared by any transaction. The data objects in the project database can be shared by the transactions belonged to the same project, while the private database can only be accessed by the designer transaction itself. Hence, the engineering design transactions in the system can only provide cooperation above the project level. However, it is not sufficient for cooperation requirement of the design transactions in the designer level, e.g., a designer may want to read the updated but uncommitted data objects in private database of another designer.

*This research is support in part by the Natural Science Foundation of China (No. 59868001) and the Research Foundation of Guangxi University (No. X022089).

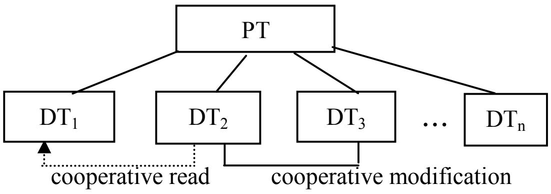


Fig 1. Transaction model supporting cooperative design

A transaction model supporting cooperative design is shown in figure 1. There are two types of transactions in the model: *project transaction PT* and *designer transactions DTs*. For the cooperative design transactions, it should have following characteristics:

- A PT is completed by two or more DTs cooperatively. (e.g., PT consists of DT$_1$, DT$_2$, ……, DT$_n$)
- Cooperative read——some transactions may read the uncommitted results belonged to other transactions. (e.g., DT$_2$ reads some data from DT$_1$)
- Cooperative modification——some transactions may modify the same design object cooperatively. (e.g., DT$_2$ and DT$_3$ modify the same object cooperatively)

However, the isolation property of transactions has been violated because of the cooperation among designer transactions in the suggested model. Therefore, in order to maintain the correctness and consistency of database using the model, we also have following constraints:

Definition 1*. If DT$_j$ has read or updated concurrently the uncommitted value of the same data object of DT$_i$, then DT$_j$ can't commit until DT$_i$ has committed. (Delay commit rule.)*
Definition 2*. If DT$_j$ has read or updated concurrently the uncommitted value of the same data object of DT$_i$, then DT$_i$ can't read or update any object updated by DT$_j$. (Avoiding waiting for each other rule.)*
Definition 3*. If DT$_j$ has read or updated concurrently the uncommitted value of the same data object of DT$_i$, then DT$_i$ can't update the object again. (Avoiding reading dirty data rule.)*

In fact, it is not necessary for definition 3. It is possible to adopt other mechanisms to avoid reading dirty data. For example, a method called savepoint [10] is a good way to implement this. Before a transaction DT$_j$ read the uncommitted value of certain data object from another transaction DT$_i$, DT$_j$ should make a savepoint first. Once DT$_i$ updates the data object again, it must inform the new value to all transactions reading the data object from DT$_i$ before. When DT$_j$ receives the notification, it should rollback to the corresponding savepoint and redoes the design according to the new value from DT$_i$. It is shown that the savepoint method could make the engineering design more flexible. For simplicity, we assume that the requirement in definition 3 is always satisfied in the following discussion of the paper, i.e., a transaction can't update the data object again if it has been read or updated by other transactions.

Obviously, the basic Two-Phase Lock (2PL) protocol can't support cooperative design, for it forbids a transaction read any data before they are committed to the project database. Therefore, it is necessary to development an advanced concurrency control protocol to help this type of transaction.

3. MULTI-GRANULARITY LOCKING PROTOCOL BASED ON ORDERED SHARING LOCKS

3.1 Preliminary

3.1.1 Hierarchical Characteristic of Engineering Design Objects

Generally, an engineering design object may consist of many parts, which can each be constituted by many smaller parts, as well. Figure 2 illustrates the nested hierarchical structure of a product. As figure 2 illustrates, a product can be constituted by a set of components, and each component can be also composed by a set of hardware.

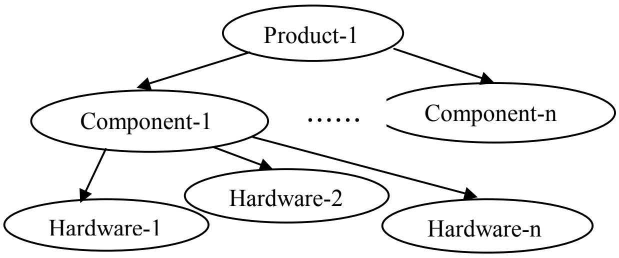


Fig 2. The Hierarchical Structure of an Engineering Design Object

3.1.2 Multi-Granularity Locking (MGL)

Reference [11] proposed a locking mechanism called multi-granularity locking. It is developed by considering the granularity of data items presented by *lock type graph*, which is used to represent hierarchical relationships between locks of different granularity. Each edge in the graph connects a data type of coarser granularity to a finer granular one. It requires that the objects in the graph can't be added any lock until all of its ancestors have been locked by corresponding locks. These corresponding locks added on ancestors are called intended locks. The mechanism can make the locking range of a transaction closely constrained in the granularity it needs or somewhat wider, which may reach a highly concurrency of the execution of design transactions.

3.1.3 Ordered Sharing Locks

A locking mode named ordered sharing is studied in the research of reference [5,6]. It allows different transactions to concurrently hold locks on the same data object, even for conflicting operations. This mode of locking allows constrained sharing in the sense that those operations of different transactions are required to be executed in the same order as the locking order. For example, it allows a transaction T$_i$ to read the modification of another transaction T$_j$ before T$_j$ has committed. But it requires the write lock of T$_j$ is obtained first. In order to ensure correctness, two rules must be obeyed in the locking mode according to following definitions:

Definition 4. *In a schedule of a set of transactions, for any two concurrent and conflicting operations on the same data object x, $p_i(x) \in T_i$ and $q_j(x) \in T_j$, if the lock operation $pl_i(x)$ of $p_i(x)$ precedes the lock operation $ql_j(x)$ of $q_j(x)$, then $p_i(x)$ must be executed before $q_j(x)$, denoted $p_i(x) < q_j(x)$. At this time it is said that the locking operation of $p_i(x)$ is ordered sharing with the lock operation of $q_j(x)$, noted $pl_i(x) \oplus ql_j(x)$. (Ordered sharing locks acquisition rule)*

Definition 5. *If a transaction T_j acquires the lock of data object x from transaction T_i with ordered sharing locks, then T_j may not release any lock until T_i has released its locks. At this time, T_j is said to be waiting for T_i orderly. (Order sharing locks releasing rule)*

The ordered sharing locks provide the supports to the engineering cooperative design transactions. The first rule makes it possible to design cooperatively. And the later rule can ensure the requirement of definition 1. If all transactions lock data objects according to 2PL, the latter rule can be described as: T_j can't commit until T_i has committed if T_j has read or updated concurrently the uncommitted value of the same data object of T_i.

3.2 Multi-Granularity Locking Mechanism Based on Ordered Sharing Locks

3.2.1 Lock Instance Graph for EDB

A set of data items that is structured according to a lock type graph is called a lock instance graph [11]. According to figure 2, it is natural to structure the lock instance graph by a tree logically (see figure 3) for the set of data items in EDB. That is, a lock on a coarse granule x explicitly locks x and implicitly locks all of x's descendants, which are finer granules "contained in" x. For example, a read lock on a product P_1 implicitly locks the components (C_1, C_2 and C_3) and hardwares (H_1, H_2, H_3, H_4, H_5 and H_6) belonged to the product with read lock.

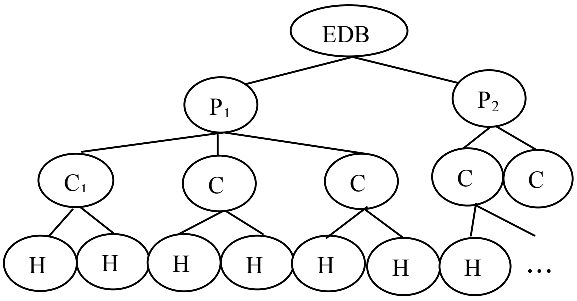


Fig 3. The lock instance graph of EDB

3.2.2 Compatible Matrix for Multi-Granularity Locking Based on Ordered Sharing Locks

To overcome the defects of the basic locks mechanism, we extend the simple read/write locks to more resorted types, including a browse lock, a read lock, a write lock, an exclusive lock and their corresponding intended locks. The definition of each lock and their working approaches are given as follows:

Browse lock (shorten in B) — use to access design data object by a read operation with browse way.
Read lock (shorten in R) — use to access design data object by a read operation.
Write lock (shorten in W) — use to access design data object by a write operation.
Exclusive lock (shorten in X) — use to access design data object by a write operation exclusively.
Intend Browse lock (shorten in IB)—use to lock the external layer node of the data object locked by B-lock.
Intend Read lock (shorten in IR)—use to lock the external layer node of the data object locked by R-lock.
Intend Write lock (shorten in IW)—use to lock the external layer node of the data object locked by W-lock.
Intend Exclusive lock (shorten in IX)—use to lock the external layer node of the data object locked by X-lock.
Given the requirement of the cooperative design and the lock types defined above, Table 1 gives the compatible matrix of those locks.

Table 1. The compatible matrix of multi-granularity locking based on ordered sharing locks

	B	R	W	X	IB	IR	IW	IX
B	+	+	+	—	+	+	+	—
R	+	+	⊕	—	+	+	⊕	—
W	+	—	⊕	—	+	-	⊕	—
X	—	—	—	—	—	—	—	—
IB	+	+	+	—	+	+	+	+
IR	+	+	⊕	—	+	+	+	+
IW	+	—	⊕	—	+	+	+	+
IX	-	—	—	—	+	+	+	+

+—compatible; —incompatible ⊕—ordered sharing
Lock holder T_i

Based on the compatible matrix above, the extended multi-granularity locking model can provide following ways to access data in CAD system.

(1) To access object exclusively, X-lock can be used. For it is exclusive with any other locks.
(2) To update object cooperating with others, W-lock can be used. For W-lock can be ordered sharing with R-lock and W-lock, the transaction who holds W-lock of an object allows its cooperative transactions to access this object with R-lock and W-lock through ordered sharing rules.
(3) To read objects and forbid other transactions to update the object, R-lock can be used.
(4) To browse object, B-lock can be used. The difference between the R-lock and B-lock is the former always hopes to get accurate results, but it is not necessary for the later.

3.2.3 Multi-Granularity Locking Protocol Based on Ordered Sharing Locks

Given the hierarchical characteristic of the design objects and the requirement of cooperation, a locking mechanism, combining with multi-granularity locking and ordered sharing locks, has been proposed. For any data item *x* in the lock instance graph as figure 3, the description of the protocol is giving in the form of following rules:

(1) A lock operation noted $al_i(x)$ must be executed before any operation noted $a_i(x)$. Where *i* stands for the transaction T_i, *a* stands for any type of operations that are allowed on the object and *x* stands for the object being operated.
(2) The operation $al_i(x)$ can be executed only if all ancestors of the object *x* are locked by the intended lock corresponding to al_i successfully. The order of the locking is from the top down to *x*.
(3) When there is no any lock al_j incompatible with al_i on the object *x* (where $j \neq i$), $al_i(x)$ can be executed. (i.e., Object *x* has been add a lock al_i.). Otherwise, it should wait till al_j is released.
(4) If object *x* has been locked by a lock al_j and al_j is ordered sharing with al_i according to the compatible matrix, and transaction T_j has not added any ordered sharing lock on the object owned by T_i, $al_i(x)$ can be executed. (i.e., Object *x* has been added a lock al_i.). Else it should wait until al_j is

released.

(5) Any lock al_i on object x can be released only if transaction T_i is committed or aborted. The order of releasing is exactly reversing to the order of locking.

(6) If transaction T_j accesses (read or update) the result of another transaction T_i through the ordered sharing locks, T_j can not release any lock until T_i successfully releases its locks.

(7) Once transaction T_i has released a lock it ever held, no new lock can be applied by T_i.

Rule (1) is a common rule in any lock mechanism. Rule (2), (3) and (4) are the acquiring lock rules in the protocol. Rule (2) is derived from the multi-granularity locking. It can prevent other transactions from adding an incompatible lock on any ancestor of the locked object. Rule (3) has the same function as the basic lock mechanism. It performs the locking operation on an object. Rule (4) is derived from the order sharing locks (definition 4). It provides partly cooperation for transactions and prevents the transactions from waiting for each other due to ordered sharing (definition 2). Rule (5), (6) and (7) are the releasing lock rules in the protocol. Rule (5) regulates the releasing time and releasing order. It can ensure the correctness of the releasing operation and avoid inconsistency in EDB. Rule (6) is derived from the ordered sharing locking (definition 5). It ensures that the ordered sharing locks can be released in correct order and avoid 'dirty data' in EDB. Rule (7) is derived from 2PL protocol, it is important to ensure the serializibility of concurrent transactions.

4. CORRECTNESS OF MULTI-GRANULARITY LOCKING PROTOCOL BASED ON ORDERED SHARING LOCKS

4.1 Preliminary

The property called serializability is a widely used criterion for ensuring the correctness of concurrent execution of transactions. And the serializability of a concurrent execution of transactions is determined by a graph derived from the execution called a serialization graph (SG). So, before giving our proof, some definitions and theorem are giving as follows first:

Definition 6. *If two operations o and o' access the same data item and at least one of them wrote to it, it is said that o conflict with o', and vice versa.*

Definition 7. *If an operation of transaction T conflicts with an operation of another transaction T', it is said that T conflicts with T', and vice versa.*

Definition 8. *For a concurrent execution H, the serialization graph SG(H) = <V, E> is a directed graph such that:*

V= {T_i | T_i is a transaction that is already started in H};

E = {($T_i \rightarrow T_j$) | there are some conflicting operations $p_i(x) \in T_i$ and $q_j(x) \in T_j$ such that $p_i(x)$ is scheduled before $q_j(x)$ where $T_i, T_j \in V$ and $T_i \neq T_j$}.

Theorem 1. *An Execution H is serializable iff SG(H) is acyclic.*

4.2 Correctness of the Protocol

We will prove the correctness of the protocol from two aspects. First, we will prove that any concurrent execution is serializable if it obeys the protocol. Then, we will also prove that no node in the lock instance graph of EDB owns a conflicting lock explicitly or implicitly except ordered sharing locks.

Lemma 1. *If the executions of a set of transactions obey the multi-granularity locking protocol based on ordered sharing lock, and a path ($T_1 \rightarrow T_2 \rightarrow \ldots\ldots \rightarrow T_{n-1} \rightarrow T_n$) exists in the SG(H) at the same time, then the following is also true:*
A releasing lock operation $u_1 \in T_1$ exists, and for any transaction T_i ($1 < i \leq n$), exists $\forall\, u_i \in T_i$, $u_1 < u_i$.

Proof: Induction method is used here.
Assume that i stands for the number of the transactions in the path ($T_1 \rightarrow T_2 \rightarrow \ldots\ldots$)
When i=1, the conclusion is held.
When i=2:
$\because$ in SG(H), $T_1 \rightarrow T_2$.
$\therefore$ There must exist at least one conflicting operation pair $p_1(x)$ and $q_2(y)$. And it is held that $p_1(x)<q_2(y)$. Where $p_1(x) \in T_1$,$q_2(y) \in T_2$, x and y are operated objects. '$p_1(x)< q_2(y)$' means that $p_1(x)$ is performed earlier than $q_2(y)$. There are three cases.
Object x is an ancestor of object y.
According to rule (2) in the protocol (in the remainder of the paper, 'rule' means the rule in the locking protocol except for specially statement), it must be held that $pl_1(x)< qil_2(x)< ql_2(y)$. Where $pl_1(x)$ and $ql_2(y)$ are the corresponding locking operations of $p_1(x)$ and $q_2(y)$, respectively. $qil_2(x)$ stands for the intended locking operation before $ql_2(y)$.There are two sub-cases.
(1a) $qil_2(x)$ is incompatible with $pl_1(x)$.
$\because$ According to rule (5) and (3), it must be held that $pu_1(x)< qil_2(x)$ and $qil_2(x)< ql_2(y)< qu_2(y)$.
$\therefore \exists\, u_1 \wedge \forall\, u_2 \Longrightarrow u_1<u_2$.
(1b) $qil_2(x)$ is ordered sharing with $pl_1(x)$.
According to rule (4) and rule (6), $\exists\, u_1$, for $\forall\, u_2$, it must be held that $u_1<u_2$.
(1) Object y is an ancestor of object x.
According to rule (2), it must be held that $pil_1(y)<ql_2(y)$. There are two sub-cases.
(2a) $ql_2(y)$ is incompatible with $pil_1(y)$.
$\because$ According to rule (5) and (3), it must be held that $pu_1(x)< piu_1(y)< ql_2(y)$ and $ql_2(x)< qu_2(y)$.
$\therefore \exists\, u_1 \wedge \forall\, u_2 \Longrightarrow u_1<u_2$.
(2b) $ql_2(y)$ is ordered sharing with $pil_1(y)$.
$\because$ According to rule (4), it must be held that $pil_1(y)<ql_2(y)$.
$\therefore$ According to rule (6), $\exists\, u_1 \wedge \forall\, u_2 \Longrightarrow u_1<u_2$.
(2) x and y are the same object.
According to rule (1) and (2), it must be held that $pl_1(x)< ql_2(y)$. There are two sub-cases.
(3a) $ql_2(y)$ is incompatible with $pl_1(x)$.
$\because$ According to rule (5), it must be held that $pu_1(x)< ql_2(y)< qu_2(y)$.
$\therefore \exists\, u_1 \wedge \forall\, u_2 \Longrightarrow u_1<u_2$.
(3b) $ql_2(y)$ is ordered sharing with $pl_1(x)$.
$\because$ According to rule (6), it must be held that $pu_1(y)< qu_2(y)$.
$\therefore \exists\, u_1 \wedge \forall\, u_2 \Longrightarrow u_1<u_2$.
$\therefore$ The conclusion is held when i=2.
Assume that all the cases where $i \leq n-1$, the conclusion is held.
That is, $\exists\, u_1$, for $\forall\, u_{n-1}$, it must be held that $u_1<u_{n-1}$.
$\because T_{n-1} \rightarrow T_n$, uses the similar method above, $\exists\, u_{n-1}$. for $\forall\, u_n$ it must be held that $u_{n-1}<u_n$.
$\therefore \exists\, u_1 \wedge \forall\, u_n \Longrightarrow u_1<u_n$.
Hence, the lemma is proven.

Theorem 2. *Every execution H is serializable, if the execution obeys the multi-granularity locking protocol based on ordered sharing locks.*

Proof: Assume that there is a cyclic in the SG(H). Not losing the generality, suppose there is a path ($T_1 \rightarrow T_2 \rightarrow \ldots\ldots \rightarrow T_{n-1} \rightarrow T_n \rightarrow T_1$) in the SG(H).

∵According to lemma 1, in the path $T_1 \rightarrow T_2 \rightarrow \ldots\ldots \rightarrow T_{n-1} \rightarrow T_n$, $\exists u'_1 \in T_1$, for $\forall u_n \in T_n$ it must be held that $u'_1 < u_n$. And at the same time, in the path $T_n \rightarrow T_1$, $\exists u'_n \in T_n$, for $\forall u_1 \in T_1$, it must be held that $u'_n < u_1$.

It shows that $u_1' \notin \{u_1\}$, a contradiction arises.

∴The assumption above is impossible.

Hence, according to theorem 1, theorem 2 is proven.

Theorem 3. *Suppose all transactions obey the multi-granularity locking protocol based on ordered sharing locks with respect to the given graph in figure 3. If a transaction owns an explicit or implicit lock on a node in the graph, then no other transaction owns a conflicting explicit or implicit lock on that node except ordered sharing locks.*

Proof: It is enough to prove the theorem for leaf nodes. For, if two transactions held conflicting (explicit or implicit) locks on a non-leaf nodes *x*, they would be holding conflicting (implicit) locks on all descendants and, in particular, all leaf descendants of *x*. Suppose that transactions T_i and T_j own conflicting locks on leaf *x*., and T_i adds locks earlier than T_j. There are following cases:

	Transaction T_i	*Transaction T_j*
1.	implicit B lock	implicit X lock
2.	implicit B lock	explicit X lock
3.	explicit B lock	implicit X lock
4.	explicit B lock	explicit X lock
5.	implicit R lock	implicit X lock
6.	implicit R lock	explicit X lock
7.	explicit R lock	implicit X lock
8.	explicit R lock	explicit X lock
9.	implicit W lock	implicit X lock
10.	implicit W lock	explicit X lock
11.	explicit W lock	implicit X lock
12.	explicit W lock	explicit X lock
13.	implicit X lock	implicit X lock
14.	implicit X lock	explicit X lock
15.	explicit X lock	explicit X lock
16.	implicit R lock	implicit W lock
17.	implicit R lock	explicit W lock
18.	explicit R lock	implicit W lock
19.	explicit R lock	explicit W lock
20.	implicit W lock	implicit W lock
21.	explicit W lock	implicit W lock
22.	explicit W lock	explicit W lock

Case 1. By rule 1 and 2, T_i owns $bl_i(y)$ for some ancestor *y* of *x*, and T_j must own $xl_j(y')$ for some ancestor *y'* of *x*. There are three subcases: (a) *y*=*y'*, (b) *y* is an ancestor of *y'*, (c) *y'* is an ancestor of *y*. Case (a) is impossible, because according to the compatible matrix shown in Table 1, T_i and T_j are holding conflicting browse and exclusive locks respectively on *y*=*y'*. Case (b) is impossible because T_j must own $ixl_j(y)$ according to rule 2, which conflicts with $bl_i(y)$. Case (c) is impossible because T_i must own $ibl_i(y')$, which conflicts with $xl_j(y')$. Thus, the assumed conflicts are impossible.

Case 2. By rule 1 and 2, T_i owns $bl_i(y)$ for some ancestor *y* of *x*. By rule 2, T_j must own ix lock for every ancestor of *x*. In particular, it owns $ixl_j(y)$, which is impossible because $ixl_j(y)$ is conflicting with $bl_i(y)$.

Cases 4,8,12 and 15 are obviously impossible according to the compatible matrix. Cases 5,9 and 13 follow the same argument as case 1. Cases 3,6,7,10,11 and 14 follow the same argument as case 2.

Case 16. If T_i owns the lock earlier than T_j, then it follows the argument of case 1. Else, if T_j owns the lock earlier than T_i, by rule 1 and 2, T_i owns $rl_i(y)$ for some ancestor *y* of *x*, and T_j must own $wl_i(y')$ for some ancestor *y'* of *x*. There are also three subcases: (a) *y*=*y'*, (b) *y* is an ancestor of *y'*, (c) *y'* is an ancestor of *y*. In case (a), T_i and T_j are holding ordered sharing read and write locks respectively on *y*=*y'*. In case (b), T_j must own $iwl_j(y)$ according to rule 2, which is ordered sharing with $rl_i(y)$. In case (c), T_i must own $irl_i(y')$, which is ordered sharing with $wl_i(y')$. Thus, T_i and T_j can not own conflicting locks on *x* except ordered sharing locks.

Case 17. If T_i owns the lock earlier than T_j, then it follows the argument of case 2. Else, if T_j owns the lock earlier than T_i, by rule 2, T_i owns $rl_i(y)$ for some ancestor *y* of *x*, and T_j must own iw lock for every ancestor of *x*. In particular, it owns $iwl_j(y)$, which is ordered sharing with $rl_i(y)$.

Case 18 follows the similar argument like case 17.

Case 19. If T_i owns the lock earlier than T_j, it is obviously impossible according to the compatible matrix. If T_j owns the lock earlier than T_i, T_i is holding $rl_i(x)$ which is ordered sharing with $wl_j(x)$ held by T_j.

Case 20 follows the same argument where T_j owns the lock earlier than T_i in case 16. Case 21 follows the same argument where T_j owns the lock earlier than T_i in case 17.

In case 22, T_i and T_j are holding ordered sharing locks $wl_i(x)$ and $wl_j(x)$ on *x* respectively.

Hence, the theorem is proven.

5. AN IMPLEMENTATION OF MULTI-GRANULARITY LOCKING PROTOCOL BASED ON ORDERED SHARING LOCKS

In this section, we show an implementation algorithm of the multi-granularity locking protocol based on ordered sharing locks. It is implemented in extended Client/Server environment [9] and programmed in C++ language. For space limit, we only introduce the algorithm of the lock manager server (LM Server) here.

In the LM Server algorithm, two kinds of lists and two kinds of tables are built to implement the multi-granularity locking protocol based on ordered sharing locks. The object relation table (ORT) is used to store the parent-child relations between objects. The possessed lock list (PLL) of an object is used to store the obtained locks on corresponding object and the transaction owning the lock, each object may have their own PLL. The ordered sharing transactions table (OSTT) is used to store the ordered sharing relations between transactions. OSTT is the set of the ordered sharing transactions pares. One pare is stored as a record, including one waiting transaction and one be-waited transaction, which means that the waiting transaction can't commit until the be-waited transaction commit (rule 6). The waiting transactions list (WTL) is used to store the transactions waiting for the lock of a specific object, thus each object may have a WTL.

The LM Server algorithm is mainly consists of two parts, the locking process and the unlocking process. The descriptions of the two parts are given as follows respectively.

Locking process

```
    When receive a lock request message,
    Decompose the message, obtain the transaction ID, requested object, operation type and other information
    Open the ORT and find all the ancestors of the object
    if (the ancestor objects exist) {
     Deal with each ancestor objects from top to down according to their parent-child relationship as follows
     Check all the locks in PLL of current object with the requested lock
  Flag1: if (compatible) {
          if (ordered sharing compatible){
              Add the current requesting transaction and transaction who own the other ordered sharing lock to the OSTT, where the current requesting transaction is the waiting transaction and the other is be-waited transaction
Add current requesting transaction and its lock to the PLL
Send lock-success message to the requested transaction
                         Wait for another request               }
          else{
Add current requesting transaction and its lock to the PLL
Send lock-success message to the requested transaction
                          Wait for another request              }
    else{
Add the current transaction to the WTL of the object
Send lock-fail message to the requesting transaction
                        Wait for another request                 }
    else Process the same steps after Flag1 above on the requested object
```

Unlocking process

```
    When receive unlock request message
    Decompose the message, obtain the transaction ID, requested object, and other information
    Open the ORT and find all the ancestors of the object
    Open the OSTT and find current requesting transaction in OSTT
   Deal with each ancestor objects from top to down according to their parent-child relationship as follows
    if (current requesting transaction exits in OSTT) {
   if (it is waiting transaction) {
          Send unlock-fail message to the requesting transaction
          Wait for another request
   }
         else {
Flag2:   Delete the corresponding record from OSTT and notify all the waiting transactions
          Remove current transaction and its lock from PLL of the requested object
          Send unlock-success message to the requesting transaction
Notify all the transactions in WTL of the object
          Wait for another request
    }
}
    else Process the same steps after Flag2
```

6. CONCLUSIONS

The computer supported cooperative design (CSCD) is a useful technology in the field of integrity manufacture. Those traditional concurrency control mechanisms are incompetent to support highly concurrency for cooperative design transactions. In this paper, according to the hierarchical structure of design objects, a new concurrency control mechanism called multi-granularity locking mechanism based on ordered sharing locks is proposed. It is developed by combining multi-granularity locking with ordered sharing locking, so that it can support cooperative design and provide a highly concurrency as well. The locking protocol is introduced and explained in the paper. And, according to the characteristic of engineering data objects, a lock instance graph with multi-granularity for engineering database (EDB) is constructed. For supporting cooperative read and modification, the compatible matrix of multi-granularity locking based on ordered sharing locks is given. In the end, we have proved that any concurrent execution is serializable if it obeys the proposed protocol and also proved that no node in the lock instance graph of EDB owns a conflicting lock explicitly or implicitly except ordered sharing locks. It shows that the locking mechanism is correct.

According to the locking mechanism and the LM Server algorithm introduced in this paper, we have performed the experiment in a transaction process system. It is developed for a cooperative design system. The system is built on the extended Client/Server structure and can provide support for team cooperation in LAN. The result of the experiment show that the protocol presented here is correct and workable. The performance will be analyzed in future reports.

7. REFERENCES

[1] Barghouti N. S. and Kaiser G. E. Concurrency Control in Advanced Database Application, ACM Computing Surveys, 1991,23 (3): 269-317.

[2] K. Aberer, S. Even, F. Faase, H, Kaijanranta, J. Klingemann, Transaction Support for Cooperative Work: An Overview of the TransCoop Project, Workshop on Extending Data Management for Cooperative Work, Darmstadt, June 6, 1997.

[3] Jürgen Wäsch, Wolfgang Klas. History Merging as a Mechanism for Concurrency Control in Cooperative Environments. Proceedings RIDE'96: Interoperability of Nontraditional Database Systems (RIDE-NDS '96), New Orleans, Louisiana, Feb. 26-27, 1996, pp. 76-85.

[4] Qi Jing, Zhang Jiaming, Zhou Boxin. A Concurrency Control Mechanism to Support Cooperative transaction in EDB. Computer Research and Development, 1998, 35(11): 987~990. (in Chinese)

[5] Agrawal D, Abbadi A. Locks with constrained sharing. In:The Proc of 9th ACM SIGACT-SIGMOD Symposium on Principles of Database Systems. Nashville, Tennessee, 1990, 85~93.

[6] Agrawal D, Abbadi A. The performance of protocols based on locks with ordered sharing. IEEE Transaction on Knowledge and Data Engineering, 1994, 6(5): 805~818.

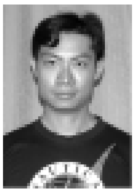
[7] Henry F.KORTH. A Model of CAD Transaction, 1985, In Proceeding of the Eleventh International Conference on Very Large Database: 25~33.

[8] Liao GuoQiong, Li TaoShen. An Optimistic Concurrency Control Method for Supporting Engineering Design Transaction, Computer Engineering, 2000, 26(7):24~25. (in Chinese)

[9] Li Taoshe, Liao Guoqiong, Chen Guoning. Transaction Management Mechanisms in Distributed Engineering Database System. Proceeding of the Seventh International Conference on Computer Aided Design and Computer Graphics, August 22-24, 2001,Kunming, China, 785~790

[10] Axel Meckenstock, Rainer Unland, Detlef Zimmer: Rolling Back in a Selective Way - an Approach to Recovery for Interactive and Long-Running Transactions, Proceedings 2nd International Baltic Workshop on DB and IS, Tallinn (Estonia), June 1996.

[11] P. A. Bernstein, V. Hadzilacos, and N. Goodman. Concurrency Control and Recovery in Database Systems. Addison Wesley, Reading, Massachusetts, 1987.



Chen Guoning is now a teacher in the College of Computer, Electronic and Information, Guangxi University. He graduated from Xi’an Jiaotong University in 1999 with specialty of information and communication engineering, and got his master degree of computer application in Guangxi University, 2002. At present, his research interests are in computer supported collaborative work, distributed processing, computer aided design.

Li Taoshen is a Full Professor and now the head of the College of Computer, Electronic and Information, Guangxi University. He graduated and got his master degree of computer application technology at College of Information science and Engineering, Central South University and now the phD candidate of the university. He is now a member of the council of Computer Society of China. His research interests are in distributed database, network security, CAD, genetic optimisations design.

A Multi-dimension Perspective to XML Databases Modeling*

Liu HongXing [1, 2]   Lu YanSheng [2]
(College of Computer Science &Technology, Wuhan University of Technology
Wuhan, Hubei 430063, P.R. China, Email: liuhx@public.wh.hb.cn**) [1]**
(College of Computer Science, HuaZhong University of Science & Technology
Wuhan, Hubei 430074, P.R. China, Email: lys@mail.hust.edu.cn**) [2]**

ABSTRACT

XML has become the standard format for representing structured and semi-structured data on the Web. To manage XML data we need to design and then use XML database systems. There is a wealth of experience for designing traditional databases; XML databases, however, are a much more recent phenomenon, raising many new requirements to the database design methodology. Although there are some researches and methods on the modeling of XML document or XML database, in order to evaluate or integrate those, we need to build a unified context. This paper presents a Four-dimension perspective to the XML database modeling, which is an extension of a Three-dimension perspective, and can be used to analyze different concepts, models and modeling methods in XML database environment.

Keywords: Database modeling, XML, Schema, XML database systems, Database design method, Multi-dimension perspective.

1. INTRODUCTION

Extensible Markup Language (XML) is fast becoming the new standard for data representation and exchange on Internet, Intranet and Extranet. With the need to process XML documents comes the need to be able to store, retrieve, and manipulate on them, these lead to the emergence of XML database systems. There are different definitions to XML databases [2]. We define an XML database (or XML document database) to be a collection of XML documents, which are managed by a system called XML DBMS. An XML database and its corresponding DBMS together constitute a XML database system. In recent years, many works have been done to the research and development on XML database systems. Currently there are three kinds of XML database systems: XML-Enabled Database (XEDB)， Native XML Database (NXDB), and Hybrid XML Database (HXDB). The differences among the three kinds, sometimes, are not very clear; for example, the Oracle 9i (V2) [13] may be labeled either XEDB or HXDB. Although having problems and being evolving rapidly, several XML DBMSs now are ready for practical use.

How to design an XML database? This is another important aspect beside XML DBMS. Database design is always the foundation and the most important aspect for the application of database. There is a wealth of experience, captured in books, methods, and tools, for designing traditional databases (especially relational database) [1], this experience has evolved over thirty years. XML databases, however, are a much more recent phenomenon, raising many new requirements to the database design methodology.

To meet different application requirements at the same time, a typical database is usually complicated in structure, so needs to be modeled by different schemas. XML has its own special characteristics and therefore adds more complications to the database modeling. Lots of researches have been done relating to the modeling of either XML documents or XML databases, but most of them are made in special ways for meeting special requirements. When putting different design concepts or models together, we may find either inconsistency in concepts or mismatches in methods. It is time to specify a general framework for XML database modeling, so the models and methods can be developed and evaluated on a consistent basis.

A four-dimension (4D) unified perspective is presented in this paper, which builds a consistent context, and can be used to describe or specify various kinds of models and the modeling approaches to XML database.

The remainder of the paper is organized as follows. Section 2 describes a “3D Perspective” to the traditional database modeling, which builds a basis for the next section. Section 3 firstly discusses XML database and the special modeling requirements, and then extends the 3D perspective to a “4D perspective”, which can be used to describe and specify the more complicated modeling to the XML database. Section 4 concludes the paper.

2. A 3D-PERSPETIVE TO THE TRADITIONAL DATABASE MODELING

Database modeling is the process to build the schemas (models) of a database for an application domain, and a schema is a representation of the database’s structure and other characteristics. In this paper we refer Database Modeling as building database’s structure, and take schema as the synonymy of structure.

To specify different kinds of schemas in different levels, people use different design methods or framework. Fig.1 is the ANSI/X3/SPARC’s Three-level database architecture [1], while Fig.2 is the three-level architecture approach which has been widely used by database designers.

The two Figs. look similar, but they view database in different ways or viewpoints. In Fig.1, the External schemas and Conceptual schema together tells that from a local or global scope we can get different views to a same database. Fig.2 shows that a database schema should have different levels of abstraction, and we may design a database through the conceptual-logical-physical levels. When putting two Figs. together we will find the inconsistencies in used terms.

* The work is supported by the Science &Technology Foundation of Wuhan University of Technology (No. XJJ2002027), and the Natural Science Foundation of China (NSFC, No. 60073057)

For example, the two “Conceptual schema” in Fig.1 and Fig.2 are different; in many cases, the “Conceptual schema” in Fig.1 may correspond to the “Logical schema” in Fig.2. Similar inconsistencies exist often in the field of data modeling.

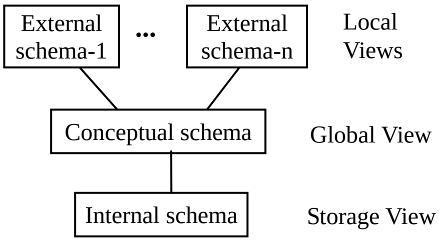


Fig.1. ANSI/X3/SPARC: Three-level DB architecture

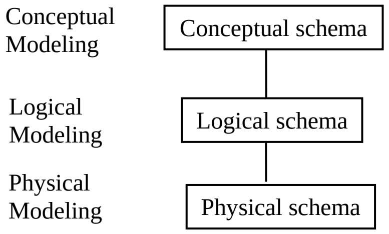


Fig.2. The three-level architecture Approach for DB design

For the purpose of discussing different database schemas and different design approaches in a unified, consistent and visual way, we present a Three-Dimension perspective as shown in Fig.3. All kinds of database schemas are put into the 3D coordinates. Each dimension represents a way to observe, design or make use of a database’s schemas, and along each axis two scales (levels) are labeled.

1) Axis X----the **abstraction dimension**: Along this axis, designers view a database in different levels of abstraction: conceptual and logical. It can be said that the idea behind Fig.2 is to design a database along X. For the purpose of the paper, another level, i.e. the physical level in Fig.2, has been omitted.

2) Axis Y----**the scope dimension**: In a typical database environment, we define a global schema, and then from which, define some sub-schemas (local schemas) by mapping, so as to support different applications.

3) Axis Z----the **Visualization dimension**: By using some design tools we can represent a schema in either a visual (graphic) or literal way.

In Fig.3 there are totally eight different kinds of schemas (i.e. the cube’s vertexes labeled by numbers). Table.1 is a summary of all kinds of schemas. Further, each edge and face on the cube does represent an approach or a view in modeling a database. In the 3D coordinates, the cube and all the elements on it，i.e. the vertexes, the edges and the faces, together constitute a unified and consistent perspective for database modeling.

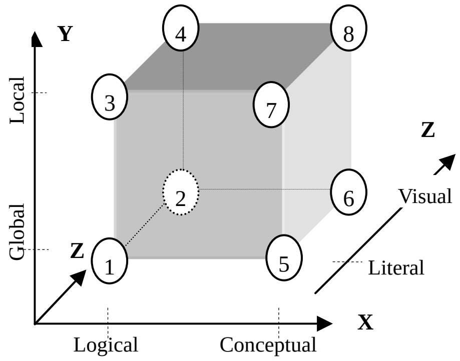


Fig.3. A 3D-Perspective to the database schemas and modeling

Table.1. A summary of all kinds of schemas

Node No.	Abstraction (X)	Scope (Y)	Visualization (Z)	Examples
1	Logical	Global	Literal	Relational DB schema specified using SQL
2	Logical	Global	Visual	PDM diagram in PowerDesigner[*]
3	Logical	Local	Literal	A part of relational DB schema specified using SQL
4	Logical	Local	Visual	A sub-diagram of PDM
5	Conceptual	Global	Literal	CDM[*] diagram in PowerDesigner specified by XMI[**]
6	Conceptual	Global	Visual	CDM diagram
7	Conceptual	Local	Literal	A sub-diagram of CDM specified by XMI
8	Conceptual	Local	Visual	A sub-diagram of CDM

* PowerDesigner [11] is an integrated modeling toolkit developed by Sybase, PDM stands for Physical Data Model, and CDM stands for Conceptual Data Model.

** XMI stands for OMG’s XML Metadata Interchange.

3. XML DATEBASE MODELING: A 4D-PERSPECTIVE

XML and its usages add new requirements to the database systems, which dedicate to the management of XML documents; also they call for some new characteristics in models and approaches for modeling the XML databases. To meet the requirements for modeling either XML documents or XML database, many researches have been done, and presented some new concepts or approaches. When putting them together there may be inconsistencies in concepts. We now present a 4D-Perspective, which can be used to integrate different concepts, models and modeling methods, towards a unified framework for the data modeling in XML database environment.

3.1 XML Document Schema and XML database Schema

A XML document has a logical structure, and a corresponding document schema may serves for describing the structure: which elements are child elements of others, the sequence in which the child elements can appear, and the number of child elements. It also defines whether an element is empty or can include data. A document may have a simple structure or a complicated, nested structure, so it is necessary to develop the methods or models to represent the document's structure or schema.

An XML database, like a general database, has its own structure, i.e. the XML database schema. Because XML database is a collection of different documents and each document may have its own schema, these necessitate that a XML database schema describe not only the relationships between the components within a documents but also the relationships across different types of documents. In some word, the schema of an XML database is the integration of all related document schema. Because XML is semi-structured, the structure of an XML database tends to be more complicated as relative to a traditional database.

Fig.4 shows the XML documents modeling, XML databases modeling, and the relations between them. Two domains are introduced in Fig. 4.

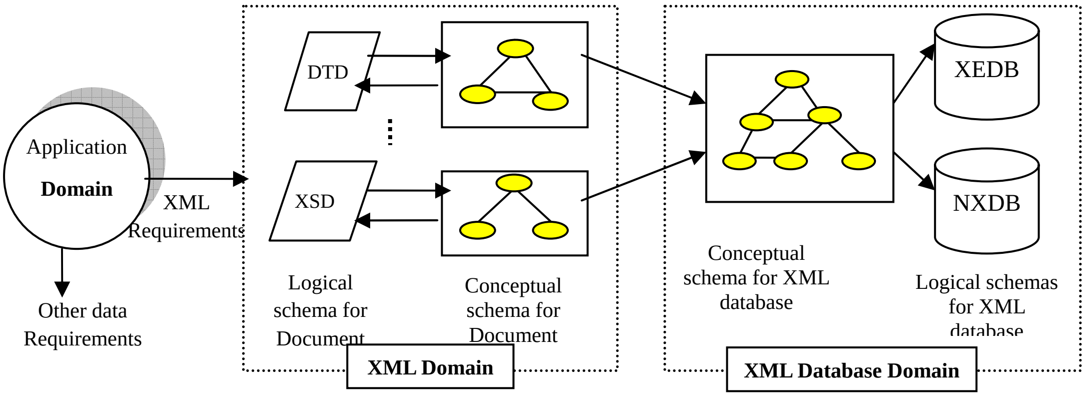


Fig. 4. XML document schemas and XML database schemas and their modeling

1) The XML Domain.

XML is original introduced for data representation and exchange, in fact, neither document instances nor document schemas need to exist as documents or be stored permanently *per se* -- they may exist as streams of bytes sent between applications. So, even if we do not consider the storage of XML documents, to design the document schemas, which itself is an important work. The "XML Domain" in Fig.4 is the area of this work.

We can use DTD or XML schema (XSD) [3] to specify the document schema, but they are focus on the logical structure of the document, and are more exposed to implementation-specific and/or implementation-oriented abstraction rather than domain specific conceptual level semantics. Recently work has been carried out to incorporate conceptual level semantics directly into XML Documents/Schema. This gives the user the power to model XML at a higher-level of abstraction using proven methodologies such as UML, ER and semantic nets [5] [6][7][8][10].

2) The XML Database Domain

When it is necessary to store XML documents permanently, we go into the area of the "XML Database Domain". To design an XML database, we should first design its conceptual schema, and then transform it into a logical schema. If using a NXDB, the transformation may be direct: while using a XEDB, the transformation may be complicated. If we emphasize the "XML-in/XML-out", a so-called native DBMS may be better [12].

3) The XML requirements and other data requirements

There are many kinds of data requirements in an enterprise or an application domain, and data can be structured or semi-structured (XML style). An ideal solution is to manage all the data in a universal database or at least to construct a universal conceptual model which can represent all the data [4]. In our options, however, all kinds of data can not be managed by a single kind of database system. So the XML database should just be used to manage the data from the XML domain, while the XML domain is just part of the application domain.

3.2 A Unified 4D perspective

By adding a new dimension, the "Domain" dimension (Dom), to the previous 3D perspective, a 4D perspective is presented in Fig.5. From this perspective we can examine both XML documents modeling and XML database modeling. It builds a unified context, and gives us different views if we view the models from different viewpoint.

In Axis-Dom we clearly distinguish the XML document and the XML database. In/among the XML domain we can examine all kinds of schemas relating an XML document. For example, by using PowerDesigner V10 [11], one can design a document's schema in a form of visual "XML Model", and then get the corresponding schema specification in the form of DTD or XSD. Similarly one can re-engineer a DTD/XSD to its "XML Model". These processes are represented in Fig. 5 by B->A and A->B link line. In the perspective some current researches [6] [8] can be considered to be along the Axis-X, i.e. from (A, B) to (C, D) or reverse.

In/among the XML database domain we can also get different schemas, however, which are used to describe the structure of an XML database rather than of an XML document. The logical schemas of a database can be relational, object-relational or native. The conceptual schemas in this domain, though very similar to the conceptual schemas in XML domain, are the specification for the XML database.

There are some researches which build mapping between the (A, B) in XML domain and the (1, 2, 3, 4) in XML Database domain [9].

Other relative work can also be discussed or analyzed in the context of that unified perspective.

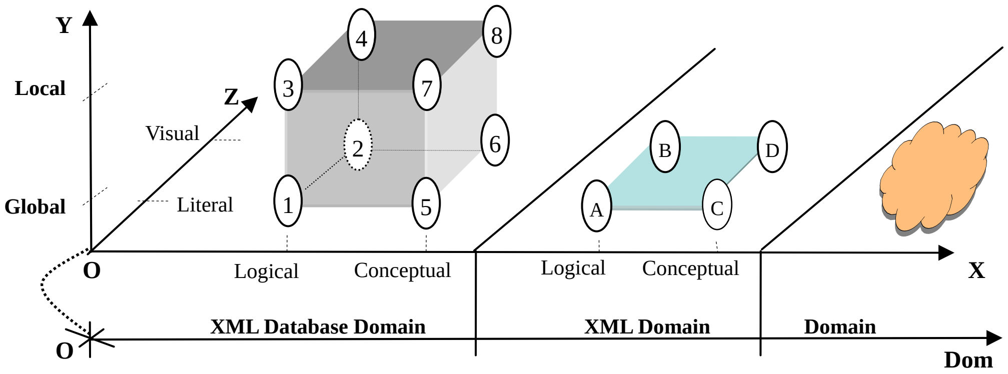


Fig.5. A 4D-Perspective to XML document schemas and XML database schemas

4. CONCLUSIONS AND FUTURE WORK

From a 3D perspective which is used to describe the traditional database modeling, a 4D perspective is presented in this paper, which can be used to explain all kinds of models relating to XML document modeling and XML database modeling. By using this perspective we can explain the phenomenon in the fields and specify or evaluate the relating approaches, models and methods.

The 4D perspective model is now presented in an informal way. Next we will define the model mathematically, and use it to analyze the requirements to an integrated XML and XML modeling CASE tool.

5. REFERENCES

[1] C.J.Date, An Introduction to Database Systems (7th Edition), Addison Wesley, 2000.

[2] Mark Graves, Designing XML Databases (Chinese edition), China Machine Press, 2002.

[3] David C. Fallside (eds), XML Schema Part 0: Primer, http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/

[4] Airi Salminen,et al., Requirements for XML Document Database Systems, ACM Symposium on Document Engineering, November 2001.

[5] R.Rajugan, et al., XML Views: Part 1, DEXA 2003, LNCS 2736,pp.148-159, Springer-Verlag, 2003.

[6] Bernadette Farias Losio, et al., Conceptual modeling of XML schemas. Proceedings of the fifth ACM international workshop on Web information and data management.pp.102-105, ACM Press,2003.

[7] Giuseppe Psaila, ERX: a conceptual model for XML documents, Proceedings of the 2000 ACM symposium on applied computing, pp898-903, ACM Press, 2000.

[8] Ronaldo dos Santos Mello, et al., A Rule-Based Conversion of a DTD to a Conceptual Schema, Proceedings of the 20th International Conference on Conceptual Modeling, pp.133-148, Springer-Verlag,2001.

[9] Popa. L,et al., Mapping XML and relational schemas with Clio. Proceedings, 18th International Conference on Data Engineering, 2002, pp.498-499, 2002

[10] Mikael R.Jensen, et al., Converting XML DTD to UML diagrams for conceptual data integration, Data & Knowledge Engineering 44 (2003) pp.323-346.

[11] Sybase, PowerDesigner(V10)-XML Model User's Guide, DC20014-01-1000-01, January 2004

[12] Michael Kay, XML Databases, Software AG, March, 2003.

[13] Oracle, Oracle XML DB, An Oracle Technical White Paper, January 2003

Liu HongXing is an associate professor in college of Computer Science and Technology, Wuhan University of Technology. His research interests are in Database Modelling, XML Database, and Enterprise Application Integration.

Construction and Maintenance of the Knowledge Base Used in GSIES-TOOL

XU Yong[a] ZHONG Luo[a] YANG Ke[b]
[a]School of Computer Science and Technology, Wuhan University of Technology, Wuhan, 430070, P.R.China
[b]College of Architecture & Civil Engineering, Wenzhou University, Wenzhou, 325027, P.R.China
Email: hfxing@sohu.com Tel.: +86 (0)27-85758496

ABSTRACT

In this paper, the construction and the maintenance of the knowledge base of GSIES-TOOL used in the domain of geotechnical engineering security inspection is introduced in detail. The knowledge representation method of this knowledge base is a kind of object-oriented rule, which is designed by us and named OORL. The major sections of this paper are as follows. First, the structure of the base is stated. According to the character of the problem solving model of this domain, the OO method and hiberarchy are used to construct the knowledge base. There are four levels in the base, project, subject, rule and formula. The lower level is the property of the consecutive upper level. Secondly, the construction of the content of each level of the base is explained. The rule and the formula are also constructed using OO method. Finally, the maintenance of the base is illuminated. The consecutive levels of the knowledge base are tightly connected with a bidirectional chain. It is safe for the user to maintain the base. The knowledge base that we have built basing on the tree structure and OO technology has better inheritability and expandability, and is easy to be managed and maintained. Its design idea can be used for reference for the relative domain.

Keywords: Knowledge base, Construction, Maintenance.

1. INTRODUCTION

GSIES-TOOL is an expert system development tool which is used in the domain of geotechnical engineering security inspection and developed by us. According to the concrete character of the security inspection in geotechnical engineering, we divide the domain knowledge into two kinds. One is how to analyze the inspection data of the measuring point and predict the future development trend. The other is how to use the first kind of knowledge to evaluate the safe performance of the whole project.

Analyzing the character of this domain, we use the OO technology to rebuild the production rule and design a new knowledge representation method. We name this method OORL, object-oriented rule, in which production rules, methods, engineering examples, theoretical criterions, logic tree and explaining engine are all encapsulated together. Using OORL, we encapsulate the methods used for analyzing the inspection data of the measuring points, and build a series of classes, the measuring point evaluation classes. Then we use these classes to build the concrete evaluation objects aimed at the measuring points. In the object, the methods are inherited from the corresponding class and encapsulated with the inspection data of the point. And so the representation of the first kind of knowledge is finished. Based on finishing encapsulating the methods and the data of the measuring points, the representation of the second kind of knowledge will be finished after the corresponding points, including their methods and data, are encapsulated together.

The knowledge base is the key and base of building the expert system [1]. Its design quality will directly affect the performance of the expert system [2]. Especially, the knowledge base of the expert system development tool should have much better expandability and inheritability [3, 4]. So after representing the domain knowledge using OORL, we use the OO technology and hiberarchy to construct the knowledge base of GSIES-TOOL in terms of the character of analyzing and solving problems of this domain.

2. DESIGN OF THE KNOWLEDGE BASE

The structure of the knowledge base is divided into four levels, project, subject, rule and formula. The project is on the upmost level, and can be separated into many subjects. And there are lots of rules included in each subject. The formula is the method used in the rule to analyze and evaluate the inspection data. The rule is the key for the knowledge base of this system, and is divided into synthesis rule and analysis rule.

OORL is the analysis rule, whose form is a binary tree. The related production rules are grouped together through a strict logic inference chain in OORL. A single OORL can solely finish a relatively absolute reasoning procedure. The analysis rule is divided into rule class and rule object. Rule class is the abstract inference method or inference chain, describes the commonness of the measuring points or instruments of a kind. The rule object is also called application rule, which is the instantiation of the rule class. The content that can not be confirmed in rule class will be reified in rule object, such as the name of the measuring point related to the application rule, the source of the inspection data, the type of the criterion, and so on. The application rule will be directly used in the inference procedure. And the inference result of the application rules will be called by the synthesis rules. The instantiation of rule class associates the actual project tightly; therefore the application rule should be constructed when the expert system of a concrete project is built using GSIES-TOOL.

The formula used in analysis rule includes compute formula and database operation formula. The formula is also constructed using OO method, and divided into formula class and formula object(application formula). Formula class describes the commonness of the formulas of a kind, and can not be used in the rule because the data is not encapsulated in it. The application formula is the instantiation object of formula class and can be directly called by the rule. The process of the instantiation of the formula class synchronizes the process of the instantiation of the rule class.

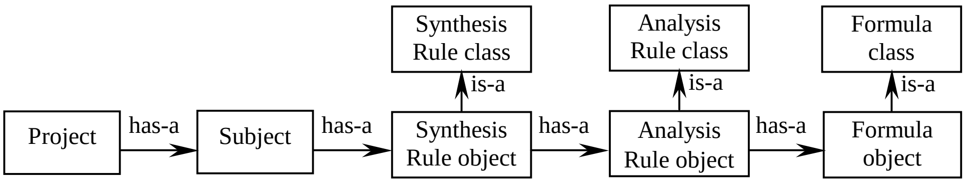


Fig.1. the structure of the knowledge base

The form of the synthesis rule is also a binary tree. In synthesis rule, the application rules are grouped together. Each node includes one application rule or more. The inference result of the application rule is the judgemental condition of the node of synthesis rule. Similar to the analysis rule, logic tree and explaining engine are also encapsulated in synthesis rule. The inference result of synthesis rule is the inference result of the last application rule in the inference chain. Corresponded with the sort of the analysis rule, the synthesis rule is also divided into rule class and rule object(synthesis application rule). The synthesis application rule is directly used in the inference machine. The instantiation process of the synthesis rule class should be processed after the application rule is built. So the synthesis application rule can be constructed only when the development process of the new system is finished

.

The construction of the project and subject are processed according to the concrete engineering and synchronizes the development of the new system too. What stored in the knowledge base of GSIES-TOOL are the synthesis rule class, the analysis rule class and the formula class. In the domain of geotechnical engineering security inspection, the analysis method of many commonness problems can be confirmed beforehand. So we can setup the analysis procedure and the inference chain in advance, and the corresponding rule classes and formula classes, and then store them in the knowledge base of GSIES-TOOL. So when GSIES-TOOL is used to develop the expert system of a concrete engineering, because of the OO character of the rule classes and the formula classes stored in the knowledge base, it is easy for the rule classes and the formula classes to be inherited into the new system. Combined with the concrete character of the engineering, these classes can derive lots of rule objects and application formulas. The knowledge base of the new system can be expanded rapidly and availably. As long as the inference data used by the formulas is gotten, the analysis and evaluation of these data can be processed immediately, and the engineering safety condition can be evaluated and predicted in time.

3. CONSTRUCTION OF THE KNOWLEDGE BASE

In GSIES-TOOL, the main construction work of the knowledge base is to build the rule and the formula. On constructing the knowledge base, the domain expert and knowledge engineer describe the domain knowledge as rule or formula through the knowledge acquisition module, which includes formula edit module and rule edit module, and store the rules and the formulas in the knowledge base, as in Fig.2.

When the formula class is being built, in order to build the compute formula, the requisite work is only to separately describe the constants, the independent variables and the intermediate variables used in the formula. After the data source of the variables have been appointed, the application formula is built.

The database operation formula is a set of database operation commands in fact. In GSIES-TOOL, we develop the database formula edit module, DBFE. We use the similar C language to describe the operation and build the database operation formulas, which will be interpretively executed by DBFE. In the knowledge base, the database operation formulas are stored in the text files separately.

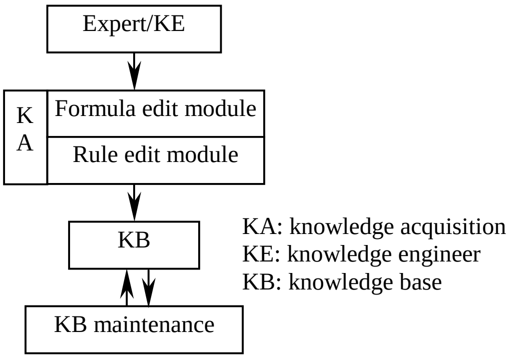


Fig.2. the implementation of KB

A concrete database operation formula is stated as follows. This formula is used to read the data of the former three fields of a table and writing the data to a new table.

```
Begin;
char FileName1[10];
char FileName2[10];
char FieldData[m][3];
char FieldType[3];
int m;
GetFileName(FileName1[10]);
GetFileName(FileName2[10];
OpenFile(FileName1[10]);
m = GetRecordNum(FileName);
GetFieldType(FileName1[10],FieldType[3]);
CreateFile(FileName2[10],FieldType[3]);
GetData(FileName, FieldData[m][3]);
WriteFile(FileName2[10],FieldData[m][3]);
End;
```

In GSIES-TOOL, a tree structure editor is provided for the user to build the application rule class according to the need of handling problem. And it is permitted that the new rule class can be inherited from the existent rule class. The application rule is inherited from the rule class or the existent application rules. In the domain of geotechnical engineering security inspection, there are many rules whose forms are same or similar. The difference between these rules only consists in using the different data and the different criterions to evaluate and predict the safety condition. So the inheritance of the rule is very important. The workload of building the knowledge base can be decreased greatly because of the inheritance. When the application rule is built, the properties of the rule nodes should be described clearly. The properties of the procedure nodes include the methods and the criterions. And the property of the end nodes includes the degree of safety, which is the quantization of the letter of the inference result.

The other editor similar to the tree structure editor used for building the analysis rules is provided for the user to build the synthesis rules. Each node of the synthesis rules encapsulates one application rule or more. So when building the synthesis rule, the user is asked to appoint the application rules used by each node and the corresponding criterion.

In our system, rule class(including analysis rule class and synthesis rule class) has the same form as the rule object(including analysis object and synthesis rule object). But in the rule class, there is only the tree structure. While in the rule object, there is not only the tree structure but also the concrete description about each node.

The analysis rule and the synthesis rule are related with the instruments or the measuring points of a kind. So in knowledge base, the rule is named using the name of the instruments or the measuring points, and stored group by group according to the kinds of the instruments or the measuring points. When the rule belonging to some instruments or measuring points is added to the knowledge base, it is easy to know if there is the same rule in the base through searching and matching the name of the existent rule. Then it is avoided to build the repetitive rules.

4. MAINTENANCE OF THE KNOWLEDGE BASE

In GSIES-TOOL, formula class, application formula, application rule, synthesis application rule are all stored in the corresponding files. The number or the name of what have been built, including the application formula, the formula class, the application rule, the analysis rule class, the synthesis rule class, the synthesis application rule, the project and the subject, are all stored in the database file. In fact, the knowledge base of GSIES-TOOL is constituted by a series of database files, formula files and rule files. In the file of a level, what recorded is not only the information of this level but also the information of being called by the upper level and calling the lower level. For example, when an application rule is recorded in the application rule file, the information that will be recorded includes not only the rule number and the name of the instrument or the measuring point which is related with but also the number of the synthesis application rules which call this application rule and the number of the application formulas which are called by this application rule. In this way, the consecutive levels of the knowledge base are tightly connected through the bidirectional chain described as above.

Because of the correlation between each level of the knowledge base, the maintenance of any level will affect the consecutive levels. Under the control of the bidirectional chain, deleting the content of some level is commanded by the upper level. Building or modifying the content of a level need not ask for the permission of the upper level. But the content of a level can be deleted only with the permission of the upper level. That is to say, the correlation chain with the upper level must be broken off firstly before deleting the content of a level. While building, modifying or deleting the content of a level is finished, the maintenance information should be written not only to the files of this level but also to the files of the consecutive lower level.

We take the maintenance of the application rule as the example for detailed explanation of the bidirectional chain. The deletion of the application rule is commanded by the synthesis application rule, which is on the upper level. When an application rule is deleting, system will automatically detect the synthesis application rules which call the application rule, and ask the user to modify these synthesis application rules firstly. Only after the modification of all these synthesis application rules finished can the application rule be deleted. And after the user finishes the maintenance of the application rule, system will automatically write the corresponding information to the application formula, which is on the lower level.

Under the control strategy stated as above, the safety of maintaining the knowledge base is insured availably. And the conflicts resulted from the maintenance of knowledge base will be avoided. And then the stability of the knowledge base is ensured. In GSIES-TOOL, the user can complete the maintenance work through the visual operation. So the maintenance of the knowledge base is very precise and convenient.

5. CONCLUSION

In the development procedure of GSIES-TOOL, we construct the OORL using the OO method to represent the knowledge in the domain of geotechnical security inspection. And the OO method is also used for building the formula. The formula and the logic tree are all encapsulated in OORL. So the judgmental knowledge and the procedural knowledge can be represented conveniently using OORL. Compared with the traditional knowledge representation methods, OORL has clear structure, concise inference, and higher modularization degree. On the basis of describing the domain knowledge using OORL, we use the OO technology and hierarchy to construct the knowledge base of GSIES-TOOL according to the classification of the domain knowledge. Because of the hierarchy and the OO idea used in the knowledge base, this base has good expandability and inheritability, and is easy to be built, managed and maintained. And the quantity of the rules and the times of repetitive judgment are decreased consumedly, the inference efficiency is increased immensely. When the GSIES-TOOL is used for developing the new expert system, what stored in the knowledge base will be conveniently inherited into the new system. So the new system can be putted into service quickly.

The global structure of the knowledge base of GSIES-TOOL is a tree in fact. The rules, including the synthesis rules and the analysis rules, are all built using the tree structure. The inference procedure of the system is a traverse procedure of the tree. If there is invariability in the structure of the knowledge base, the inference machine needs not be modified. So it is achieved that the knowledge base is independent of the inference machine. And then it is insured that GSIES-TOOL can be generally used in the domain of geotechnical engineering security inspection.

6. REFERENCES

[1] K. W. Chau, F. Albermani, “Expert system application on preliminary design of water retaining structures”, Expert Systems with Applications, Vol.22, Issue.2, February 2002, pp. 169-178.

[2] XIA Hong-xia, SONG Hua-zhu, ZHONG Luo, “The Design of the Knowledge Base in CARRES98”, Mini-Micro Systems, Vol.20, No.12, December 1999, pp.937-940.

[3] Alain O. Villeneuve, Jane Fedorowicz, “Understanding expertise in information systems design, or, What's all the fuss about objects?”, Decision Support Systems, Vol.21,Issue.2, October 1997, pp. 111-131.

[4] Richard C. Hicks, “Knowledge base management systems-tools for creating verified intelligent systems”, Knowledge-Based Systems, Vol.16, Issue.3, April 2003, pp. 165-171.

Xu Yong is the lecturer of Centre of Information and Network in Hubei Administration Institute. He graduated from Wuhan University of Technology in 1996; obtained his master’s degree in computer application technology from Kunming University of Technology in 2000. Presently he is specializing in doctor’s degree in Wuhan University of Technology. His research interests are in artificial intelligence, network security and e-government.

Zhong Luo is a Full Professor .He graduated from Wuhan University in 1982. His research interests are in intelligent technology, software engineering, and image graphic.

Create Distributed Application with Java RMI to Manipulate BLOBs *

Wang Jingyang,Wang Jianxia, Zhang Xiaoming, Qin Min, Fu Dong
College of Information Science & Engineering, Hebei University of Science and Technology
Shijiazhuang Hebei 050054, China
Email: jingyangw@hebust.edu.cn **Tel.:** 0311-8613336

ABSTRACT

This article introduces Java Remote Method Invocation (RMI) architecture, principle and the main steps of developing distributed Java programs with RMI. We present a method of creating distributed programming with Java RMI based on JDBC (Java Database Connectivity) to manipulate BLOBs (Binary Large Objects) stored in database, illustrate the working principle and working process of the method, and provide an example manipulating BLOBs stored in Oracle's LONG RAW field.

Keywords: Java RMI, Manuscript, Distributed Programming, BLOBs, JDBC.

1. INTRODUCTION

Most Java programmers have used JDBC to create and access tables containing the familiar address book data types, such as strings, integers, floats, and dates. However, many modern applications require the management of much larger data objects, from images, which may require tens of kilobytes of storage, to video clips, which may run into the hundreds of megabytes. The earliest approach to handing large objects was to store them as files in the underlying operating system, using the database to store only the file path and letting the application code manage the file. While this approach works fine, anyone who has tried maintaining a complex Web site knows all about broken links that can result from someone moving a file or renaming a directory. Nowadays most databases also support so-called large objects. There are several commonly implemented large object data types. Tow of the most widely used are BLOBs (such as OLE object data type in Access, image data type in SQL server, LONG RAW data type in Oracle) and Character Large Objects (CLOBs). BLOBs and CLOBs are particularly useful because they are a great way to handle objects such as images and documents within your database. How to operate these data efficiently is a difficult problem. This paper introduces a method of creating distributed programming with Java RMI to manipulate BLOBs stored in database based on JDBC, provides an example manipulating BLOBs stored in Oracle's LONG RAW field.

2. RMI

The Java 2 Enterprise Edition (J2EE) remote method invocation (RMI) framework enables you to create virtually transparent, distributed services and applications. RMI-based applications consist of Java objects making method calls to one another without regard for their location. This allows one Java object to invoke methods on another Java object residing in another virtual machine in the same manner in which methods are invoked on a Java object residing in the same virtual machine [1].

The RMI system consists of three layers: Stub/Skeleton Layer Client-side stubs (proxies) and server-side skeletons; Remote Reference Layer Reference/invocation behavior (e.g., unicast, multicast); Transport Layer set up and management and remote object tracking [2]. The RMI architecture is shown in Figure 1.

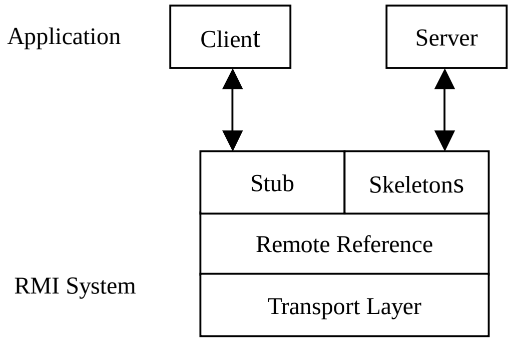


Figure1 Java RMI architecture

Java RMI relies on a protocol called the Java Remote Method Protocol (JRMP)**.** Java relies heavily on Java Object Serialization, which allows objects to be marshaled (or transmitted) as a stream. Since Java Object Serialization is specific to Java, both the Java RMI server object and the client object have to be written in Java. Each Java RMI Server object defines an interface, which can be used to access the server object outside of the current Java Virtual Machine (JVM) and on another machine's JVM [3]. The interface exposes a set of methods, which are indicative of the services offered by the server object. For a client to locate a server object for the first time, RMI depends on a naming mechanism called an RMIRegistry that runs on the Server machine and holds information about available Server Objects. A Java RMI client acquires an object reference to a Java RMI server object by doing a lookup for a Server Object reference and invokes methods on the Server Object as if the Java RMI server object resided in the client's address space [4].

Developing a distributed application using Java RMI involves the following steps [5]:

1. Define a remote interface
2. Implement the remote interface
3. Develop the server
4. Develop a client
5. Generate Stubs and Skeletons, start the RMI registry, server, and client

* This paper is supported by the fund of Hebei University of Science and Technology (Grand NO. 2003132)

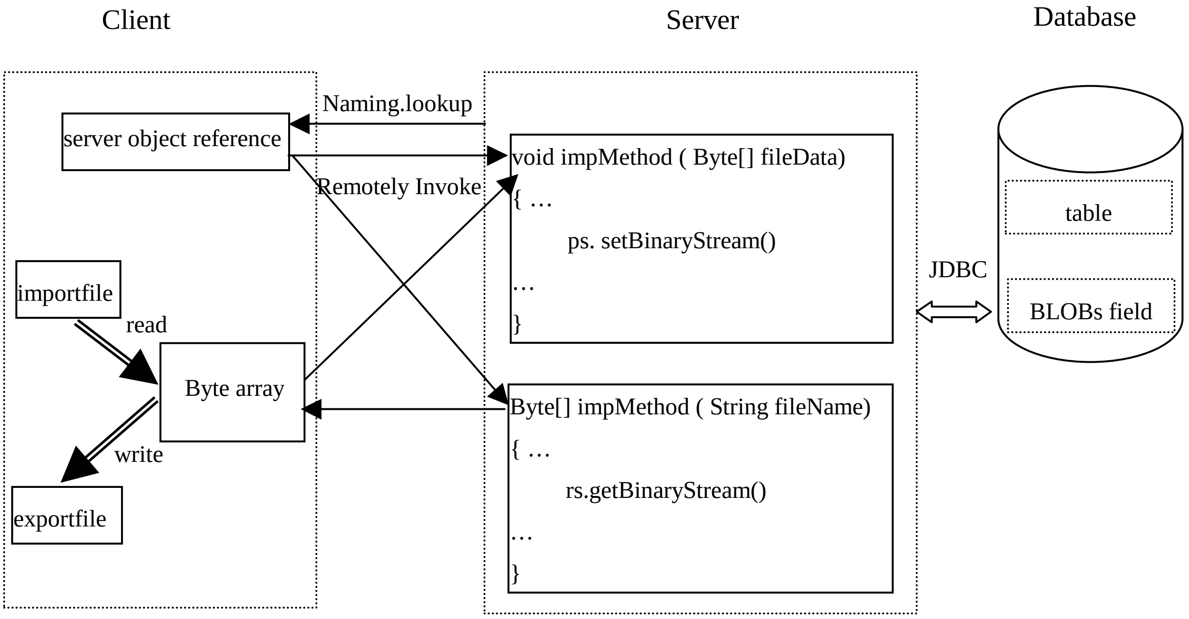


Figure2 The principle of manipulating BLOBs with Java RMI

3. MANIPULATING BLOBs

Because common fields store fixed length data such as integer, float etc. or variable length string, however BLOBs field can store variable length binary data, we can't manipulate it as common fields data. We need store BLOBs using setBinaryStream method of JDBC's prepareStatement object and get it using getBinaryStream method of JDBC's ResultSet object.

3 .1 Importing BLOBs to Database

Importing BLOBs to database using RMI involves the following steps:

Client:

1. Construct FileInputStream object with file to be imported;
2. Read the BLOBs to byte array by FileInputStream object's read method;
3. Acquire Java RMI server object reference using Naming class's lookup method;
4. Invoke Java RMI server object method importing the BLOBs in byte array to database through the reference;

Server:

1 Create connection with database using JDBC;

2 Construct ByteArrayInputStream object with the BLOBs in byte array from Client;

3 Write BLOBs in ByteArrayInputStream object to database using prepareStatement object's setBinaryStream method;

3.2 Exporting BLOBs from Database

Exporting BLOBs data from database using RMI involves the following steps:

Client:

1. Acquire Java RMI server object reference by Naming class's lookup method;
2. Invoke Java RMI server object method exporting the BLOBs from database to byte array through the reference;
3. Construct BufferedOutputStream object with exported file;
4. Write BLOBs in byte array to exported file by write method of BufferedOutputStream object.

Server:

1 Create connection with database using JDBC;

2. Read the BLOBs stored in database to byte array using ResultSet object's getBinaryStream method;

3.3 Principle

The principle of manipulating BLOBs using Java RMI is shown in figure 2.

4. EXAMPLE

This example implements importing all kinds of files (BLOBs) in client to Oracle's LONG RAW field through java RMI server. The table name of the example is bin_data, which has two fields, one is name (VARCHAR2 type) the other is data (LONG RAW type).

4.1 Define a Remote Interface

The remote interface LongrawInterface provides a importFile method which takes two arguments one is byte array (the data of imported file) the other is String (the name of the file). The code is shown as follow:

```
    public interface LongrawInterface extends Remote {
        public void importFile(byte[] fileData, String
fileName) throws RemoteException;    }
```

4.2 Implement the Remote Iinterface

The following implementation class illustrates how the LongrawInterface interface could be implemented to define a valid remote object:

```
    public class LongrawImpl extends
    UnicastRemoteObject implements LongrawInterface {
      public LongrawImpl() throws RemoteException {
```

```
      super();
    }
    public void importFile(byte[] fileData, String
fileName) {
//connect the Oracle database using JDBC
//register the driver of Oracle

                        Class.forName("Oracle.jdb
                        c.driver.OracleDriver");
//create Connection interface object
     Connection conn =
DriverManager.getConnection(
          "jdbc:Oracle:thin:@127.0.0.1:1521:hbkj",
"system", "manager");
//construct ByteArrayInputStream object with fileData
      ByteArrayInputStream kk = new
ByteArrayInputStream(fileData);
      PreparedStatement ps =
conn.prepareStatement(
           "INSERT INTO bin_data (name,data)" +
"VALUES (?, ?)");
//set the value of 'name' field in bin_data with
fileName
        ps.setString(1, fileName);
//set the value 'data' field with the data in kk
        ps.setBinaryStream(2, kk, (int)
     fileData.length);
        ps.executeUpdate();
        ps.close();
  }
```

4.3 Develop the Server

There are three things that the server needs to do:

1. Create an instance of the RMISecurityManager and install it.
2. Create an instance of the remote object (LongrawImpl in this case).
3. Register the object created with the RMI registry.

The main code is shown as follow:

```
public class LongrawServer {
  public static void main(String argv[]) {
      //Create an instance of the
      RMISecurityManager //and install it
      System.setSecurityManager(new
RMISecurityManager());
      //Create LongrawInterface object fi
      LongrawInterface fi = new LongrawImpl();
      //Register fi with "//127.0.0.1/FileServer"
      Naming.rebind("//127.0.0.1/FileServer", fi);
  }
}
```

4.4 Develop a Client

The client remotely invokes any methods specified in the remote interface (LongrawInterface). To do so however, the client must first obtain a reference to the remote object from the RMI registry. Once a reference is obtained, the importFile method is invoked. A client implementation is shown in following. In this implementation, the client accepts two arguments at the command line: the first one is the name of the file to be downloaded and the second one is the address of the machine from which the file is to be downloaded, which is the machine that is running the file server.

```
public class LongrawClient {
  public static void main(String argv[]) {
      String name = "//" + argv[1] + "/FileServer";
      //acquire Java RMI server object reference fi
      LongrawInterface fi = (LongrawInterface)
Naming.lookup(name);
      //construct files with the first argument
      File files = new File(argv[0]);
      //create fis with files
      FileInputStream fis = new
FileInputStream(files);
      byte[] fl = new byte[ (int) files.length()];
      //read the data of fis to f1
      fis.read(fl);
      //remotely invoke importFile method
      fi.importFile(fl,argv[0];
  }
}
```

5. CONCLUSIONS

This paper introduces a method of creating distributed programming with Java RMI to manipulate BLOBs stored in database based on JDBC, It is easy for programmer to implement this method and the method is efficient. But the method is weak to support different programming languages. Both the Java RMI server object and the client object have to be written in Java. If these objects of this method written in different programming languages, we can develop distributed programming with The Common Object Request Broker Architecture (CORBA).

6. REFERENCES

[1]. Pallavi Jain,Shadab Siddiqui. J2EE Professional Projects [M]. Premier Press, April 2002

[2]. Yan Feng. Developing distributed applications based on CORBA and RMI [J], Journal of Changchun Institute of Optics and Fine Mechanics, 2002,3:45-47 (in Chinese)

[3]. Qusay H.Mahoud. Distributed Programming with Java [M]. National Defence Industry press, April 2002

[4]. David Reilly, Michael Reilly. Java Network Programming and Distributed Computing [M]. Person Education Press, March 2003

[5]. Yang Hui, Lu Wei. The Research of RMI and Its Application on EJB [J], Journal of Sichuan University (Natural Science Edition), 2003, Vol.40 No.01:45-50 (in Chinese)

Wang Jingyang (1971-), male, he is a lecturer of Hebei University of Science and Technology. He graduated from Lanzhou University with specialty of Computer Software Science in 1995. He has published two books, over 10 Journal papers. His research interests are in Network and Database application technology.

Cooperation Agent Applications for MKA Based on Grid Computing*

Xia Huosong
Wuhan Institute of Science and Technology, P. R. China, 430073
E-mail: bxxhs@sina.com **Tel:** +86(0)27-87800491

ABSTRACT

The enterprise marketing knowledge acquisition is more important for effective marketing management. This paper discusses the ways and mechanism of the marketing knowledge acquirement and retrieval based on grid computing. The cooperation agent, multi-agent and mobile agent will have profound impact on MKA. To solve complex problems, these agents must work cooperatively with other agents in a heterogeneous environment. The ant algorithm for the applications of cooperation agent in marketing knowledge acquirement is presented. And a multi-agent distributed marketing knowledge acquirement scheme in ant algorithm is constructed according to the hybrid different structure mechanism based on grid computing. The component of the architecture is analyzed.

Keywords: Marketing Knowledge Acquisition (MKA), Grid Computing, Cooperation Agent, Ant Algorithm.

1. INTRODUCTION

Marketing knowledge acquisition for the marketing management and electronic commerce over the cyberspace has become more and more important. It has been estimated that the amount of information stored on the Internet doubled every 18 months and the speed of increase of home pages doubled every 6 months or even shorter [1]. In the same time, in the field of data mining, knowledge discovery in database or Web and research on data warehouse, the algorithm for automatic knowledge acquirement has made progress continuously. Agent is a kind of software pack for creating knowledge tirelessly on Internet with fulfilling repeated and anticipated task. That called agent, intelligent agent, software agent or software robot, in fact means the same procedure. It can present individual or organization to complete information collecting, filtering and adjusting (performing on the backstage). Whenever you need, it can offer help to reduce the work of finding, comparing, negotiating and purchasing when people buying products or services on Internet, as well as the work of the knowledge searching or acquiring. So it has advantages of long-term utilization, half-automation, participation and conformation. In some areas, like Massachusetts Institute of Technology, Stanford University, National Computer and Half-automation Institute (France), have carried out the research work on multi-agent and have won initial achievement. Dr. Pattie Maes, the founder of Massachusetts Institute of Technology's software agent group, created the first and real successful software agent for making information filter correctly. There have been software agents like AuctionBot, BargainFinder, Firefly, Kasbah and Excite Jango; there have also been tools for building models agent-oriented (testing edition), as IBM's ABE (Agent Building Environment), Aglets & Gensym's ADT (Agent Development Tooket) and Stanford University's JAT Litle (Java Agent Template, Litle), etc. Much of the current research related to intelligent agents has focused on the capabilities and structure of individual agents. However, to solve complex problems, these agents must work cooperatively with other agents in a heterogeneous environment [8], related to the method and mechanism of market knowledge retrieval and acquirement based on web computing, the research for definition and application in this field has become a challenge to the enterprise and researcher. Therefore, cooperation agent applications for MKA based on grid computing is the most important and exciting areas of research and development.

2. COOPERATION AGENT AND ANT ALGORITHMS

An agent is a hardware or (more usually) software entity with (some of) following kinds: it is an entity working autonomously and continuously; it is an entity that can apply agent's communication language correctly; it is an entity with spirit state. Generally, it's considered as the computing entity that is designed to complete some kind of task, give full play independently under certain circumstance, and have its life cycle. Agent is a kind of abstract entity, which can act on itself and its environment, and make response to the environment with knowledge, target and ability. The basic characteristics of agent are autonomy, target-forced, responsive, ongoing execution and reproduce; while the non-basic characteristics are environmental awareness, awareness, adaptive, mobility, intelligence and anthropomorphism. In agent of program, the definition means: A×A (Autonomous Agent) = (M,K,A,I,L,S,F,G,R,C). Here, M means method, K means knowledge, A means attribution, L means language, S means sending information operation, R means receiving information protocol, G means general knowledge, F means forward mechanism and C means complex service. Multi-agent is a collection made up with many agents applicable for web computing. It can give answers according to the problems, and it is also a distributed intelligent agent that can correspond, communicate, and cooperate with other agents to accomplish solving the same problem. Under the computer web cooperation supported (CWCS), it expresses as a scheme, which can share computer source (hardware, software, DB, KB and etc.) to the largest length, as well as knowledge, task and middle result; and can also complete problems with high effect, high quality and high level cooperated with the unit's intelligence, expert group's experience and knowledge.

Mobile agent, as an outcome coordinated between agent and Internet's web technology, can be replaced from a main computer to another main computer on the Internet autonomously, according to the user's specified task to retrieve, filter and collect information as well as standing for

*Supported by HuBei Province Fund of China(2003B002,2003aa404B08,2003Q003)

the user to carry out business activity. The classical method for building agent is to consider it as a special kind of conscious system, specifying its manners like belief, desire, intention, obligation and promise as well as creating complex entity and action to predict its behavior. In the perspective of agent's structure, there are structures as knowledge management system, external information acquirement, solving problems with field knowledge and dynamic information with storing function, negotiating and behavior evaluation.

To apply multi-agent system successfully, the following questions must be answered: software system supporting building models for agent (including languages for editing procedures, knowledge expressing, communication and frame of model for higher level); internal structure and system structure for agent; mechanism for system programming and dispatching, cooperating and negotiating, conflict dispelling and dead-lock checking; intelligent mechanism for evaluation and prediction of dynamic behavior and function; self-organizing, self-studying and self-maintaining; security tactics; building system model with complex and different structure and so on.

Marco Dorigo and other ones, the founders of ant algorithm, inspirited by the ant group's social behavior in natural world, proposed a new kind of distributed imitation algorithm. The basic idea is imitating ant's expressing behavior when communicating depended on information element. On the basis of overall agent, guided by greedy method to form process of self-catalyzed, directing each agent's action, this is a kind of common-used sounding-out ways in chance. We consider this algorithm applied for enterprise web computing to acquirement. In multi-agent system, the demand for each agent may not high. As large-scaled computing carried out at the same time, it makes ant group to avoid portion-best; as positive-responsive mechanism is used, it accelerates restraining speed and as structure greedy algorithm is applied, it finds better applicable answers when searching in early stage. The most important key part of ant algorithm is the solution of trace intensity. As the algorithm executing, the trace in shorter path is increasing its density, and related branch path has more possibility to be chosen by ant. For each way visited by every ant saved, forbidden table is designed. It is very worthy to make use of characteristic of cooperative intelligent ant group system into marketing knowledge acquirement based on web computing. For example, the matter of shortest way in logistic sending: r_{ij} equals *(i, j)* branch path's information element trace; d_{ij} means the distance between them and the definition of visibility is: $y_{ij}=1/d_{ij}$. When choosing the knot of branch, it executes as follow possibility distributed: $p_{ij}{\sim}r_{ij}^{a}\ y_{ij}^{b}$. If "*y*" hasn't been visited, the possibility is 0. There' s also trace renewing equation: $r_{ij}^{new} = s{*}r_{ij}^{old} + (\triangle r_{ij}^{1} + \triangle r_{ij}^{2} + \ldots\ldots + \triangle r_{ij}^{m})$. Here, 1-*s* means the degree of trace expressing and m means the quantity of ants. If ant "*k*" has visited the branch path *(i, j)*, then $r_{ij}^{k} = Q/L_k$. Q means the total information element each ant leaving behind in a cycle time and L_k means the length of complete and closed path made by ant. By doing this, we can solve the problem of shortest way with continuous and repeatedly computing. There are also algorithms like Ant Group System (AGS) or maximum and minimum ant system (MMAS) for marketing knowledge acquirement based on web computing. When an ant finds the standard document suited for searching, it will share the document with other ants in order not to check the same document. For the index machine, the cooperation between ants makes it possible of each index to protect source depending on distributed index loading in different browser. In the course of searching, user can search for the field that is restricted for local ant or the entrance spread by ant.

3. COOPERATION AGENT ARCHITECTURE FOR MKA BASED ON GRID COMPUTING

It is a tendency to design a kind of architecture with analyzing and acquiring marketing knowledge automatically, classifying customers and describing customers' behavior demand and buying pattern automatically. Agent can reach the purpose. The basic parts made up for web computing are: customer/server (a multi-customer computer, multi-server computer and operating system and communication system supporting distributed computing within enterprise); distributed database; data warehouse; web and correspondence; web and system management and all kings of web application. In the perspective of platform development, information management system has produced four kinds of patterns: ultimate main- computer pattern, document server pattern, customer/server pattern (C/S) and web browser/server pattern (B/S). Recently, the most popular research of grid computing pattern will become the mainstream pattern for platform of MIS in next 20 years. The architecture for grid system includes: grid structure layer, grid service layer, grid applying tools layer and applying layer. The most distinguish between information service grid and web is becoming integral, which means what users are looking at is not countless and multi-kinds websites, but only the entrance and the system casting light upon. What is includes is not only computer and WebPages, but also all kinds of information source. They all link together to form integrity, so the whole web is just like a huge computer offering integral service to each user. Agent is important technology of information platform integrally. In the multi-agent system based on web computing, the relationships among agents, some are cooperation while some are competition. So with the complex and intricate relationships among agents, it is quite difficult to describe this kind of behavior of hybrid different structure system. According to the distributed artificial intelligence, agent can be divided into four kinds: deliberative agent, reactive agent, hybrid agent and belief-desire-intention agent. The internal structure and behavior pattern in system of each agent have some differences. In compliance with Nwana's classification based on its different roles and effect in application, there are cooperation agent, interface agent, mobile agent, information agent, respective agent and hybrid agent. The cooperation agent architecture for MKA based on grid is as Figure 1 showing.

Interface agent emphasizes autonomy and study in order to serve users. It's used mainly for acquiring knowledge of user models the time long or short for browsing, key words for searching and the degree to be interrelated so as to optimize the interface between person and computer of system Information Agent manages and operates to collect relative information in the environment of distributed web computing. According to the user's interest demand, it can retrieve relative information monitor exchange of information source, program task making use of its controlling mechanism knowledge and so on, answer question and report to users initially, timely and automatically.

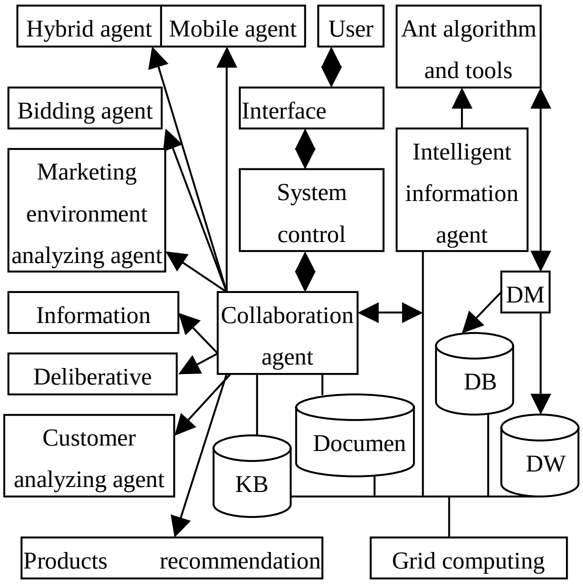


Fig. 1 Cooperation agent architecture for MKA based on grid computing

Deliberative Agent completes negotiating to agree protocol as task so that it can acquire knowledge of user's value tendency in the course of coordination actively with users by using ant algorithm (solution for trace intensity) and sharing information and knowledge deliberatively.

Mobile Agent can carry out knowledge acquirement in long distance servers and stand for users to execute task, including searching products with satisfactory price and quality automatically in purchasing activity, collecting information and analyzing user' interest long or short changing frequency and so on. Hybrid Agent means two or more than two kinds of Agent formal as different structure agent system. Here one kind is when we carrying out agent analysis on buying-product interest with information and interrelation the expression shown for the degree of interest for different products, another kind of is agent for knowledge for knowledge acquirement with products knowledge.

Intelligent information agents are autonomous computational software entities that are especially meant to (1) provide pro-active resource discovery, (2) resolve information impedance of information consumers and providers, and (3) offer value-added information services and products [1]. The next generation Internet era is grid-computing era. The "grid" was coined in the mid 1990s to denote a proposed distributed computing infrastructure for advanced science and engineering [9]. Now, it has emerged as an important new field. Grid computing can be defined as coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations. The features of grid computing are orientation to how to solve changing problems and how to deal with complex resource structure and multiple resource management. Grid computing is the infrastructure for the next generation Internet. Grid computing, an important method to solve complex marketing knowledge acquisition application, is growing more successful in distributed computing.

Base and user domain knowledge in non-database system .By means of biding agent, we gain knowledge of large customers distributed sellers and sellers with strengthen; By ways of products recommendatory agent, we can recommend products definitely as well as gain customer knowledge we can recommend products and prediction for new products; Customer analysis agent, in the way of knowledge acquirement by data mining stored in knowledge base, make deeper analysis about the customer pattern at the same time being able to "bargain" with customer and agree on the trade price (Now the method for EC are only two: fixed price and auction price) and Market Environment Analysis Agent is used for analyzing macro or micro environment knowledge of internal of external enterprise.

Database mainly means user and product database. The process of data mining generally includes data washing, data collecting, data choosing, data changing, data mining, pattern evaluating and knowledge expressing as its segment. Data mining means exploring interacting knowledge from large-quantity data stored in database, data warehouse or other information base.

Take example for this, by ways of sending inviting bid demand of this system agent, besides spreading ways. There is also methods aimed at related units to sending calling for bid automatically to make evaluation of submitted biding document based on knowledge base next selecting two kinds of building document; one passes each items of evaluation rules; another disobey each rules with a little lager difference and at last submitting user's decision by Dividing Agent. What's more, by Using agent to advertise with definite aim (If what advertisement and promote sales have made just satisfied customers' individual demand, customers will inform the products which he is really interested in. By dong this, it saves personal time, gain satisfactory service and enterprise saves its cost) and discover customer, find more accurate customer models for enterprise. Analysis users' demand more collectively, carry out selling for related product, promote attribution for products as well as its buying quantity and selling quantity. And keep good relationship with customers.

High-level collaboration agent with other agent can rely, for example, on service brokering, matchmaking, negotiation, and collaborative filtering, whereas collaborating with its human users mainly corresponds to the application of techniques steaming from human-computer interaction and affective computing.

4. CONCLUSIONS

Agent make satisfied expression and special characteristic in the field of MKA, retrieval and distributed imitating algorithm. In architecture of web computing, by using its characteristic put forward separately, Marketing knowledge retrieval and acquirement offer a basic prototype with creativity. As java realizing its function with feature of applying platform panned, we are applying PB7.0 and JSP realize this prototype system.

5. REFERENCES

[1] Christopher C. Yang, Jerome Yen, Hsinchun Chun, Intelligent Internet searching agent based on hybrid simulated annealing [J], Decision Support Systems 28(2000) 269-277

[2] Dorigo M, Maniezzo V, Colorni A. Ant system: Optimization by a colony of cooperating agents. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 1996 26(1):30-41.

[3] Filippo Menczer, Complementting search engines with online web mining agents, Decision Support Systems, 2003,35:195-212

[4] Mary E. Zellmer-Bruhn, Interruptive Events and Team Knowledge Acquisition, Management Science ,Vol. 49,No.4 April 2003:514-528

[5] Mattias Klusch. Information agent technology for the Internet: A survey. Data & Knowledge Engineering,2001,36:337-372

[6] Kwok-Wai Cheung et al. Mining customer product ratings for personalized marketing, Decision Support Systems ,2003,35:231-24

[7] Scott A. Deloach, at. el. Multiagent Systems Engineering. International Journal of Software Engineering and Knowledge Engineering,2001,1(3):231-258

[8] Xia Huosong. The methods and systems of knowledge acquisition and knowledge sharing in MKMS based on EC , A Dissertation for the degree of doctor of Huazhong University of Science and Technology,2002.4.29

[9] The Globus Project, http://www.globus.org

Xia Huosong is a professor at the Department of information management and information systems of the Wuhan Institute of Science and Technology. He studies as Ph.D. from *Management School, Huazhong University of Science and Technology. His main research interests are: data* mining and knowledge management, MIS,DSS, e-Commerce. He is the author or the first author of more than 60 papers in journals. He has participated in many MIS projects. E-mail Box: bxxhs@public.wh.hb.cn

Accessing BLOB Data Stored in Database Based on ADO in Visual C++ *

Qin Min, Wang Jingyang, Zhang Xiaoming, Yang Kuihe
College of Information Science & Engineering, Hebei University of Science and Technology
Shijiazhuang Hebei 050054, China
Email: qm@hebust.edu.cn **Tel.:** 0311-8613184

ABSTRACT

BLOB fields in database can store BLOB (Binary Large Object) such as graphics files, audio files and large format text files etc. This article introduces a simple and efficient method to access BLOB data in database by internal objects of ADO (ActiveX Data Object) in Visual C++. We also illustrate the working principle and working process of the method, which provides an efficient approach for the developing of multimedia information systems.

Keywords: BLOB, ADO, SAFEARRAY, Database, Visual C++.

1. INTRODUCTION

With the rapid development of the computer science and multimedia technology, database should store not only standard type data but also LOBs (Large Objects) data (such as graphics files, audio files and large format text files etc.), which can be used in the applications. LOBs can be stored in database in two forms. One is BLOB, and the other is character type. Different database provides corresponding data type to support the storage of LOBs. The typical BLOB data type is OLE object in ACCESS, image data type in SQL SERVER, LONG RAW and BLOB data type in Oracle; the typical character data type is text and next data type in SQL SERVER, CLOB and NCLOB data type in Oracle. How to operate this data efficiently is a difficult problem. This article discusses LOBs data stored with BLOB form, introduces an efficient method to manipulate BLOB data in database by the internal ADO objects in Visual C++.

2. ADO INTRODUCTION

OLE DB defines a group of COM (Component Object Model) interfaces, which are low-level data accessing interface. It encapsulates the function of ODBC, and through it we can access data stored in different database by uniform method. Microsoft UDA (Universal Data Access) is based on OLE DB.

ADO is a high-level programming interface based on OLE DB, it inherits the virtue of OLE DB, encapsulates the interface of OLE DB, defines ADO objects, and makes developing program simple. From data of relation database to E-Mail file system to Html files to ISAM files to Excel files and custom files can be accessed by ADO. The programs with ADO technology have good transplantability and extensibility. ADO provides a group of efficient interfaces in which the fundamental interfaces are _ConnectionPtr, _CommandPtr and _RecordsetPtr.

_ConnectionPtr is used to create a connection with database or execute a SQL statement without any result.

_CommandPtr returns a record set. It provides a simple method to execute a storage process and a SQL statement to return a record set. When using _CommandPtr interface, we can use global _ConnectionPtr interface or connection string directly.

_RecordsetPtr is a Recordset object. It provides more function to control record set, such as locking record and controlling cursor etc compared with _ConnectionPtr and _CommandPtr.

3. OPERATING BLOB DATA BY ADO

ADO exchange Data with application program using COM data types. COM data types have VARIANT, BSTR, SAFEARRAY and HRESULT etc. Related COM encapsulation classes have _variant_t, _bstr_t and _com_error etc.

VARIANT is a kind of structure including data type and data value. VARIANT structure can store all kinds of data, including standard type data, COM point and all COM type data. _variant_t class encapsulates VARIANT structure, and provides methods to manage VARIANT structure.

BSTR is a kind of structure including string length and buffer. _bstr_t class encapsulates BSTR structure, and provides methods to manage BSTR structure.

SAFEARRAY is a kind of structure describing an array composed by other data type, including the dimension and bound of the array which can limit program to access beyond bound. We can use it to transport a lot of data in same type by a parameter in ADO interface.

3.1. Importing ADO Type Library

ADO type library including ADO class's definition is stored in msado15.dll, it describes automatism and COM vtable interfaces in C++. In order to use ADO in Visual C++ program, we must import ADO type library firstly by #import statement. Using no_namespaces property in #import statement can let us use ADO type without scope limitation. But it may cause name confliction with EOF key words. In order to solve this problem, we can use rename property of #import statement. The codes are as follows:

```
#import "c:\program files\common files\system\ado\
msado15.dll" no_namespaces rename ("EOF", "adoEOF")
```

This statement assumes msado15.dll stored in "c:\program files\common files\system\ado" directory. If it is not in such directory, you should change to correct path. If using ADO in whole project, we should put this statement in stdafx.h file.

* This paper is supported by the fund of Hebei University of Science and Technology (Grand NO. xl200122)

3.2. Initializing OLE/COM

ADO library is a group of COM dynamic link libraries. This means you must initialize OLE/COM before using ADO in program. In MFC program, it is a better method to initialize OLE/COM in InitInstance member function of application class.

```
  //the project’s name is ADOTest
 BOOL CADOTestApp::InitInstance ()
 { …
       AfxOleInit (); //Initializing OLE/COM
…}
```

3.3. Connecting Database Using ADO

In Visual C++ we need define a _ConnectionPtr object, instantiate the object and execute the object’s Open function to connect database by ADO. Open function is used to open a data resource and create a session. The prototype of the method is as follows:

HRESULT _ConnectionPtr: Open (_bstr_t ConnectionString, _ bstr_t UserID, _bstr_t Password, long Options);

ConnectionString parameter is a string including the information of connection. UserID parameter is the name of login user. Password parameter is the password of the user. Options parameter is options of the connection, which is used to specify connection object’s authority of updating data in database. If UserID and Password is not need or included in ConnectionString parameter, the two parameters can be a zero length string (not NULL). We should execute _ConnectionPtr object’s Close method after finishing the operation of database. The following example of creating a SQL SERVER connection is given.

```
 //declare a connection variable
 ConnectionPtr m_pConnection;
 // create a connection object
 m_pConnection.CreateInstance (_uuidof (Connection));
 //open database connection
m_pConnection->Open("Provider=SQLOLEDB.1;Passwo
rd=wjy;Persist Security Info=True; UserID=wjy;Initial
Catalog=wang;Data ource=wjy","","",adModeUnknown);
 …
m_pConnection-> Close (); //close the connection
```

3.4. Opening Recordset by ADO

We should execute _RecordsetPtr object’s Open method in order to open a Recordset by ADO in Visual C++. Its prototype is as follows.

HRESULT Recordset15::Open (const _variant_t & Source, const _variant_t & ActiveConnection, enum CursorTypeEnum CursorType, enum LockTypeEnum LockType, long Options);

Source parameter is a query string, such as SQL statement. ActiveConnection is a created connection (needs Connection object point to create a _variant_t object). CursorType parameter is the type of cursor. LockType parameter is locked type. Options parameter is the data type of SQL command. We should execute _RecordsetPtr object’s Close method after finishing the operation of Recordset. The main codes of opening Recordset by m _pConnection object are as follow:

```
 //declare a Recordset variable
 _RecordsetPtr m_pRecordset;
 //create a Recordset object
 m_pConnection.CreateInstance (_uuidof (Connection));
 //open a Recordset
m_pRecordset->Open(“select xh,xm,picture from student”,
variant_t((IDispatch*)m _pConnection, true), adOpenStatic,
adLockOptimistic,adCmdText);
```

3.5. Accessing BLOB Data in Recordset

We can use m_pRecordse object’s GetCollect and PutCollect method to store or load data of the ordinary type fields such as integer, float, string and date etc.

Because common fields store fixed length data such as integer, float etc. type data or variable length string, however BLOB field can store variable length binary data, we can’t store it as common field. We need store BLOB data using AppendChunk method of Field object. The prototype of the AppendChunk method is as follow:

HRESULT AppendChunk (const _variant_t & Data);

Its function is to Append data to a large text or binary data Field or Parameter object. The first AppendChunk call on a Field object writes data to the field, overwriting any existing data. Subsequent AppendChunk calls append the new data to existing data. The pivotal problem is how to put the BLOB data to VARIANT structure. We can read the BLOB data stored in graphics files or audio files etc. to a buffer by ReadHuge method of MFC Cfile class. There is only SAFEARRAY in COM types can express byte array. Although VARIANT can’t express it, we can get byte array data though two conversions. Main process is as follow:

(1) Read BLOB data stored in files to buffer

```
 CFile InputFile; //define file object
 //open the file object
 InputFile.Open ("ZhangSan.bmp", CFile::modeRead)
 char* m_pBMPBuffer;
 //get the length of BLOB data
 long m_nFileLen= InputFile .GetLength();
 //allocate buffer
 m_pBMPBuffer=new char [m_nFileLen+1];
 //read BLOB data to buffer
 InputFile .ReadHuge (m_pBMPBuffer,m_nFileLen);
```

(2) Create the dimension and bound of SAFEARRAY

```
//define one dimension SAFEARRAY
SAFEARRAYBOUND        rgsabound [1];
//set the low bound of SAFEARRAY is 0
rgsabound [0].lLbound = 0;
//set the element count of SAFEARRAY is
//m_nFileLen
 rgsabound [0].cElements = m_nFileLen;
```

(3) Create SAFEARRAY variable

```
SAFEARRAY    *psa;
//the SAFEARRAY is one dimension and the
//element data type is byte, the property of the
//bound is in rgsabound
psa = SafeArrayCreate(VT_UI1, 1, rgsabound);
```

(4) Get and release the pointer of buffer in SAFEARRAY

```
//get the pointer of buffer and lock it
SafeArrayAccessData(psa,(void**)&m_pBMPBuffer);
 //release the lock
 SafeArrayUnaccessData (psa);
```

(5) Encapsulate SAFEARRAY to VARIANT variable

```
        VARIANT    varBLOB;
        //set the data type of the SAFEARRAY is byte array
        varBLOB.vt = VT_ARRAY | VT_UI1;
         //put psa to varBLOB’s parray
         varBLOB. = psa;
```

Thus we completed the creation of a VARIANT variable. Then we should execute the AppendChunk method of _RecordsetPtr ‘s Field object to move BLOB data to BLOB field in database.

```
         //store the value of varBLOB to picture field
        m_pRecordset->GetFields () ->GetItem ("picture") ->
        AppendChunk (varBLOB);
```

To read the value of BLOB variable we should execute the GetChunk method of field object, the prototype of the method is _variant_t GetChunk (long Length). The function of the method is to return all or part of the contents of a large text or binary data Field object. The method Returns a VARIANT type variable. After getting the byte pointer pointed to the VARIANT’s parray by SafeArrayAccessData method, we can do some manipulations such as export and display BLOB data using the byte pointer. The main codes are as follow:

```
        //get the length of picture field
        long lDataSize=m_pRecordset->GetFields()->GetItem
    ("picture")->ActualSize;
        _variant_t        varBLOB; //define export variant
        //get BLOB data
        varBLOB=m_pRecordset->GetFields()->GetItem("pict
       ure")->GetChunk(lDataSize);
        char* pBuf;
        //allocate buffer
        pBuf =new char[lDataSize+1];
        //judge if the type of varBLOB is byte
        if(varBLOB.vt == (VT_ARRAY | VT_UI1)
        {      char* pBuf=NULL;
               //get the pointer of varBLOB to pBuf and lock
         data
               SafeArrayAccessData(varBLOB.parray,(void **)
               &pBuf);
        //export BLOB data of current picture field to OutFile
               OutFile.WriteHuge(pBuf,lDataSize);
               //release the pointer and unlock buffer
                SafeArrayUnaccessData(varBLOB.parray);
    }
```

Through above discuss, we find out the basal working principle of accessing BLOB data in database by Field object’s AppendChunk or GetChunk method of ADO’s _RecordsetPtr object in Visual C++, which is shown in fig 1.

The method based on ADO technique, so it has high efficiency, rapid speed and widely application.

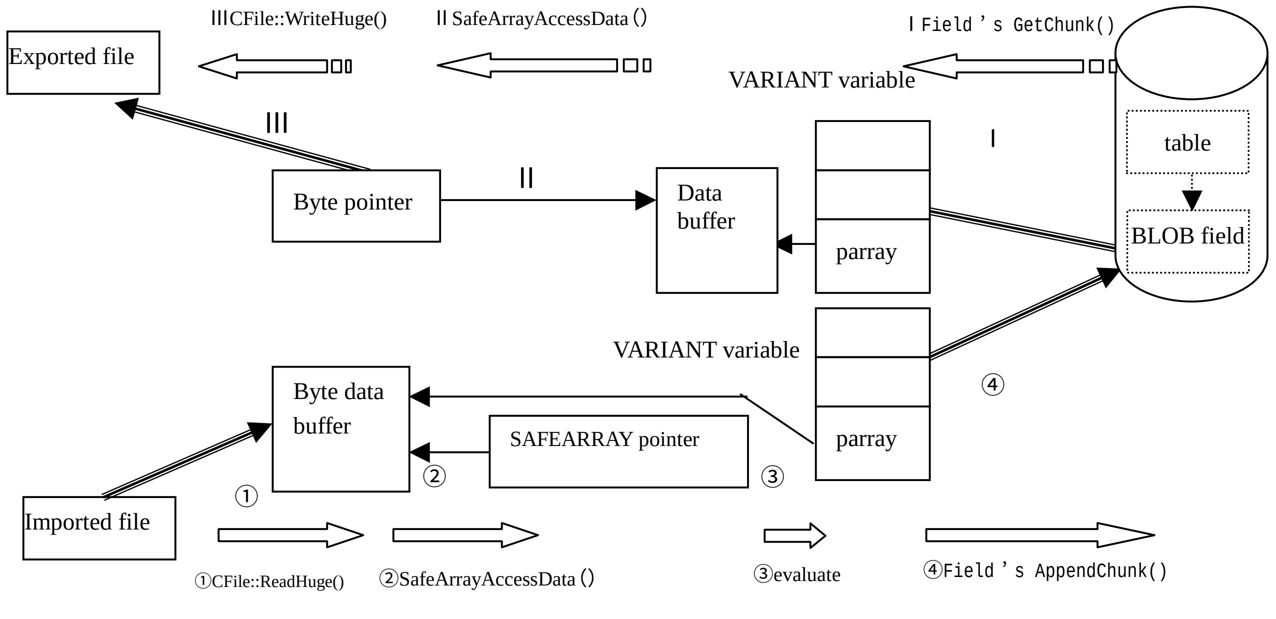


Fig. 1 the basal working principle of accessing BLOB data in database by internal objects of ADO

4. CONCLUSIONS

This article discusses a simple and efficient method to manipulate BLOB data in database by ADO in Visual C++, which provides an efficient approach for the developing of multimedia information systems. The idea of the method to manipulate BLOB data in database by ADO in this article can extend to some other program languages who support COM such as Visual Basic, ASP (Active Server Page), DEPHI etc. There are only small distinguishes between those program languages to realize the method, so the work we should do is to describe it in corresponding language.

5. REFERENCES

[1]. Microsoft Visual C++ 6.0 MFC Library Reference. Microsoft Press, August 1998

[2]. Beck Zaratian. Microsoft Visual C++ 6.0 Programmer’s Guide. Microsoft Press, August 1998

[3]. Jim,Maloney. Distributed COM Application Development using Visual C++6.0. Tsinghua Press,

Beijing, 2000.9 (in Chinese)

[4]. Zhang hua,Li zhi-zhu.Access Multimedia Database by ADO Component. Application Research Of Computers 2001, Vol.18 No.01(in Chinese)

[5]. Ren dong,Liu lian-zhong.Research on Information Manipulation Technique about Image and Binary Text in a JSP Web Site, Application Research Of Computers,2002,5:155-158(in Chinese)

[6]. Wang cong-hua. The Instance Anatomy of Accessing Database by ADO, Application Research Of Computers, 2002,5:159-160(in Chinese)

[7]. Cha feng. ADO of Visual C Applied in Constructing a Comprehensive Platform for Querying Heterogeneous Data Sources. MICROCOMPUTER APPLICATIONS, 2002 Vol.18 No.5 (in Chinese)

Qin Min (1972-), female, is a lecturer of Hebei University of Science and Technology. She graduated from Hebei University of Technology with specialty of Electrical Engineering & Automation in 1995, gained master degree from Hebei University of Technology in 1999. She has published three books, over 10 Journal papers. Her research interests are in Network and Database application technology.

Research on How to Connect Database in FORTRAN

Xia Hongxia , Jin Peng, Yuan JinLing
Computer Science, Wuhan University of Technology
Wuhan, Hubei (Province) 430070, China
Email: boolking@mail.whut.edu.cn **Tel.:** 027-87211983

ABSTRACT

In this paper, the author describes how to use a database in FORTRAN language. The key is how to use FORTRAN and C/C++ mixed language programming to make FORTRAN access database easily. The other is how to access database through ODBC or ADO.

Keywords: FORTRAN, database, ADO, ODBC, COM, DLL.

1. INTRODUCTION

FORTRAN language is famous for its powerful function in the region of mathematic calculation, so most of program in scientific and engineering calculation is written in FORTRAN in the past. Though FORTRAN is not fast than C language, and not easier using than some script language like Perl, FORTRAN is still widely used for its fully optimized function library and its ability to fulfill all kinds of mathematic calculation. And, as well as the developing of the new standard of FORTRAN, this status is changing.

At the same time of the increasing requirement of mathematic calculation, the amount of data of result and the data to be calculated is increasing much faster. How to keep this data becomes the most requirements and the best choice is Database Management System (DBMS).

In our project of health investigating and controlling of Shenzhen Civilization Center, we need the technique of accessing DBMS in FORTRAN language. In this project, we are cooperating with the experts of structure, and the experts process the data we gathered using their own FORTRAN program before they provide the result to use. Using the result, we estimate the stability of the web structure of Shenzhen Civilization Center rooftree by our intelligent system.

In this process, our structure experts use the FORTRAN language to make their static and dynamic calculation. But for the purpose of further investigation of the structure status continuity, we have to keep all the data in our database. So we have to find a method to access the database in FORTRAN language.

At the beginning, we chose the production of Canaima Software Co., f90SQL, which is a general database-accessing module. While we were evaluating this production, we find it is impossible to teach our structure experts to use database through f90SQL from FORTRAN in a limited period because this production is made for the programmers who have had the concept of database. Secondly, our structure experts are accustomed to use files as the intermediates of data transferring. If we fix the program to adapt the database accessing, there will be a lot of work for us. Thirdly, the price of f90SQL is much higher than what we can accept.

Because of lack of the information in this field, we have to research by ourselves. In the end, we find some method to achieve our goal. We make some C/C++ functions to replace the former FORTRAN subroutines used to read/write files, and accessing the database in these C/C++ functions. So the structure experts can access the database in the same way to access files.

This scheme should be separate into two steps, the first is how to link FORTRAN and C/C++ together, and the second is how to access the database. For the first step, we can link FORTRAN and C/C++ together using mixed language programming and dynamic link library, and for the second step, we can access the database using ODBC and ADO.

As the example, we provide two of these methods to explain the whole process.

2. ODBC DATABASE ACCESSING USING FORTRAN AND C MIXED LANGUAGE PROGRAMMING

Introduction

ODBC is a general database accessing method provided by Microsoft, it is a kind of interface of C language, so it makes it possible to access various database management systems from the same interface. ODBC makes the application can be independent on any DBMS. Users can access database through a component called driver. ODBC is called in the way of Windows API, which is convenient for the C language.

We will only describe the process of FORTRAN and C mixed language programming, and the ODBC can be referenced to MSDN.

Implement

In the mixed language programming, the most important is that the caller and called function should obey the same rule, or the confliction will fail the compile process. The rules include the naming convention, the calling convention, parameter passing convention and return value convention.

The naming convention determines how a language alters a symbol name as it places the name in an .OBJ file. This is an issue for external data symbols shared between modules as well as for external routines. Parameter names are never affected. The reasons for altering the name include case sensitivity or lack thereof, type decoration, and other issues. If naming conventions are not reconciled, the program cannot successfully link. You will receive an “unresolved external” error.

The default naming convention of FORTRAN is to uppercase the symbol names, so if calling a default FORTRAN subroutine from VC++, the caller should make the call from an uppercased name. And if the caller in VC++ makes the call from a lowercased name, the FORTRAN program should alter the name to lowercase using the property of C and STDCALL. But if the caller uses mixed case name to make the call, the FORTRAN must use the ALIAS property to solve the naming confliction. All the naming convention is shown in the next page:

Tab.1 Comparison of the Naming Convention

language	Attributes	Name translated as	Case of name in OBJ file
Fortran	CDEC$ ATTRIBUTES C	_name	All lowercase
Fortran	CDEC$ ATTRIBUTES STDCALL	_name@n	All lowercase
Fortran	default	_name@n	All uppercase
C	cdecl (default)	_name	Mixed case preserved
C	__stdcall	_name@n	Mixed case preserved
VC++	default	_name@@decoration	Mixed case preserved

In the other side, we can use the method shown below to solve the problem:

```
#ifdef __cplusplus
extern “C” {
#endif
void Pythagoras( float a, float b, float *c)
{
    *c = sqrt( a * a + b * b );
}
#ifdef __cplusplus
}
#endif
```

The calling convention determines how a program makes a call and where the parameters are passed. In a single-language program, calling conventions are nearly always correct, because there is one default for all modules and because header files enforce consistency between the caller and the called routine. In a mixed-language program, different languages cannot share the same header files. It’s easy to link FORTRAN and C modules that use different calling conventions, and the error isn’t apparent until the bad call is made at run time, causing immediate program failure. Therefore, you should check calling conventions carefully for each mixed-language call.

The following table summarizes how C and FORTRAN calling conventions work (Tab.2).

In C and C++ modules, you can specify the FORTRAN calling convention by using the __stdcall keyword in a function prototype or definition. The __stdcall convention is also used by window procedures and API functions. For example, the following C language prototype sets up a function call to a subroutine using the FORTRAN calling convention:

Tab.2 Comparison of Calling Convention

Calling convention	Parameter passing	Stack cleared by
C/C++	Pushes parameters on the stack, in reverse order (right to left)	Caller
Fortran (__stdcall)	Pushes parameters on the stack, in reverse order (right to left)	Called function

```
extern void __stdcall fortran_routine (int n);
```

Instead of changing the calling convention of the C code, you can adjust the FORTRAN source code by using the C attribute, enclosed in brackets ([ ]). For example, the following declaration assumes the subroutine is called with the C calling convention:

```
SUBROUTINE CALLED_FROM_C [C] (A)
  INTEGER*4 A
```

It should be clear that calling conventions need only agree between individual calls and the called routines, and that the conventions must be the same: Both caller and called routine must use the C/C++ convention or both must use the __stdcall convention (the FORTRAN default).

Note: In programs written for the graphical user interface of Windows, PASCAL, WINAPI, and CALLBACK are all defined with __stdcall. But the C language default is still cdecl.

The following table summarizes how to specify calling conventions. You can always specify calling conventions explicitly rather than relying on the default, which is a good technique for mixed-language programming.

Tab.3 Comparison of Calling Convention

Language	C calling convention	Fortran calling convention
C/C++	cdecl (default)	__stdcall
Fortran	C attribute	STDCALL attribute (default)

When using mixed language programming, which means to use object files compiled in VC++ and Fortran separately and link these object files together later, it is important to use the same single-threaded or same multithreaded library at both side in one time, or it will lead to some error when linking.

3. FORTRAN INVOKE THE DYNAMIC LINK LIBRARY WRITTEN BY C++ TO ACCESS DBMS

We recommend a font of 9 points or greater. This document is set in 9-point Times, that means all the characters are set in the 9-point Times, except specifically pointing out. If absolutely necessary, we suggest the use of condensed line spacing rather than smaller point sizes. Some technical formatting programs print mathematical formulas in italic type, with subscripts and superscripts in a slightly smaller font size. This is acceptable.

Introduction

Though the mixed language programming can solve some problem, the code have to be re-compiled when the code are changed because the code is link statically. And if the amount of code is big and duplicated many times, it will be a waste of resources while dynamic linking can solve all the problems.

As well as the rise of COM, Microsoft abandoned ODBC, and turn to ADO. And because of the convenient and accessing the remote DBMS seamlessly of ADO, it becomes the best choice to develop the database application.

FORTRAN supports COM at same time of developing. But in spite of the complication of COM itself, using COM directly to develop an application should manage the reference count of a COM object manually. But in C++ language, we can use the statement of #import to generate the smart pointer to manage all the trivial things automatically, and make our work easier. So we decide to use C++ language to make a dynamic link library to encapsulate the accession to the database and use this dynamic link library from FORTRAN language.

Implement

To use ADO in C++, we should use “#import” to import the type library of ADO:

```
#pragma warning( disable:4146 )
#import "c:\program files\common
files\system\ado\msado15.dll"\ no_namespace
rename("EOF","adoEOF")
```

Then, construct the smart pointer to the ADO object:

```
_ConnectionPtr m_pConnection;
_RecordsetPtr m_pRecordset;
```

and construct the function to access the database for the Fortran language:

```
extern "C"
{
extern int __declspec(dllexport)    Connect2DB()
{
return 1;
HRESULT hr;
try
{
hr = m_pConnection.CreateInstance("ADODB.Connection");
//create the Connection object
if(SUCCEEDED(hr))
{
m_pConnection.CreateInstance(__uuidof(Connection));
m_pConnection->Open("Provider=SQLOLEDB;User\
ID=name; Password=pass;Initial Catalog=REALTIME;Data\
Source=127.0.0.1","","",-1);
//connect to the database
}
}
catch(_com_error e)//catch the abnormal
{
abort();
}
}
}
```

The database read-write and database disconnection function is like this.

The usage of dynamic link library in FORTRAN is shown below:

First import kernel32 module to use some Windows API:

```
use kernel32
```

then, declare a “INTERFACE” and a “pointer” to invoke the function in the dynamic link library:

```
INTERFACE
        SUBROUTINE rtn_DBC()
        END SUBROUTINE
END INTERFACE
POINTER (q,rtn_DBC)
```

Load the dynamic link library, and get its export-function address, and we assume its name is test.dll:

```
INTEGER p
p = loadlibrary("test.dll"C)
IF (p == 0) THEN
type *, "Error occurred opening test.dll"
         type *, "Program aborting"
ENDIF
q = getprocaddress(p, "Connect2DB"C)
IF (q == 0) THEN
type *, "Error occurred finding Connect2DB in test.dll"
type *, "Program aborting"
ENDIF
```

Now, we can use subroutine rtn_DBC to connect to the database:

```
CALL rtn_DBC()
```

Finally, we should release the dynamic link library after we finished using it:

```
Freelibrary(p);
```

4. CONCLUSIONS

After our description, we can conclude that if we use the mixed language programming, it is not necessary to load the dynamic link library, and it seems same to write the normal FORTRAN program. But if you are not familiar with the whole procession, you will always get some errors while linking. If you use dynamic link library, OS executes the procession of linking automatically. So you should choose appropriate one.

Nevertheless, in either method, you have to pay attention to the matching of naming and calling convention. But we recommend you to chose the second method because the convenience of linking and the usage of ADO. To choose this method, you can improve the efficiency and save the time.

5. REFERENCES

[1] Jiang youzhi, Zhu xiaobo, Zhang shaoqiang, “The basic application of ADO in VC++”, Modern Computer, Vol.4,2003. （in Chinese）

[2] Zhou zhenghong, Yan guohong, Wu hongjuan, “A Study of Fortran and Visual C++ Programming”,Journal of Wuhan University, Vol. 4, 2001. （in Chinese）

Xia Hongxia is an associate professor. Her research interests are in database application technology, distance education and intelligent method.

Jin Peng is a graduate of Computer Science Department of WUT. His research interests are computer graphics and database application technology.

Methods of Applet Querying Database Through Servlet *

Qin Min, Wang Jingyang, Wang Jianxia, Zhang Xiaoming
College of Information Science & Engineering, Hebei University of Science and Technology
Shijiazhuang Hebei 050054, China
Email: qm@hebust.edu.cn **Tel.:** 0311-8613184

ABSTRACT

This paper firstly introduces the mechanism of Applet querying database through Servlet, then briefly illustrates several common methods, which include delivering query results by text stream, delivering query results by record set, delivering query results by Vector of objects and delivering query results by two-dimension Vector. Moreover it evaluates these methods. Finally it gives the codes of the method of delivering querying results by two-dimension Vector.

Keywords: Servlet, Applet, JDBC, Vector.

1. INTRODUCTION

At present we always use Java Applet and Servlet simultaneously to design web application program with multiple-tier structure. Applet provides convenient mechanism to establish powerful and dynamic interface while Servlet provides efficient approach for web servers or other application servers to process requests. The best approach to develop enterprise-level Java application under Java2 platform is to use Applet, html in front while to use Servlet and other components in back [1]. The key is how Applet at client access database through Servlet at server.

2. MECHANISM OF APPLET QUERYING DATABASE THROUGH SERVLET

Applet can communicate with Servlet over HTTP socket connections [2]. If Applet query database through Servlet, the following will be done: firstly the Applet send query statement to Servlet, then the Servlet query database by JDBC and return the results to Applet. Its working mechanism is shown as figure 1.

The Applet sends query statement to the Servlet by sending a GET method. The Applet simply has to open a connection to the specified Servlet URL. Once this connection is made, then the Applet can get an output stream or input stream on the Servlet[3].

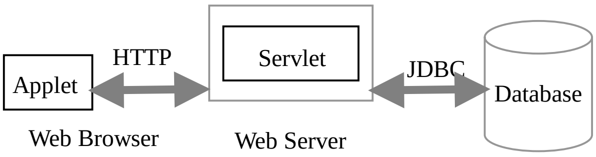


Fig.1 Applet query database through Servlet

The Servlet accesses database through JDBC, then puts the querying result into output stream to transmit to Applet. The output stream can be text stream or object stream here [4].

3. METHOD OF APPLET QUERYING DATABASE THROUGH SERVLET

There are some common methods of Applet querying database through Servlet, which include delivering query results by text stream, delivering query results by record set, delivering query results by Vector of objects, or delivering query results by two-dimension Vector. Here are the introductions of these methods respectively:

3.1 Delivering query results by text stream

The simplest method to exchange information between Applet and Servlet is through HTTP text stream. Servlet will write records of result set into text stream line by line while Applet will read them from input stream on line sequentially.

The main weakness of this method is that Applet would not understand the data information directly rather to convert them into a useful format. But the conversion will become uncontrollable when we try to process a more complex data or object.

3.2 Delivering query results by record set

It is a better way to deliver query result by record set. Servlet will execute query after receiving Applet’s query statements and construct CachedRowSet object with result set. Then Servlet put the CachedRowSet object into object output stream to send to Applet. Applet will read the object from object input stream and then handle it as operating record set.

Because what Servlet send to Applet is record set, Applet terminal could operate it using field name, which differs from text stream, which Applet reads values by line and does not know its content. The readability of the program is be improved, however you have to download CachedRowSet and netcape package and adjust safety policy at Applet terminal.

3.3 Delivering query results by Vector of objects

Another common method is to deliver query result set using Vector of objects. Firstly a certain class should be defined corresponding to a table or several tables, then records of query result is encapsulated into objects of this class, Finally these objects are put into the Vector and this Vector are put into object output stream to be sent to Applet. When the Servlet returns the Vector of objects, there is no need to iterate through the Vector and serialize each object individually. The Servlet can simply serialize the entire Vector in one step, since the class java.util.Vector also implements the java.io.Serializable interface. Applet terminal will read the object from object input stream, thus Applet can get any object in the Vector to operate.

* This paper is supported by the fund of Hebei University of Science and Technology (Grand NO. xl200122)

Because object has been used, this method overcomes the defect that Applet can't directly understand the data information in text stream. Properties value of the object is equal to field value. No need type conversion, No need install new class package. But this method is inconvenient because it must define a class for each query. It is impossible for Applet to use one Servlet for all queries because different query needs different objects.

3.4 Delivering query results by two-dimension Vector

In order to explain this method, we have established an Oracle database. There is a table named type in this database, it has two fields, one is named typeid (NUMBER) and the other is named typename (VARCHAR).

1) Implementation

Firstly Applet defines a SQL query statement and then delivers it to Servlet with parameters. Secondly Servlet reads and executes it, then put every field of a record in result set as an element into a Vector by sequence. Thus every record will become a Vector. Thirdly all of these Vectors as element will be putted into another Vector. Finally this Vector will be send to Applet. Figure 2 shows this delivery course.

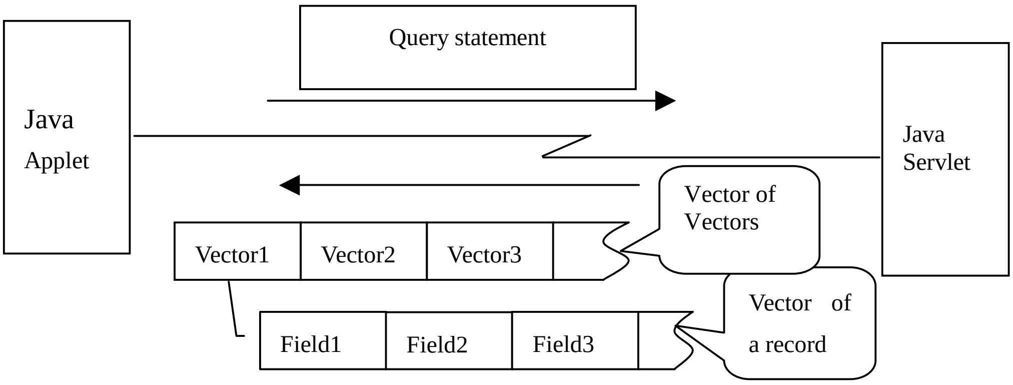


Fig.2 Delivering query results by two-dimension Vector

The mainly code of Servlet as follow:

```
//Create a connection to database
Class.forName("Oracle.jdbc.driver.OracleDriver");
Connection cnn = DriverManager.getConnection("jdbc:
Oracle:thin:@127.0.0.1:1521:hbkj", "system", "manager");
//get the query parameter  and execute it
String sql = request.getParameter("message");
stmt = cnn.createStatement();
ResultSet  resultSet = stmt.executeQuery(sql);
ResultSetMetaData  metaData = resultSet.get-
MetaData();
//Create object output stream
ObjectOutputStream out = new ObjectOutput
-Stream(response.getOutputStream());
// Put resultSet into two-dimension Vector
Vector v = new Vector();
while (resultSet.next()) {
    Vector row = new Vector();
    for (int i = 1; i <= metaData.
        getColumnCount();i++){
        row.addElement(resultSet.getObject(i));
     }
    v.addElement(row);
  }
//Put Vector in object output stream
out.writeObject(v);
stmt.close();
```

At Applet terminal input stream is ObjectInputStream, the Applet will read the object from ObjectInputStream and send them into Vector, in this time every element in the Vector v standard for a record in the result set. Then Applet will read an element from v and write to another Vector vRow, in this time every element in vRow standard for a field value in a record. Because there is one-to-one correspondence between element in the vRow and a field value of a record, the value of a certain field in this record can be got as soon as Applet get the elements in Vrow through index. The Applet's main code is as follows:

```
//Create the connection to Servlet, deliver the query statement
//by parameter "message", chatURL is the address of web
//server
messageSql="select typeid typename from type";
String queryString = "VectorServlet?message="+
URLEncoder.encode(messageSql);
URLConnection connect=(new URL(chatURL,
queryString)).openConnection();
connect.connect();
//Create ObjectInputStream, read object from ObjectInput
 //-Stream, put this object into a Vector
ObjectInputStream in = new ObjectInputStream
(connect.getInputStream());
Vector v=new Vector();
v=(Vector)in.readObject();
if (v.size() != 0) {
//every element in the Vector v standard for a record in the
//result set, read a element from v and write to another Vector
//vRow, every element in vRow standard a field value for a
//record
for (int i = 0; i < v.size(); i++) {
    Vector vRow = (Vector) v..get(i);
     Int temtypeid = Integer.parseInt(vRow.get(0).toString());
     String temtypename = vRow.get(1).toString();
}//end for
}//end if
```

2) Characters

In this method, SQL query statement is sent to Servlet, Servlet will return query result with 2-dimension Vector. Different query statements could use the same Servlet, but this method has not encapsulate record into object, so you must remember every field's type and meaning, At Applet terminal type-conversion is needed after the field value is got. Furthermore this method must use index and cannot use field name to get field value. If query statement has been modified or field sequence has been changed, many codes in the Applet have to be changed.

4. CONCLUSIONS

This paper gives a brief introduction of several methods of Servlet returning query result set to Applet, and compares their characteristics. Especially the implementation of Servlet returning query result set to Applet by two-dimension Vector has been discussed. Each of these methods has its advantages and disadvantages, programmers should choose appropriate one according to different situations.

5. REFERENCES

[1]. Harvey M.Deitel, Paul J. Deitel. Advanced Java 2 Platform How to Program[M]. Prentice Hall press, 2001

[2]. Tan Jun-shan, Wu Chang-sheng. The Communication Technique between Applet and Servlet[J], Information Technology, 2002,Vol.26 No.12

[3]. George Reese. Database Programming with JDBC and JAVA, Second Edition, China electric power publishing house, March, 2002 (in Chinese)

[4]. TANGJian-pin, LIUXiao-ling. The Protection of the Web Page of Document Based on the Communication Techniques between JavaApplet and Servlet[J]. Acta Scientiarum Naturalium Universitatis Nei Mongol, 2003, vol.34, No.1 (in Chinese)

Qin Min (1972-), female, is a lecturer of Hebei University of Science and Technology. She graduated from Hebei University of Technology with specialty of Electrical Engineering & Automation in 1995, gained master degree from Hebei University of Technology in 1999. She has published three books, over 10 Journal papers. Her research interests are in Network and Database application technology.

The Research of an Inventory Control Information System Based on the Internet

Xiao Hanbin, Mo Lili, Zeng Xiangfeng
Logistics Department of Wuhan University of Technology
Wuhan, Hubei, 430063, China
Email: xhb@mail.whut.edu.cn **Tel:** 13507134385

ABSTRACT

Most produce enterprises in our country face with problems overstocking and high cost to maintain inventory. In order to deal with problems mentioned above, this paper develops an integrated inventory control information system according to principles of MRP and JIT and adopts B/W/D/C mode. The system consists of inventory management module, purchase management module, supply agent analysis and Web query module. We put the first three modules in C/S interface to meet the minority (functional departments) and put the last module in B/S interface to meet the majority (clients). Both interfaces use Web database in common. In the system, it utilizes Delphi 6 as C/S interface foreground development tool, Visual Interdev 6.0 as Web interface development tool and Microsoft SQL Server 2000 as database in bottom.

Keywords: Inventory Control, Information System, MRP, JIT, B/W/D/C Mode, Internet.

1. INTRODUCTION

At present, most produce enterprises in our country face with problems overstocking and high cost to maintain inventory. It costs more than the one in the same kind of foreign enterprises. For example, plan of purchase materials is not rational, control to semi-manufactured goods is not enough, products are overstocking and choice to supply agent is often affected by subjective factors. In order to deal with problems mentioned above, this paper develops an integrated inventory control information system according to principles of MRP (material requirement planning) and JIT (just in time). It is based on enterprises' existing Intranet and adopts the B/W/D/C mode. The system consists of inventory management module, purchase management module, supply agent analysis and Web query module. We put the first three modules in C/S interface to meet the minority (functional departments) and put the last module in B/S interface to meet the majority (clients). Both interfaces use Web database in common. In the system, it utilizes Delphi 6 as C/S interface foreground development tool, Visual Interdev 6.0 as Web interface development tool and Microsoft SQL Server 2000 as database in bottom.

2. INVENTORY CONTROL METHODS BASED ON MRP AND JIT

2.1. 1MRP Analysis Of Inventory Control

The basic principle of MRP (material requirement planning) is to produce and purchase according to real requirements of materials. MRP process: MRP's data processing adopts the pattern from top to bottom. It utilizes a main produce schedule to make certain requirements of final items and adopts material lists and advance date to dispart them into requirements of different produce process step by step. The requirements include all kinds of assemble items, accessory, outsourcing products and raw and processed materials. Its basic process is to do treatments circularly step by step, item by item. Begin with requirements of final items and end up with final required outsourcing products and raw and processed materials, so it can get total requirements of outsourcing products and raw and processed materials. And then, find out net requirements based on existing inventory and safety inventory. Later, obtain its material requirement plan according to purchase period. At last, give order commands or produce commands.

2.2. JIT Analysis Of Inventory Control

The basic principle of JIT (just in time) is to produce products according to required quantity and time. In other words, system's process pattern, quantity and time are decided by the next working procedure. Supply agent's consignment patterns, quantity and date are decided by requirements of produce schedule. So it can satisfy requirements of time and quantity no matter in each working procedure and stage of produce process, or in goods' transfer and supply agent's consignment.

2.3. Integrated Mode Based On MRP And JIT

MRP's advantage is that it can realize centralized information management and control reasonless inventory effectively while JIT's advantage is that it can shorten produce response time and decrease goods inventory effectively. Synthesizing MRP and JIT's advantages can deal with enterprises' overstocking effectively. That is to say, based on sale volume and market forecast, use MRP as the plan of whole system and control to input materials while use JIT as control goods and produce pace. The flow chart of integrated mode is illustrated in Figure.1.

3. WHOLE DESIGN OF INVENTORY CONTROL INFORMATION SYSTEM

3.1. Inventory Control System Mode

In this paper, we choose a four-layer structure of Web database (B/W/D/C) as inventory control information system mode. (B/W/D/C) is improved on three-layer structure of Web database. In other words, we put C/S structure into database server, so we can regard application system as a system made up of 4 modules. The first module is a layer to realize user interface in browser and it is also called client browser. It mainly offers browse interface or execute application interface. The second module is a Web application and function layer and its function is data processing. It is called Web server/database client computer or network integrated server and it not only control user's visit to application, but also as the database server client computer of the third layer to run database server. The third module is a function layer to store data and it is also called database or other host computer servers. It includes restriction to all data. The fourth module is a client interface function layer and it is also called database client interface. Its mainly function is to manage database such as data insert, delete and modify and it uses C/S structure to connect with database server.

The four-layer structure of Web database (B/W/D/C) has advantages as follows: first, it exerts B/S and C/S structure's advantages adequately and has deep consideration to user, that is to say, it not only ensures user's manipulation expedient, but also makes system more simple, more flexible and easier to manipulate. Second, information release interface adopts B/S structure, so all operation systems of client computer and servers are unified completely, and then all problems in client interface can be dealt with. Third, database interface uses C/S structure and is connected by ODBC/JDBC, so it can deal with problems from systems only use C/S structure and overcomes many detects of B/S structure.

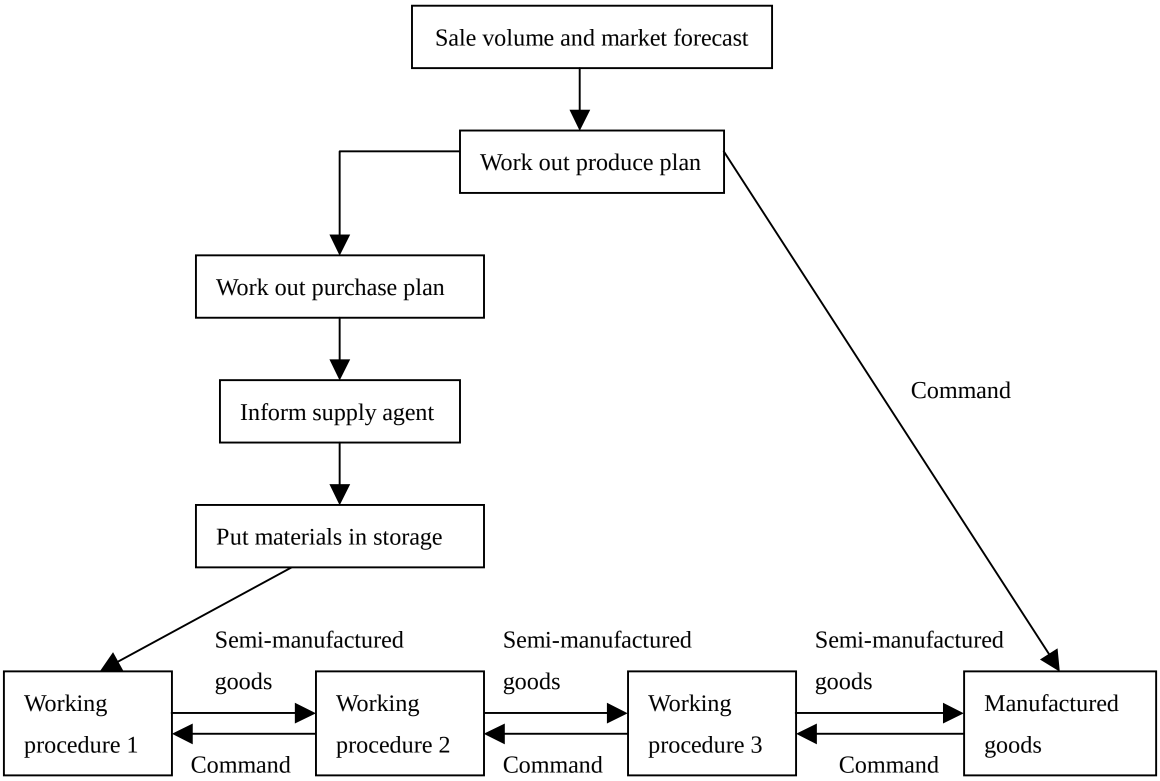


Figure. 1 Inventory information flow chart based on MRP and JIT

3.2. Choice Of Development Tool

3.2.1. Choice of database in bottom

Database is the basis of whole inventory control information system and as core of data processing. It is closely related to application, so the choice of database has great significant to whole system. This paper chooses the SQL Server as database development tool. SQL Server uses integrated distributed structure to manage whole enterprise's server. Using a management interface based on Windows not only control multiple servers simultaneous, but also can copy data, control server, diagnose, adjust and establish special database and realize remote management to many tasks. At last, SQL Server can greatly enhance system's performance and expansion by using parallel inner database. It also utilizes parallel structure's advantages to support large-scale database.

3.2.2. Choice of foreground development tool

Delphi is the preferred tool in developing conventional Windows application, database application and Internet application, so we choose Delphi as development tool of foreground in this system. Compared with other development tools, it has more advantages as follows: Delphi's IDE (integrated develop environment) is open completely, so clients can not only plan development environment adapting to their habits, but also can put other tools they design into Delphi's IDE. Delphi's database function is great and it supports ODBC (open database connection) and mode of client computer/server, so it can effectively operate kinds of database in local and large-scale distributed network. Delphi applies SQL into its database to let user develop more effective database by SQL.

3.2.3. Choice of web development tool

In this paper, we choose Visual InterDev 6.0 as development tool and it consists of whole design, database development, nodes management, satisfy development tool. It has advantages such as integrated develop environment, supporting compile interactive server application, powerful and integrated database tool, development tool of nodes management and content development, open and expansion.

3.3. Whole Design Of Inventory Control Information System

We have mentioned above that mixture mode of MRP and JIT can effectively deal with problems overstocking and unreasonable inventory structure and how to choose supply agent to effective control purchase cost, so we use MRP and JIT to design this inventory control information system. Whole structure of inventory control information system is illustrated in Fingure.2.

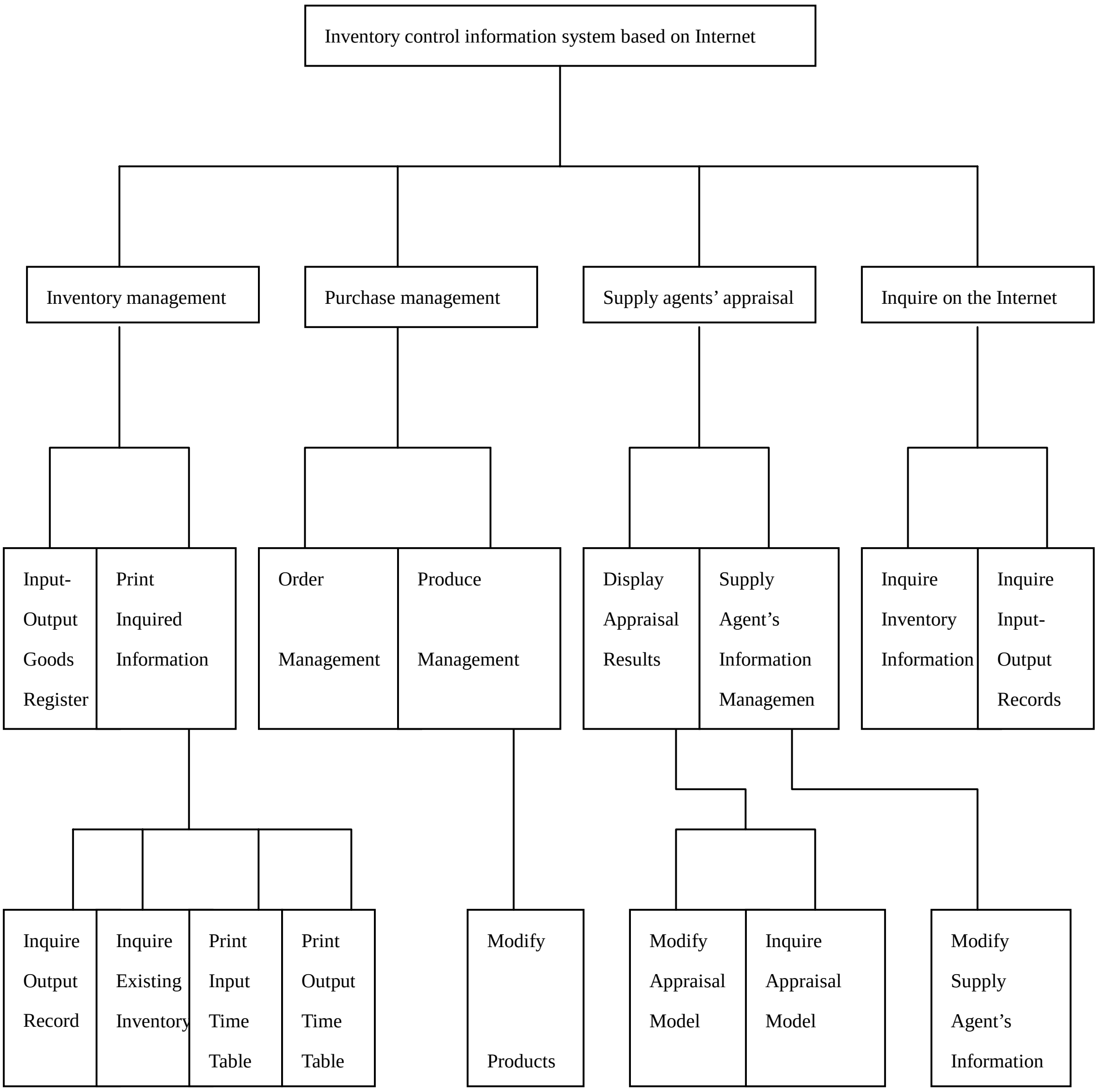

Inventory control information system based on Internet
Inventory management
Purchase management
Supply agents' appraisal
Inquire on the Internet
Input-Output Goods Register
Print Inquired Information
Order Management
Produce Management
Display Appraisal Results
Supply Agent's Information Managemen
Inquire Inventory Information
Inquire Input-Output Records
Inquire Output Record
Inquire Existing Inventory
Print Input Time Table
Print Output Time Table
Modify Products
Modify Appraisal Model
Inquire Appraisal Model
Modify Supply Agent's Information

Figure. 2 Whole structure of the system

According to different functions, this system can be divided into 4 parts: inventory management module, purchase management module, supply agent assistant module, information publish and inquire on the Internet module.

(1) Inventory management module: It mainly manages goods' input-output register and kinds of existing inventory conditions to keep integrity and real-time update of database. The managed data has 3 main bodies, they are raw and processed materials, parts, semi-manufactured goods and manufactured goods. Input-output goods has 2 forms: Input-output goods in the enterprise and Input-output goods between different enterprises.

(2) Purchase management module: This module is established according to MRP and its basis is real-time and exact inventory information. Purchase plan management is made through follow steps: first, ascertain products' hiberarchy. Then obtain correlative requirements according to produce plan. Next, find out total material requirement by synthesizing self-requirement material. At last, make final purchase plan after considering safety inventory based on existing inventory.

(3) Supply agent assistant module: Firstly, manage every supply agents' appraisal model and modify it constantly according to problems that enterprises face with in order to reflect enterprise's truly desire. Let experts give mark for enterprises in order to appraise every enterprise fairly and reduce subjective factors' influence in the course of appraise.

(4) Information publish and inquire on the Internet module: In enterprise's local area network, clients can inquire of enterprise's real-time and exact data according to their different purview. This method overcomes detects that utilizing file transmission results in low speed of information transmission and information scattered.

In this system, functional departments in enterprise use mode of client computer/server to deal with information and kinds of complicated affairs. The clients use mode of B/S to inquire affairs, in this way, it can decrease times of system maintenance.

4. CONCLUSION

This paper analyzes how to control produce enterprises' inventory by decreasing inventory and by control purchase cost. It brings forward a mode that mixture MRP and JIT to decrease inventory and a method that choose rational supply agent to control purchase cost. Based on 2 principles mentioned above, the author develops an inventory control information system based on B/W/D/C four-layer system mode and applied to Hubei Carbuilder. The information system consists of inventory management module, purchase management module, supply agent assistant module, information publish and inquire on the Internet module. Inventory management module, purchase management module and supply agent assistant module are used by minority, so they are put in C/S interface, while information publish and inquire on the Internet module is needed to process by Web and is visited by majority, so it is put in Web interface. C/S interface and Web interface use a database in common. The development of this system utilizes Delphi 6 as C/S interface foreground development tool, Visual Interdev 6.0 as Web interface development tool and Microsoft SQL Server 2000 as database in bottom.

This system utilizes Web to realize clients inquire more real-time and exact information and in a certain extent overcomes enterprises' information isolated problems and detects of files transmission such as low speed and half-baked information. It also can increase the information responsive speed of enterprises.

5. REFERENCES

[1]. Zhang Guofeng ect. Management information system. Bejing. China Machine Press. 2000.

[2]. Wang Shan ect. Database system. Beijing. Higher Education Press. 1991.

[3]. Liu Chunying,Zhang Guoxuan. Computer integrated manufacture system based on B/C/S. Application of micromputer. 2001.5:27-29.

[4]. Huang Yiguo ect. New ways of inventory management—MRP. Beijing. China Machine Press. 1987.

[5]. Song Hua ect. Modern logistics and supply chain management. Beijing. Economic and Management Press. 2000.5.

Xiao Hanbin is a professor, assistant head of equipment fault diagnosis lab and assistant director of logistics school of wuhan university of technology. In 2003, he obtained a doctor's degree. He has edited 3 books and published over 30 Journal papers. His research interests are in port machinery monitor state and fault diagnosis, logistics equipments' virtual diagnosis technology. He is the principal of over 10 tasks such as the research of MQ4033 gantry crane in Hainan, large-scale drums' synthetic experiment device.

Mo Lili is a postgraduate of logistics school of wuhan university of technology. She was born in 1982. In 2003, she obtained a bachelor 's degree. She has published 3 Journal papers and her research interests are in logistics plan and emulation, logistics engineering.

The Warehouse Management System Based on the Distributed Database

Xiong Guohai, Wan Junli
College of Electric Engineering & Information Science, China Three Gorges University
Yichang, Hubei, 443002 China
E-Mail: xiongguohai0096@sina.com.cn **Tel:** (86)13647174593

ABSTRACT

It is discussed that the warehouse management system is realized by using the features of the distributed database which is considered to be an independent one but scatters in different places. The principle of the distributed warehouse management system, ODBC interface of PowerBuilder and the distributed structure of ORACLE is analyzed. The primary function of the warehouse management system is illustrated. The efficiency of the enterprise's warehouse management has been improved because of the realization of systematism, standardization and automatization of the warehouse management.

Keywords: distributed database; warehouse management; network environment; data exchange; Interface

1. INTRODUCTION

The business warehouse management is often complex and troublesome. Because of the commodity's variety, the differences of booking ways, management ways and distributing ways, the differences of management system and the chart variety, we must use computer management. What's more, we must work out the suitable plans to realize systematism, standardization and automatization of the warehouse management. Then we can improve the business warehouse management efficiency. Here we use Powerbuilder to be the front developing tool of the database, and ORACLE to be the back database.

2. THE DISTRIBUTED DATABASE SYSTEM

The distributed database system is made up of a group of data which scatter in different computers on the network. Every node can be processed separately for local purpose and also can be used for the unitary purpose through the communication branch system.[1] A distributed database system must have the following characters: data spread on the different network node, and every node can have separate management ability with its own DBMS for local purpose; every node can cooperate very well to complete the greater whole system use. Just for the whole system use, although data scatter at the different places physically, the whole system users can't feel this logically. It seems that all data are from one local database.

3. THE DISTRIBUTED DATABASE TECHNOLOGY IN THE SYSTEM STRUCTURE

The company and its branches are in different area. They deal with their own business and data, but they also need to exchange and deal with the data. In the old management system, every branch only dealt with its own data and exchanged the statistic data with other branches in a certain time, so the data appeared disturbing. Distributed database system fits for the separate branches because nodes of the system can show the logical structure of the company. It allows each branch to store the often used data in the local place for local use so that communication cost can be reduced and responding speed can be improved. The distributed database can distribute date on many nodes, so it can improve the reliability of the system. If just one database and net can be used, part of the whole database can be used, the whole operation will not be stopped or the bottleneck won't appear just because of one database failure. The trouble can be recovered on single node. The software can be advanced separately by the node. Every part database has a data dictionary.

Distributed database has segment independence. If the certain relations are divided into segments, it can improve the process performance of the system. The data can be stored by segment on the places where we often use it, so most of the operation is part operation which can eliminate the information flux of the net. If segment independence can be supported by the system, the users will feel that the data doesn't seem to be the segment. Data copy independence means that the certain relation can be stored in different places through different store copies on the physical level. The copy independence should be supported by the system which supports copying data, so the users can work as it doesn't store the copy. Firstly, between the relation of the company and its branches, because branch data is the subset of the company business data, we use the paralleled segment and recomposed the relation through the calculation. Secondly, between company database server and Web database server, the data is divided according to its application function, so we use the perpendicular segment.

Data synchronous methods are transaction replication and merge replication according to the system needs. As the branch only stores its own data, the data management and analysis function is realized by the database server of the company. The branch can only send the new data to the company's database. We use transaction replication to have the synchronous business data. The branch's database is regarded part B, while the company's database is regarded part A. The branch's data were given a snapshot and the synchronous information is recorded in the branches database. Every branch that uses transaction replication has its own log read agent operating on part B. The transaction task for instant synchronous operates on part A with its own agent and is related to part B.

Transaction replication supports two types of object copy: table and storage procedure. In part B, definition of a part of the database or all the data can be copied with some storage procedures. When the branch's data is changed, log agent will deliver the information to the corporation's database. The replication based on the storage procedure has better function so that the communication current on the network may be reduced. Transaction log is used to watch the data change of the database. The transaction of the distributed query is supported by three kinds of query in the distributed database system: local query, distant query and overall query. The local query and the distant query is just related to the single node's data (local or distant) and their query optimization technology had been used in the concentrated database. The process and optimization of overall query is more complex because it is related to many nodes' data. The distributed transaction management is supported and includes recover and collision control. In the distributed system a transaction is related to code execution and renew in many places so that every transaction is made up of many "agents" which stand for the execution procedure of the given transaction in the given place.

4. ODBC INTERFACE OF POWERBUILDER AND THE DISTRIBUTED STRUCTURE OF ORACLE

Power Builder can be connected to various database through ODBC interface and it can be divided into four layers. [2] The first layer is Power builder application program which delivers SQL sentence to the database and obtains its result by transferring ODBC function. The second layer is ODBC driving management program which manages various database driving program, reflects original name of data to particular dynamic link of driving program, deals with the ODBC initialization, offers the ODBC entrance and checks the ODBC parameter for every driving program. The third layer managed by the second layer is driving program which deals with ODBC function transferring, delivers the SQL order to particular data source and returns the result to application program. The fourth layer is data source that is source of application program and it can be various database management systems. The distributed structure of ORACLE is base on two kinds of modes: Client/Server and Server/Server. By using these two kinds of modes, ORACLE offers the efficient data share in network environment, and becomes the DBMS which has the superiority over the traditional DBMS which uses the concentrated management modes.

5. THE DESIGN OF WAREHOUSE MANAGEMENT SYSTEM

The following functions of warehouse management system of a company should be fulfilled:[3] to input various warehouse management information including putting in, putting out, returning and requiring; to inquire, revise and maintain information; to finish the equipment purchase list; to monitor the goods in the warehouse by adding the highest store and the lowest store segment of the system; to manage the material requests of subordinate company; to manage the log; to help to use the warehouse management system. The function module graph of the system refers to graph 1. The function module of the warehouse management system in the subordinate company can be designed according to the actual case.

6. CONCLUTION

Distributed database as an independent one lies is in different places. The places may be installed in anywhere, from the local office to the other side of the world. Distributed database acts as a separate system in the Internet in which different nodes are connected together, which is very convenient for users. With the continuous popularity of Internet application, the distributed information management system based on the Internet is developing quickly. The development of the future enterprise information system will be developed in different systems and in different platforms. To provide a united program interface independent from the special database management system and a universal visiting way based on the common database is the objective of realizing the distributed database.

7. REFERENCES

[1]. Zhou Jianfang, A Sample of Constructing a Distributed Database by Sybase Replication Server, Journal of Computer Engineering and Application, No.18, September 2003, pp. 183-184. (in Chinese)

[2]. Zhang Shaozhong, Research on PB-based Accessing Heterogeneous Distributed Database System, Journal of Computer Engineering and Application, No.18, September 2003, pp. 199-201. (in Chinese)

[3]. He Xuhong, PowerBuilder Database System Open Out Guidance of Example, Bei Jing: Posts &Telecommunications Press, 2003. (in Chinese)

The Design and Realization about the Campus Information Management System Based on the Data Warehouse

Wang jianxia, Zhou wanzhen, Qin min, Fu Dong, Wang Jingyang
College of Information Science & Engineering
Hebei University of Science and Technology, Hebei Shijiazhuang 050054, China
Email: wang_jianxia@hebust.edu.cn **Tel**: 0311-8613336

ABSTRACT

The paper discusses the technology of data warehouse and data mining, states its application in the campus information management system, and provides the model of the management system based on the data warehouse technique. Finally this paper gives the developing steps of the campus information management system based on data warehouse.

Keywords: Data warehouse, DM, OLAP, Decision analyze, Management of campus

1. INTRODUCE

Now, the technique of the data warehouse is already applied in many realms. With the increasing of the data managed, it is necessary to use the strong function of the data warehouse in the campus information management to search and pick-up the worthy information. The campus information management system is a complicate system. The management work is tedious, complicity, long-time. This management system also requires high decision analysis. In recent years, it is increasingly urgent to establish the sound campus information management system with assistant decision in each college. During the period of information technique developed quickly, to establish the campus information management system with assistant decision will have the guidance and decision meaning to estimate student education quantity, to establish and to perfect the setting of courses, to improve the teaching method, to perfect practice teaching. In addition, it will have the very important and profound meaning to the existent and development of the school.

2. DATA WAREHOUSE AND DATA MINING TECHNIQUE

2.1 Data Warehouse

The data warehouse (DW) is a subject-oriented, integrated, non-volatile (stable) and time-variant collection of data in support of management's decisions. Here the subject is a standard to categorize data on higher level and each subject point is corresponding to a macro analysis realm. Integrated characteristic of data warehouse is that the data must be processed and integrated before the data enter to the data warehouse. First, the antinomy in the primitive data should be unified and the primitive data should be changed from application-oriented to subject-oriented. The stability of the data warehouse is that the data in warehouse is the history data, is not the data that produced by the day business handle. After the produced and gathered data enter the data warehouse, it is hardly modified. The data of warehouse is gathered from the different time data, and it requests the data in the warehouse to keep generally for 5 ~10 years in order to satisfy the proceeding trend analysis and making decision analysis. And the data in the warehouse should be marked clearly its history period.

2.2 Data Mining

Data Mining (DM) is picking up the knowledge interested by people from the data of the large database. The knowledge is value information that it is latent, unknown beforehand. The picking up knowledge can be concept, regulation, mode, rule, control, visual etc. The DM is decision sustain process of looking for some mode in some facts or observing data gather. The object of DM not only is a database, but also a document system, data gather or DW. The DM technique based on DW is to find the knowledge, which is not found from DW yet now. For those decision makers, the known information can use search, OLAP or other tool to obtain directly, but some relation or trend concealed in the large quantity data need using DM technique to be obtained. DM is a mainly application technique of DW. Using the technique of DM we can find out the real and worthy information and knowledge from the DW, Which let us make better quantitative analysis and prediction about the development of university.

3. THE WHOLE FRAME OF CAMPUS DATA WAREHOUSE SYSTEM

At present many universities enlarge their scale, the number of students in school is over ten thousands or even more, the number of teachers is over thousands. Thus the database with enormous data is formed. The common database system of campus have the campus information management, the student status management, the personal management, the educational administration management, the management of the teaching and scientific research work of teacher, and so on. Using data warehouse technology, we can reuse history data to management campus. Also it provides decision sustain system (DSS). The system would load the data in history database to the data warehouse by data collection and process, user use data in data warehouse by data mining (DM) and online analysis process (OLAP). Its realization process is shown as fig.1.

4. THE FUNCTION AND MODEL OF THE DATA WAREHOUSE IN CAMPUS INFORMATION MANAGEMENT SYSTEM

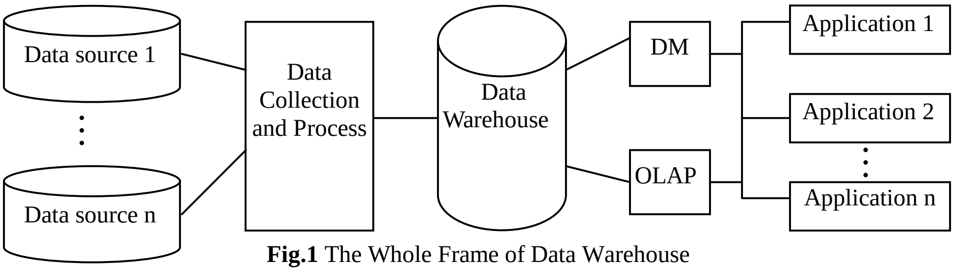


Fig.1 The Whole Frame of Data Warehouse

4.1 Processing complexity and the variety of the campus information data

The data of campus management system is complexity and variety. This system not only need manage students' score, prize, punishment and the information of graduated students and so on but also manage teaching, scientific researching and personal information. Moreover this system need give decision sustain function such as evaluation and analysis to teaching quality, scientific research ability and so on. So the traditional management can’t meet the requirement. The application of campus information management system based on data warehouse not only relieves the burden of manager, raises the efficient of educational administration, but also makes the management become normalization, systematization.

4.2 Providing the Decision Sustain for Analyzing The Diathesis of Students and Searching Information

To solve the difficulty of students’ synthetically characters difficult and the gauge uncertain, and to index, analysis, optimize the more satisfaction of decision support result, general database technique is not competent. Using the data mining technique to develop the campus information management system, support to make policy the system, by student's character of completely analyze, establish the system of target and the synthesize the valuation's mathematics model with student's character. Based on these, the associate policy software should be programmed, and provide the analysis relevant information of the student character, student employment information and adoption multi-dimensions data store of the data warehouse method, can contain the large quantity data about analyzing decision, and guarantee the data's consistency, integrality and safety, analysis and query high-efficiency and interactive GUI can high quantity, quickly satisfy user demand information.

4.3 Providing the Technique Support for Establishing Campus Information Management System

Now, the campus information is so much, processing them is difficult using traditional database technology. DW provides key technique for developing campus information management system. DW can store, manage and use data effectually. Data warehouse is a new technique brown up in recent years, but it has already been used in extensive realms such as telecommunication, finance and so on. At the same time, data warehouse technique also acquired development. Although DW is in the start stage on campus management, it has application potential in knowledge discovers, data mining, organizing and analyzing etc aspect. In the development of campus information management system, meta data defined and automatically extracted，a large mount of resource and distributed database quickly accessed ， the implementation of inter-operation of the distributed resource repository all need DW technology. Data warehouse is also necessary and key technology to implement the rapid crossing-database query of different level distributed resource repositories and rapid query engine based on parallel process.

4.4 The Model about the Campus Information Management System Based on Data Warehouse

We put together the warehouse, data mining, OLAP, model etc. the technique knot, designed a model of the campus information management, such as figure 2 shown.

Operational type data is the student’s basic data of lately semester in the university, once the data is overdue and then the data will be imported to the data warehouse. Large numbers of data conversion usually take place when operation type data is transferred to data warehouse. Large quantity of basic information of the students are put in the data warehouse, and by data mining from large quantity of random student data we can pick up useful information that is not known by people beforehand. By analyzing student education quantity and the feedback information of employer, university can establish perfect courses setting, improve teaching method and improve practice teaching. The graduated student's information put in the disk, magnetic tape...etc. is generally put into the data

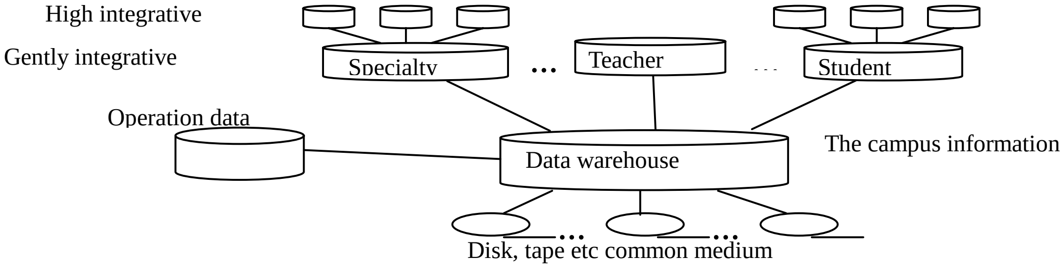


Fig.2 The Structure of the Data Warehouse

warehouse when it is needed. We can synthesize gently the information in data warehouse, for example, some professional student's score, some course score in different class, each course score about some students and grades etc. Based on the gently synthesize the user can synthesize highly, such as the pass rate of some courses of some professions, some score of some courses of some students in some classes etc. Data warehouse technique not only provided place to accept the large quantity information but also provided technique support for mining with deep layer and analysis for information resources timely. The application of data warehouse and mining technique can discover the real and worthy information from the large quantity the complicated data, such as the synthesizing character level of student and the rationality of course establishing etc.. We can predict that along with the development and maturing of data warehouse technique continuously and the application in the campus information management this system will make the query ability of the campus information management and the ability of decision support to get further improvement, and will also make a foundation to campus information management.

5. THE REALIZATION OF THE CAMPUS INFORMATION MANAGEMENT SYSTEM BASED ON DATA WAREHOUSE

The campus information management system based on data warehouse primarily includes gathering all kinds of the source data, storing and managing data, obtaining necessary information. The detail implement includes steps as following:

1) System analysis and programming
This step includes analyzing the requirement about the campus information management system based on the data warehouse, Establishing the target of developing this system, making the engineering plan, analyzing and establishing the technique environment, choosing the software and hardware resources of the data warehouse of campus information management system, which includes the developing platform, database management system (DBMS), network communication, terminal interview tool and establishing the service target etc..

2) System design
This step includes making sure the topic of data warehouse of campus information management system, setting up the data model, confirming the topic according to the policy and requirement of the campus information to choose the data source and design the logical construction according to the warehouse's data and organization, designing the database in the warehouse of campus information management system, developing the data's physics store construction in data warehouse according to the decision requirement and emphasis on a certain topic and developing data conversion programming. Data conversion occupies a large amount of workloads in data warehouse development. Data source cannot be directly added to the data warehouse. It must be converted before it is loaded to data warehouse. The ways of convert include extracting, deleting, connection identify, expanding, verifying, renewing, loading and so on. This step also includes managing meta data, defining meta data, in another word defining data and the relation of each component in the system. The meta data describes the data and environment about the data warehouse, which include the key word, attribute, data description, physics data construction, source data construction, mapping and convert the rule, synthesize algorithm, code, default, safety request, variety and data time limit etc. The meta data is usually divided into two types. One is management basic data, which describes source data and its contents, data warehouse’s topic, data convert and the each operation information. Another type is user basic data, which helps the user to search the information, comprehend result and understand data and organization in the warehouse. Finally we develop the tools of analyzing data about the student search manage and decision, and establish the decision sustain of the structure require, implement and use the tool about data analyze of the warehouse, which includes the optimize of the inquire tool, statistics analyze tool, C/ S tool, OLAP the tool and the data mine tool etc. The decision sustain is implemented by this analysis tools.

3) System test
Testing the backstage data warehouse to insure the maturity of the data warehouse; testing the application procedure, search tool, long range register procedure and analyze tool etc. to insure the haleness of the forestage’s program.

4) System maintenance
System maintenance is mainly managing the environment of data warehouses. Data warehouse must be managed as other systems, which includes the quantity examination, managing decision sustain tool and application programming and renewing data periodically to make the data warehouse run in normal.

6. CONCLUSIONS

This text brings up a way for developing campus information management system based on data warehouse. In normal data warehouse provides basic information source for new decision sustain. Higher decision support needs data mining (DM) in the data warehouse (DW). DW not only can provide place to accept the large quantity information, but also can mine out the real and worthy information from large quantity data. This system implements the analysis of the development trend of department and specialty, the evaluation of the teaching quality and the ability of scientific research, the requirement analysis to the resource of person with ability, and the analysis of school cost and economy benefit. Along with its continuous development and progress, this technique will advance the university’s competition and level greatly.

7. REFERENCES

[1] W.H.Inmon. Building the Data Warehouse [M] wang zhihai.translate. Peking: China Machine Press,2000.

[2] Jiawei Han. Micheline Kamber . Fan ming and Meng Xiaofeng translated. Peking: China Machine Press, 2000.

[3] Hu yan . According to the decision support of the data warehouse the tool's comparison research [J]. Computer application, 2000,20(6):20 22.

[4] Li Wei ,Li Wan zhou. According to the data warehouse technique of into the cancel and store the system's design with realize [J] Computer engineering and application. 2001(10):53~55.126

Wang jianxia (1970-), female, is a lecturer of Hebei University of Science and Technology. She graduated from Hebei University of science &Technology with specialty of Electrical Engineering in 1994, gained master degree from yanshan University of Technology in 2003. She has published three books, over 10 Journal papers. Her research interests are in Network and Database application technology.

Application of Distributed Database of Electric Power Management Information System

Wu Wei[1], Wu Jie[2], Hu Peng[3]
[1]Department of Computer Science, Wuhan Polytechnic University
Wuhan, China, 430023
Email: wu_72@163.com
[2]Yangtze River Scientific Research Institute
Wuhan, China, 430010
Email: wujie1965@sina.com
[3]School of Remote Sense Information Engineering, Wuhan University
Wuhan, China, 430072
Email: hupeng19764@sohu.com

ABSTRACT:

In this paper, we designed a Electrical Power Management Information System(EPMIS), which facilitates a integrated retrieval of data from several distributed database on the network. We propose the EPMIS structure which relies on a global replicated data and a local information database. And several key techniques in the EPMIS are illuminated.

Keywords: Distributed Database, Data Replication, Data Collision

1. DISTRIBUTED DATABASE TECHNOLOGY

Being a new branch originated from database technology in later 1970s, distributed database is a data aggregate logically as a whole but physically distributing saved at computer network nodes, and characterizes distributing property and logic harmony property as compared with traditional centralized database. As concerning distributing property, it means that the data aggregate is divided into stated data subset according to need, and saved separately at different places (nodes), instead of at some single computer. For logic harmony property, it means that the data subset saved at different nodes are limited each other by rigorous restriction rules, so the data subset also form an organic integrity logically.

Being in speedy growing, distributed database system has become an important field of information processing subject, and has several advantages as below:

(1) Can solve data conformity problem when the organization is dispersed. For example remote education system, its data centre and substations are located in various cities and area, in operation they need processing data severally and exchanging data each other, therefore distributed database system should be employed.
(2) When an organization expands through increasing new relative branches, distributed database system could be extended under some circumstances.
(3) Load equilibrium. Because of adoption of decomposed data, local applications reach their maximal level, corresponding to the mutual interference among processors drop to minimum degree.
(4) High reliability. Thus it can be seen that the distributing property of distributed database is different from that of centralized database, and its logic harmony property is out and away excellent than that of decentralized database. But centralized database is essentially the basis of distributed database, and computer network is the necessary environment of distributed database.

Having been employed in varied systems such as computer local area network, management information system, etc, distributed database has achieved remarkable economical benefit and society benefit. In this paper, the primary problems of application of distributed database on electric power management information system are discussed.

2. DATA DISTRIBUTION DESIGN OF ELECTRIC POWER MANAGEMENT INFORMATION SYSTEM

A certain power corporation is a provincial enterprise. The corporation governs 5 subcompanies which respectively dominate numerous power plants, transformer substations and electricity transmission lines. Management of the provincial corporation includes many aspects such as personnel administration, financial management, production management, building and structure safety management (asset management), etc. At the present time, the actuality of the computer-based management of the corporation is: each functional department among the corporation develops respective management system, for example finance management system for financial department, personnel information management system for personnel department and production management system for production department and so on. And, these management systems are mostly based on one-machine environment or department local area network environment, so resource sharing and message exchange cannot be achieved among these systems. With continual development of modern management technique, this kind of management mode results in overmuch resource engrossing, and cannot meet the requirements of open information management, sharing and mutual data access.

Based on the demand stated above, we have developed a kind of Electric Power Management Information System (EPMIS). Using this system, data processing and real-time data sharing can be achieved between the provincial corporation, the subcompanies and idiographic power plants. Due to the management universality of the corporation and the distributing property of its underlying units, multilevel distribution structure should be adopted for the system. Several key problems are described as below:

Determination of Data Distribution Plan

Alike centralized database, distributed database is also made up of two parts, viz. physical database and descriptive database. Physical database is the data aggregate for application, and descriptive database is about the definition of data structure and the distributed description of global data. The data in distributed database is divided into local data and global data. Local data is the data for local station application, and global data is for global application (every station can visit the data) in despite of being physically saved at various stations. Data distribution denotes that the data is divided into many logic segments according to need, and these data segments are distributed at every station according to stated rules, in stead of saving at single station. The rationality of data distribution is much influential to performance, reliability and efficiency of the system, so the establishment of data distribution plan is very important to the management information system.

1) Basic principles of data distribution

Generally, the data distribution problem is allocating a given group of data segments to network nodes according to manipulation (inquiry and updating) and there use frequency of the data segments, and getting a certain minimum expense function value (usually denoted by communication fee or communication time). If the logic data segments are distributed to some network nodes without redundancy, the data distribution problem would be very simple. But in the interest of improving the usability and reliability, reducing communication expense and increasing data access efficiency of the distributed database, it is necessary that multi copies of same data segment should be redundantly saved at multi network nodes, therewith making the data distribution problem more complicated. In distributed database designing, the principles of data distribution are generally that data should be at the node where it is used, and the global system performance checked by the load equity method should be optimal.

2) Data distribution plan of EPMIS

According to above data distribution principles and the management status of the corporation, for establishment of distributed database system of EPMIS, database management systems should be installed at the provincial corporation and every subcompany. And, these database management systems are not isolated but related. These database systems distributed physically all over the province are logically related, they all together compose an integrated distributed database system. Owing to hugeness of data of the corporation, the performance of database management system software and hardware should be excellent. Excepting being granted, a certain subcompany cannot freely access the data of other subcompanies. In view of cost and safety of the system, each subcompany should only save the data belonging to itself, and the provincial corporation should hold whole data. Because the data is saved at the site where it is frequently used, both system response time and expense of network communication are reduced. In distributed database, data copies would objectively increase reliability of data. In addition, owing to holding whole data, the provincial corporation is easy to carrying on statistical and analytic work facing to whole enterprise. In designing of the distributed database, the database of the subcompanies are set up as slave database (guest), and that of the provincial corporation are set up as master database (host). So the whole distributed database shows a kind of master-slave structure which is illustrated as below.

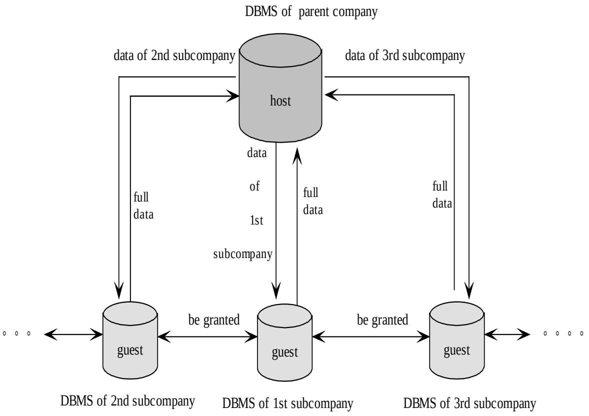


Fig 1. DBMS of Parent Company

3. SEVERAL KEY TECHNIQUES

3.1 Maintenance of Database Consistency

Once the master-slave database is built up, it is necessary to keep the consistency of every database. When you maintenance the database, you must consider the data quantity and the bearing capacity of network bandwidth. Because the distributed database we used is SQL Server, it is possible that we can use its new quick data refresh technique to update data timely between the master database and the slave database, so the consistency of the master-slave database is kept. Because each slave database is relatively independent and the master database accesses the data of the slave database only when the information of the congener structures are compared, so we divide the database tables needed duplicating into groups according to their correlation and possible data updating quantity. While quick refreshing, the transferred data quantity of every group should be farthest same. For the sake of saving time and increasing the efficiency of the network bandwidth, every group is refreshed at the same time.

3.2 Influence of Weak Consistency

Because quick data refreshing is time-lapse, the data consistency of the whole distributed database cannot be completely ensured between two refreshing. This kind of consistency is called weak consistency, and the result is that the database copies at different network nodes are possibly inconsistent sometimes, and the inquiry results are possibly untruthful. For this problem, you can start the database refreshing program to solve it.

3.3 Data Replication

Another important problem of data distribution and data processing is keeping data synchronization. For example, if a certain subcompany modified the data of its slave database, and the modified information didn't arrive at the master database in good time, another subcompany would get outmoded data differed from the true data at the former subcompany when it accessing the master database to obtain the modified data. This phenomenon did result from unsynchronized modification between the slave database of the former subcompany and the master database. The replication service offered by SQL Server is able to achieving data replication from database to database. Copying selected data to another network node, this kind of replication propagation is called snapshot refreshing which is divided into complete refreshing mode and quick refreshing mode. Here we choose quick refreshing mode. Its principle is that at the same time the record in database is modified, the quondam value and the updating value of the record are registered, then search the corresponding record in the replication database via the quondam value, and replace the quondam value with the updating value.

3.4 Influence of Lag of Processing Time

As a result of that the quick refreshing of database is time-lapse, the data and transactions committed to the corporation by subcompanies cannot be immediately processed, the returned decision-making data cannot come at once. So the lag of data transmission results the lag of transaction processing. Acceptable lag degree by users is important basis on setup of manipulation cycle of quick refreshing. For this system the requirement of lag of data processing time is not very high, so it is usually not necessary to consider the influence of the time lag. Only before decision-making, the database refreshing program should be started to confirm refreshing.

3.5 Data Collision

Owing to adopting time-lapse quick refreshing technique, there is a lack of concurrency control mechanism. Data collision is likely caused by appending, modification and deletion manipulation of the master-slave database. Data grouping method is the effective means which can solve the problem of data collision. In distributed database, data grouping has different grain size in terms of table and record.

1) Grouping in terms of table is entirely granting the modification authority of tables to a certain subcompany or the provincial corporation. For the tables that can be completely modified by one user, the data collision problem can be simply solved by grouping the tables needed duplicating. Regarding database tables, one modification member is designated for every table. The modification number may be the provincial corporation or a certain subcompany. Regarding the provincial corporation or the subcompanies, every modification member is assigned a modifiable table aggregate. As long as the intersection of the table aggregates is a null set, the data collision problem would be avoided.
2) Solving data collision problem by grouping in terms of record. For some database tables, it is impossible that one and only modification member is designated. Many users need to modify the records of the tables, so grouping in terms of table is unable to solve data collision problem. For these circumstances, grouping in terms of record should be employed. Grouping in terms of record means that the modification authority of a certain kind of records of a certain table is only granted to a certain modification member. In stead of all records of the table, a specified user can only modified a certain kind of records of the table.
3) Solving data deletion collision problem by using deletion mark. While one of two copies of one record being deleted and another being modified at the same time, data deletion collision would take place. This kind of data collision problem can be solved by using data deletion mark. Before being deleted from database, the records are firstly marked. The marked records would be really deleted from database after definite time. Because nobody would modify any marked records, deletion of the marked records wouldn't cause any data deletion collision problem.

4. EPILOGUE

In this paper, we have discussed several key techniques and implementation method of application of distributed database on EPMIS. Along with development of distributed database technology, many problems still need to solve for its application on various environment. We believe that development of distributed management information system based on distributed database would be much more stirring.

5. REFERENCES

[1] M.Tamer Özsu ,Patrick Valduriez,(2002),Principles of distributed database systems. Prentice-hall,Inc.
[2] Doreen L.Galli.(2003),distributed operating systems concepts & practice. Prentice-hall,Inc.
[3] Hector Garcia-Molina, Jeffrey D.Ullman, Jenninfer Widom. (2003), Database systems :the complete book. Pearson educaton ,Inc.

The Feature Parameter Extraction in Palm Shape Recognition System

Wang Jianxia, Qin Min, Zhou Wanzhen, Wang Jingyang, Zheng Guang
College of Information Science & Engineering
Hebei University of Science and Technology
Hebei Shijiazhuang 050054, China
Email: wang_jianxia@hebust.edu.cn **Tel**：0311-8613336

ABSTRACT

By feature parameter of palm shape to recognize palm is a common way. This paper introduces the recognition process of this way, brings forward the method of picking-up feature parameter of palm shape. Then eight feature parameters are picked up. They are the length of pinkie, the length of ring finger, the length of middle finger, the length of forefinger, the length of thumb, the width of ring finger, the width of middle finger, the width of palm. Experiment indicates using these eight feature parameter to recognize palm can reach high veracity and rapidity.

Keywords: Palm shape recognition, Preprocessing, Edge detecting, Feature extraction

1. INTRODUCE

Because biological characteristics have personal stability and individual otherness, presently, identification with biological characteristics is a hotspot in pattern recognition. Palm shape recognition uses palm shape as the recognition object of pattern recognition. This method extracts parameters from the outline of a palm, and compares these parameters with the data in the database to get the recognition result. Compared with other body recognition technology, palm shape recognition has its peculiar advantages such as low demand to equipment, small calculating quantity, quick speed and easy to get parameters.

2. THE PROCESS OF PALM SHAPE RECOGNITION

The system of palm shape recognition includes getting palm image, preprocessing images, extracting palm outline, exacting feature parameters and recognition. The process of this system is shown as Fig.1.

First we scan the palm to get a bitmap of this palm. Then computer preprocesses this bitmap, including image grayscale, binary, filtrate, image buildup and so on. By edge detection, edge thinning and outline tracing we can get perfect palm outline. Then palm feature parameters can be extracted. Finally we compare with data in database to recognize this palm.

3. THE METHOD OF EXTRACTING FEATURE PARAMETERS

Extracting feature parameters is the key in palm recognition. The process of extracting feature parameters includes image regularization, looking for extreme point, extracting feature parameters and calculating feature parameter.

3.1 Image Regularization

Though we can adjust angle of scanning and screening to diminish the gradient of palm, we can not assure that all palms are vertically placed. It increases difficulty to palm recognition. Because the lean will bring warp in pick-up feature parameter and make a mistake in palm recognition, it is essential to regularize palm image. Regularization such as circumrotate and horizontal remove and so on can deal with palm image of random placement to insure the middle finger upright. Thus extreme point (finger tip and finger root) can be at correct place. We can regularize palm image by circumrotating image. The circumrotation is relative to the center point of image. X-coordinate of center point is obtained by summing all X-coordinate of pixel point of palm edge, dividing the total of pixel. Y-coordinate of center point is obtained by summing all sums Y-coordinate of pixel point of palm edge, dividing the total of pixel. Palm image need circumrotation or not is based on the angle between middle finger and vertical direction, the experiment result indicates that image need circumrotate when the angle greater than 5 degree. The angle less than 5 degree will not influence extracting feature parameters. We will not rotate image when the angle is less than 5 degree. Because rotating image need large numbers of triangle operation, operation speed will become slower; the recognition time will become longer.

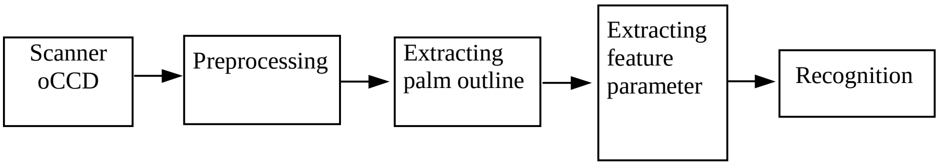


Fig.1 the System Chart of Palm Shape Recognition

Fig. 2 shows the image before regularization and the image after regularization.

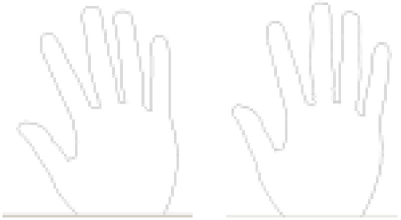


Fig. 2 (a) the image before regularization
(b) The image after regularization

3.2 Looking for Extreme Point

After obtaining palm outline, we need find every extreme point in palm outline in order to pick-up feature parameter. We use the comparison method in program that this point is maximal point if y value of one point more than y value of its neighboring point. Contrarily it is minimal point. Fig. 3 shows the feature point extracted in palm outline by above method.

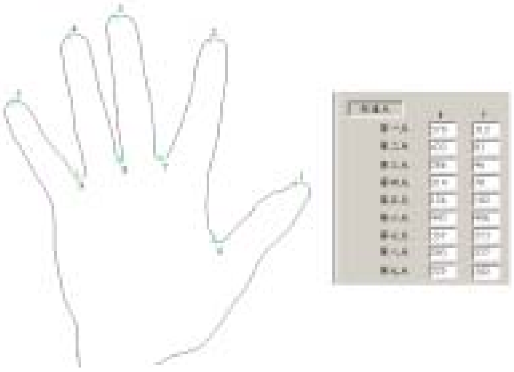


Fig.3 The Character Point of the Palm Shape

3.3 Feature Parameter Extraction

Digital processing of palm image is key in palm shape recognition system, processing well or not directly influence the error of feature parameter extraction, thereby it determines exactness ratio of palm shape recognition. We use scanner as input equipment in experiment, distinguish ratio in scanning is 84dip. After observing and analyzing large number of experimental results, we determine palm shape parameters as fig.4 shown. In fig.4, parameters L1, L2, L3, L4, L5, L6 and L7 are defined as the distance between each finger tip and each finger root one side or both sides. Parameters W1, W2, W3 and W4 are defined as the middle positive section projection width between finger tip of pinkie, middle finger, ring finger, forefinger and their finger root. Parameters H1 and H2 are defined as the distance between finger tip of ring finger and forefinger and the midpoint of their two sides finger root. Parameter W5 is defined as H2 extend toward palm center to 1.3 times of H2, at point of 1.3 times of H2 we draw a line vertical with H2, this line cut across the palm edge. Parameter W5 is defined as the distance between two points of intersection. In same condition, we choose parameters of ten natural patulous palms as palm target stylebook, we select a palm from these target stylebook random, individual stylebook is obtained by measuring this palm time after time in same condition.

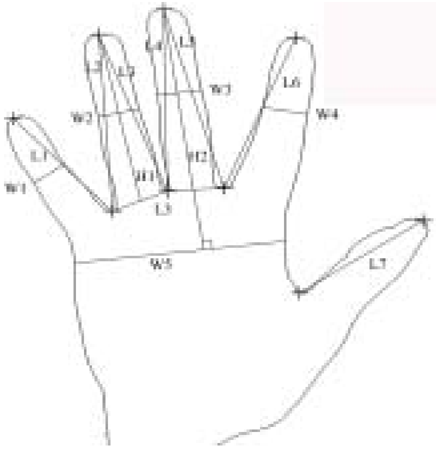


Fig.4 The Chart of Extracting Feature Parameter

3.4 Feature Parameter Calculation

After looking for nine extreme points in palm outline, we can calculate eight feature parameters. Feature parameters in palm shape are confirmed finally, they are L1（the length of pinkie）,H1（the length of ring finger）,H2（the length of middle finger）,L6（the length of forefinger）,L7（the length of thumb）,W2（the width of ring finger）,W3（the width of middle finger）,W5（the width of palm）.we redefine these feature parameters according to picking-up order in program and obtain eigenvector of palm shape as X={x0，x1，x2，x3，x4，x5，x6，x7}.They are the length of middle finger（x0）,the length of forefinger（x1）,the length of ring finger（x2）,the length of thumb（x3）,the length of pinkie（x4）, the width of palm（x5）、the width of middle finger（x6）、the width of ring finger（x7）。Fig. 5 shows eight feature parameters. x_0 is the distance between the third point and the tenth point, the tenth point is the midpoint between the seventh point and the eighth point; x_1 is the distance between the second point and the seventh point; x_2 is the distance between the fourth point and the eleventh point, the eleventh point is the midpoint between the eighth point and the ninth point ; x_3 is the distance between the first point and the sixth point; x_4 is the distance between the fifth point and the ninth point; x_5 is the distance between the eighteenth point and the nineteenth point; x_6 is the distance between the fourteenth point and the fifteenth point; x_7 is the distance between the sixteenth point and the seventeenth point. The coordinates of nine extreme points are stored in array P, eight feature parameters x_0~x_7 are stored in array X.

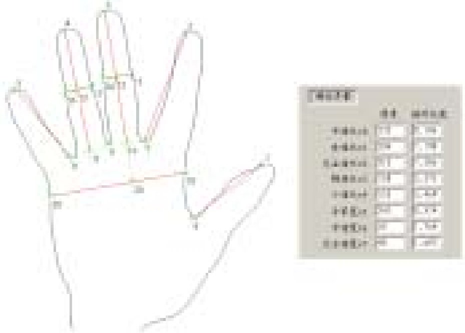


Fig.5 The Chart of the Characteristic Parameter

4. EXPERIMENT AND CONCLUSION

After eight feature parameters are confirmed, we can look upon every palm as a pattern. We should make large numbers of experiments to confirm feature parameter of each palm because error is appeared randomly. We extract randomly six images of the same palm as data source of this palm, and calculate these eight feature parameters of every palm image. Then we average every feature parameter of six images of the palm to get eight feature parameters of this palm. Underside is six experiment results of two palms, shown as table1 and table2.

From above two tables, take example for the length of middle finger, the average x_0=8.31 in Table1, the average x_0=8. 87 in Table2, we can see that the palm shape has personal stability and individual otherness. The experiment indicates feature parameters picked up using this method is close to the result practice measured; recognition ratio by this method is up to 95% upwards.

Table1. The testing result of the first group

Parameter Name	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
The first breadth image	8.32	7.23	7.68	5.81	5.69	8.47	1.69	1.69
The second breadth image	8.32	7.35	7.68	6.14	5.75	8.56	1.69	1.66
The third breadth image	8.26	7.32	7.68	5.84	5.78	8.59	1.72	1.63
The forth breadth image	8.32	7.38	7.68	5.54	5.78	8.62	1.69	1.63
The fifth breadth image	8.35	7.23	7.74	5.78	5.75	8.50	1.72	1.66
The sixth breadth image	8.26	7.38	7.65	5.63	5.75	8.62	1.69	1.66
Average	8.31	7.32	7.68	5.63	5.75	8.55	1.70	1.65

Table 2 The testing result of the third group

Parameter Name	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
The first breadth image	8.80	7.83	8.31	5.72	6.14	9.68	1.85	1.63
The second breadth image	8.98	7.80	8.26	5.60	6.02	9.41	1.82	1.66
The third breadth image	8.89	7.83	8.32	5.96	6.20	9.47	1.79	1.60
The forth breadth image	8.89	7.68	8.35	5.63	6.14	9.41	1.79	1.60
The fifth breadth image	8.89	7.68	8.35	5.84	6.14	9.56	1.82	1.63
The sixth breadth image	8.77	7.86	8.32	5.57	6.14	9.29	1.79	1.63
Average	8.87	7.78	8.32	5.72	6.13	9.47	1.81	1.63

5. REFERENCES

[1] Li chunming,Zhuang qingde,Zhang ying,Wu ruihong. The Producing and Digitalization Technology of Palm Image Files for Identity Recognition. Electronic Measurement & Instruments.2001.15: 410-413

[2] Tao Huang. Structural Characterization of Grain Pattern Diversity in Parametric Space. Journal of Materials Science.1999,34(18):4551-4561.

[3] Thomas H. Kolbe, Lutz Plümer, Armin B. Cremers. Identifying Buildings in Aerial Images Using Constraint Relaxation and Variable Elimination. IEEE Intelligent System &Their Applications.2000,15:33-39.

Wang Jianxia (1970-), female, is a lecturer of Hebei University of Science and Technology. She graduated from Hebei University of science &Technology with specialty of Electrical Engineering in 1994, gained master degree from Yanshan University of Technology in 2003. She has published three books, over 10 Journal papers. Her research interests are in Network and Database application technology.

Commercial Bank Credit Risk Real-Time Value-at-Risk Computing System

Rong Lan[1, 2] **Shouqi Zheng**[2]
[1]School of Economic and Finance, Xi'an Jiaotong University
[2]School of Electric & Information, Xi'an Jiaotong University
Xi'an, Shanxi, China
Email: wu_lan@mial.xjtu.edu.cn **Tel.:** 029-85267141

ABSTRACT

The main business in bank is risk and credit risk is the principal challenge for managers, especially in our country's national commercial banks. VaR is a risk assessing methodology and CreditMetrics is a framework to realize credit risk VaR calculation. They are all the advanced and mature techniques in modern financial risk management. As VaR calculations are highly computational intensive, there is a increasing demand for better computing resource. Distributed computing is one way to realize effective financial computing. After analyzing the principle of VaR and CreditMetrics, combing our national commercial bank hierarchical organization, we design and realize a credit risk real-time VaR computing system and make simulation computation. The main aim of this project is make a usability research, including both the methodology and our financial management system.

Keywords: Risk, Credit Risk, Value-at-Risk (VaR), CreditMetrics, Distributed Computing.

1. INTRODUCTION

The main business of bank and insurance companies is risk. Here we defined risk as the degree of uncertainty of future net return. Based on the source of the underlying uncertainty, one common classification of risks are market risk, credit risk, liquidity risk, and operational risk. So, the evaluation of individual and integrated risk is essential for both bank and insurance companies.

"Transparency is the key to effect management" [1]. To promote greater transparency of risk, we need a framework to quantifying risk in company asset.

During the 1990's there has been established a measure for risk in finance theory as well as in practice, the Value-at-Risk, VaR. It was mainly popularized by J. P. Morgan's RiskMetrics, a database supplying the essential statistical data to calculate the VaR of derivatives. Currently, VaR is being embraced by corporate risk managers as an important tool in overall risk management.

Globally, financial institution are taking on an increasing amount of credit risk, credit risk has become perhaps the key risk management challenge. As credit exposures have multiplied, the need for more sophisticated risk management techniques has increased. Now, risk managers are seeking to quantify and integrate the overall credit risk assessment within a VaR statement to realize "all the answer is in numbers."[6]

In our country, the financial risk caused by the inability of counterpart to meet its obligations is very serious in all financial institutions, especially in national commercial banks. But the credit risk management technique is very backward, and" there is even no credit risk quantifying methodology is being used."[7] This situation is not fit in with our rapidly developing economic environment. To study advanced and mature credit risk management technique and set up applying test environment, is very important for us to summarize and build our-self's credit risk management system.

After analysis J. P. Morgan credit risk VaR framework—CreditMetrics, combining our country's commercial bank hierarchical organization, we use Java technique to build a loan credit risk VaR distributed computing and reporting system. We also have using J. P. Morgan free date set to make simulating calculation. All our aim is to make a usability research, including both the methodology and also bank management system.

2. THE CONCEPT OF VaR

2.1 The Concept of VaR

VaR is a measure of the maximum potential change in value of portfolio of financial instruments with a given probability over a pre-set horizon. VaR answers the question: how much can I lose with ***x%*** probability over a given horizon.

2.2 Parameters in VaR Calculation [3]

The sample space $\Omega$of the expected rates of return r on the portfolio W in some arbitrary assets is mathematically represented by the set *R*. We assume that the expected rates of return $r(t)$ with respect to the time horizon t of the investment is a random variable determined by the distribution function $\boldsymbol{F}$ **:** *Ω→[0,1]*,

$$F(x) = \int_{-\infty}^{x} p(r)dr \tag{1}$$

where p is the corresponding probability density. This means in particular that the expected rate of return $r(t)$ will achieve a value less than *x% (x∈R)* after time t with probability $P(r(t) \leq x) = F(x)$. Let $\tilde{\Omega} = R$ be the sample space of currency-valued returns $R = r(t)W$. The expected *loss* $L(t)$ with respect to the time horizon t of the portfolio W then is given as the negative difference between the return and the mean value, $L(t) = \mu W - R(t) = (\mu - r(t))W$. It is a quantity in currency units (cu). Note that any return less than the expected one means an effective loss, even if it is positive. Positive values of $L(t)$ mean a loss after time t, negative ones a gain. Its distribution function F *:Ω→[0,1]* is simply given by $\tilde{F}(L) = 1 - F(L/W - \mu)$, or

$$\tilde{F}(L) = 1 - \int_{-\infty}^{L} p(L'/W - \mu)dL' \tag{2}$$

with the probability density $-P(L^{'}/W-\mu)$. The *Value-at-Risk* with respect to the time horizon *t* of the portfolio then is defined as the maximal expected loss $L(t)$ not exceeded with probability *(1-α)*:

$$P(L(t)\le VaR)=1-\alpha, \quad 0\le\alpha\le1 \qquad (3)$$

α is the default or downfall probability of the Value at Risk, the common number is *1%* or *5%* (mainly depending on the time horizon). Here the portfolio can be a single or a multiple of different instruments. If there are more than one instrument, the sample space are the all scenario of different combinations and the correlation among them must be analysis.

Typically the data required for the calculations of VaR are statistical parameters for the 'underlying' and measures of a portfolio's current exposure to these underlying. So, to complete VaR calculation, we need the base data set includes:

1) The time horizon;
2) The portfolio;
3) The sample space *Ω* of the expected rates of return on the portfolio *W* and its distribution function or probability density;
4) The loss space and its distribution;
5) Confidence degree *α(1%,5%);*
6) Correlation.

When we use VaR methodology to quantity different kinds of risk, the way to obtain these dates is different. RiskMertic and CreditMetrics are frameworks to calculate market risk and credit risk, respectively. They all proposed by J. P. Morgan.[1, 2]

3. USING CretidMetrics TO REALIZE CREDIT RISK VaR CALCULATION

3.1 CreditMetrics

CretidMetrics is a benchmark for understanding credit risk. By assessing portfolio risk due to changes in debt value caused by change in obligor credit quality to computing finance risk. It was proposed by J. P. Morgan in 1997, co-sponsors by five leading banks (Bank of America, BZW, Deutsche Morgan Grenfell, Swiss Bank Corporation, and Union Bank of Switzerland) and a leading credit risk analytics firm, KMV Corporation. All these firms have spent a significant amount of time working on their own credit risk management issues, and provide related data set to support the development of CretidMetrics.

Now, they update data set periodically and provide to market for free. So, CretidMetrics has become one mature model to analyze credit risk. CretidMetrics framework can be described in the Fig. 1

To complete each part of calculation, we need different data to support. The main dataset are four pars: transition matrix, yield curves, spread, correlation.

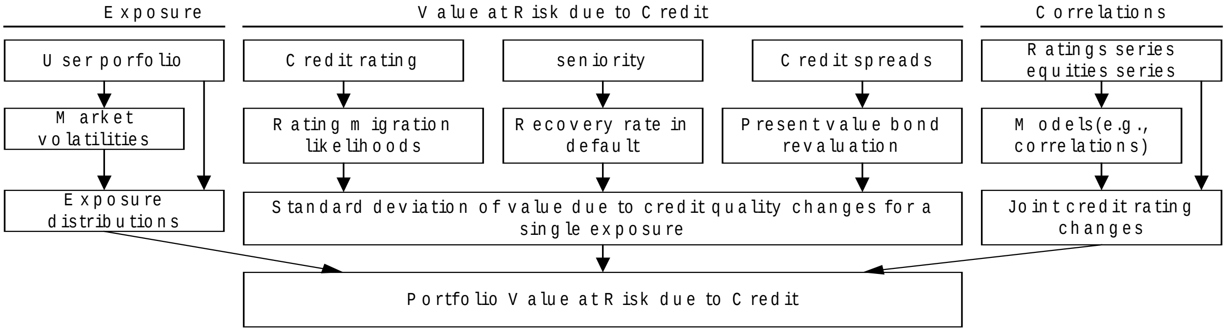

Exposure
Value at Risk due to Credit
Correlations
User porfolio
Credit rating
seniority
Credit spreads
Ratings series equities series
Market volatilities
Rating migration likelihoods
Recovery rate in default
Present value bond revaluation
Models(e.g., correlations)
Exposure distributions
Standard deviation of value due to credit quality changes for a single exposure
Joint credit rating changes
Portfolio Value at Risk due to Credit

Fig.1 CreditMetrics framework

3.2 Using CretidMetrics Framework to Calculate Credit Risk VaR

To realize credit risk VaR calculation, we need the sample space for the return and loss of a portfolio and its likelihood. But these data can't be obtained by directly observe, CretidMetrics provide a methodology to construct them. It includes three key steps.

Step1: Using a credit rating or grade system to obtain portfolio credit quality migration likelihood matrix;
Step2: Using forward zero coupon yield plus spread yield curve (matrix) to revalue portfolio as credit quality change;
Step3: Combining likelihood from step1 and the value from step2 to calculate volatility of value due to credit quality changes, and the expected value(mean), standard deviation.

3.2.1 Credit Rating

A credit rating or grade is assigned to firm as an estimate of their creditworthiness. It is usually is done by rating agencies. The famous of them are Standard & Poor's and Moodey's. Standard & Poor's rate businesses are as one AAA, AA A, BBB, BB, B, CCC, or default. Moodey's use Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C. The credit agencies continually gather data on individual firm and will, depending on the information, grade or re-grade a company according to well-specified criteria.

3.2.2 Transition Matrices

The likelihood that a given company migrates to another rating in a given forward time horizon is given by a transition matrix. In CretidMetrics framework, the time horizon is one year, and the likelihood of any credit rating migration in the coming period is conditioned on the senior unsecured credit rating of obligor. So, the matrix is time-dependent, forming a Markov process as time passes. One Standard & Poor's credit transition matrix is as Table 1.

Table 1 one-year transition matrix *(%)*

Initial Rating	Rating at year-end *(%)*							
	AAA	AA	A	BBB	BB	B	CCC	default
AAA	90.81	8.33	0.68	0.06	0.12	0	0	0
AA	0.70	90.65	7.79	0.64	0.06	0.14	0.02	0
A	0.09	2.27	91.05	5.52	0.74	0.26	0.01	0.06
BBB	0.02	0.33	5.95	86.93	5.30	1.17	0.12	0.18
BB	0.03	0.14	0.67	7.73	80.53	8.84	1.00	1.06
B	0	0.11	0.24	0.43	6.48	83.46	4.07	5.20
CCC	0.22	0	0.22	1.30	2.38	11.24	64.86	19.79

3.2.3 Forward zero coupon yield plus spread yield curve

The CreditMetrics dataset consists of the risk-free yield to maturity for several currencies. It contains yields for maturity of 1, 2, 3, 5, 7, 10, and 30 years. Additionally, for each credit rating the dataset gives the spread above the riskless yield for each maturity. Thus the spreads denote the differences between rickless bonds and a rated credit. Evidently, the riskier the credit the higher the yield: Higher yield for risky credits is compensation for the possibility of not receiving future coupons or the principal. J. P. Morgan gives one- year forward yield curve is in Table 2.

Table 2 one-year forward zero curve
by credit rating category *(%)*

Category	Year1	year2	year3	year4
AAA	3.60	4.17	4.73	5.12
AA	3.65	4.22	4.78	5.17
A	3.72	4.32	4.93	5.32
BBB	4.10	4.67	5.25	5.63
BB	5.55	6.02	6.78	7.27
B	6.05	7.02	8.03	8.52
CCC	15.05	15.02	14.03	13.52

For example, if there is a bond with the following attributes:

Seniority	Maturity	Coupon *(%)*	Face value	Grade
Senior unsecured	5	6	100	BBB

We can revalue the bond at one- year end with grade 'A' by:

$$\begin{aligned} v &= 6+\frac{6}{(1+r_1+s_1)}+\frac{6}{(1+r_2+s_2)^2}+\frac{6}{(1+r_3+s_3)^3}+\frac{106}{(1+r_4+s_4)^4} \\ &= 6+\frac{6}{(1+3.72\%_1)^1}+\frac{6}{(1+4.32\%)^2}+\frac{6}{(1+4.49\%)^3}+\frac{106}{(1+5.32\%)^4} \\ &= 108.66 \end{aligned}$$

Here, r_i denotes the risk-free rate expected after year *i*, and s_i is the corresponding annual credit spread. If the loan is downgraded, the spread s_i will rise, and thus the present value will fall.

As there are *8* possible states at one- year end, we need repeating the calculation *8* times for all likelihood scenarios. Here, if the credit quality migration is into default, the likely residual value net of recoveries will depending on the seniority class of the debt. One recovery rate in the sate of default provided by J. P. Morgan is given in Table 3.

Now we can form the possible one-year forward value for a BBB bond plus coupon and likelihoods.

To calculate percentile level credit risk (VaR), we start from the state of default, and move upwards towards the AAA rating state. We keep a running total of the likelihood as we move up until the percentile lever equal to confidence degree α, then the VaR is equal to mean subtract the grade value.

Table 3 Recover rates by seniority class (*%* of face value)

Seniority	Mean(%)	St. deviation(%)
Senior secured	53.80	26.86
Senior unsecured	51.13	25.45
Senior subordinated	38.52	23.81
Subordinated	32.74	20.18
Junior subordinated	17.09	10.90

3.2.4 Correlation

In this example, the portfolio has only a single bond, if there are more than one exposure, we need to elaborate the joint likelihoods in credit quality co-movements. The simplest way of obtaining the joint likelihoods is to just assume that these are the product of the individual likelihoods. So, if there are *n* exposures in a portfolio, then we can make 8^n different scenario, and each scenario has a joint likelihood *p*, and

$$p = \prod_{i=1}^{n} p_i \cdot$$

4. DISTRIBUTED COMMERCIAL BANK CREDIT RISK VaR COMPUTING SYSTEM

Using Java technique, based on CreditMtrics framework, we have design and realized a commercial bank loan credit risk VaR computing system.

4.1 The Reasons to Use Distributed Computing Technology

1) The organization of commercial bank is a hierarchical construction. There are central bank and branch bank. Central bank is the owner. He has many managers to be in charge of each branch. Branches have specific business. The whole asset and risk are distributed in each branch. To realize effective supervision, they must start from bottom.

2) Simplifying computing complexity and improving computing efficiency. Each branch is an independent economic entity to developing their business. So, they have many different credit asset. Meaning while, each exposure has different credit grade. If all branches send them portfolio to central bank to complete VaR calculation, the computing complexity will increase index, and the demand for computing resource will increase greatly.

4.2 The Benefits of Credit Risk Distributed Computing

1) Turn from loan amount distribute to risk limitation management.

Now, central bank assign annual loan amount to each branch, but no limitation on risk-based. This management model can easily make huge potential risk and also have made an incentive problem. Using credit risk VaR real-time computing system, the owner can combine loan amount and VaR limitation to distribute each branch (next year loan) task. By VaR limitation, the owner can realize effective management. Meaning while, we can use the rate of risk/ asset to evaluate manager's achievement. As Fig. 2

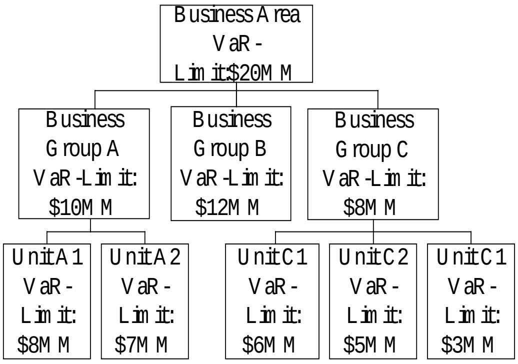

Business Area VaR-Limit$20MM
Business Group A VaR-Limit: $10MM
Business Group B VaR-Limit: $12MM
Business Group C VaR-Limit: $8MM
UnitA1 VaR-Limit: $8MM
UnitA2 VaR-Limit: $7MM
UnitC1 VaR-Limit: $6MM
UnitC2 VaR-Limit: $5MM
UnitC1 VaR-Limit: $3MM

Fig.2 Hierarchical risk limitation

2) To promote greater management transparence.

Based on uniform standard and algorithm, using local credit instruments database to complete different branch's VaR calculation. It is a benchmark to compare different branches potential risk and manager's achievement.

4.3 The System Architecture and Computing Process

Using C/S model, we set up the distributed computing system (Fig. 3).

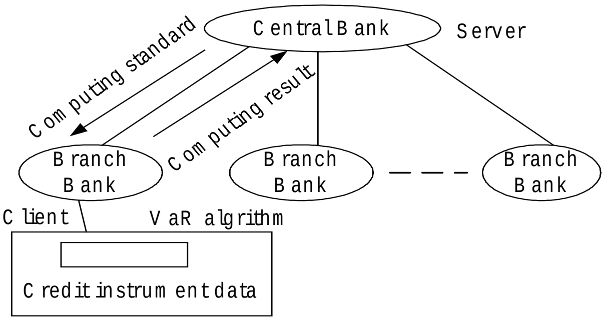


Fig.3 Distributed credit risk VaR computing sysytem

Central bank server makes financial market data statistic to form calculating standard (including transition metrics, forward yield curve, and default recover rate) and send them including JDBC, object serialization, socket, threads and synchronization technologies.to each branch bank client to complete VaR calculation. Many of the advanced Java features have been used,

Now, we have set up a simulating system including one central bank server and four branch clients. Using J. P. Morgan free dataset (Table 1, 2, 3), we complete four credit products real-time VaR computing. Here, each branch has one product, see Table 4. Though it is only a simulation computing, but we also can get a lot of information. For example, the reason why a loan grade can't below BBB.

Table 4 Credit asset and risk analysis table ($: million)

Branch	Grade	Seniority	Maturity	Coupon(%)	Amount($)	VaR($)	Risk/asset (%)
Branch1	A	Senior unsecured	4	4	10	0.2149	2.15
Branch2	BBB	Senior subordinated	5	5	15	0.1785	1.19
Branch3	AA	Senior secured	3	2.5	8	0.0139	0.174
Branch4	BB	Senior unsecured	4	6.5	20	5.2	26
Asset=53, VaR=5.6073, Risk/asset=10.58%							

5. CONCLUSIONS

1) We need these financial risk quantifying methodologies. Through in our country, a mature financial system hasn't set up, many parameters needed in model computing can not be obtained directly from financial market, there is no precision to say, but it is not effect these methodologies using. By these quantifying techniques, we can intensify financial organization internal management, and improving managing transparence. This is very important in speed up the process of financial reform.

2) Intensive financial computation is a way to increase competitive advantage.

3) There is a large space for all kinds of advanced computing techniques to be used. Later 1970's, our country had begun the project of financial electricalization. Thirty years has passed. Now, we have sep up an advanced financial information infrastructure, it is time for (all) computer scientists and financial specialists to combine to resolve financial problems. Web computing, distributed parallel computing, grid computing is now being used in financial analysis. It is a challenge and a chance for all of us.

6. REFERENCES

[1] Gupton, et. al. *CreditMetrics—Technical Document*. J. P. Mongan, April 2, 1997

[2] J. P. Morgan / Reuters. *RiskMetrics^{TM}—Technical Document.* Fourth Edition (December 1996)

[3] Andreas de Vries. *The Value at Risk*. International Journal of Theoretical and Applied Finance **4**(3), June 2001, 535-543

[4] Zvi Bodie, Robert C. Merton. *Finance*. BeiJing: Person Education, Inc. 2002

[5] Bruce Eckel. *Thinking in Java*, (Second Edition) BeiJing: China Machine Press, 2002

[6] Jay Wheeler. *The answer is in the numbers.* Risk.April. 1999

[7] Zong Liang, Jiang Hua. *International bank risk management developing and main new methodologies.* Financial Paper Online, see: http://efinance.yeah.net

Lan Rong is a associate professor in School of Economic and Finance, Xi'an Jiaotong University. She graduated from Northwest University in 1988 and obtained computer science master degree. Her main researching areas include distributed parallel computing, Finance Engineering. Now, she has published over 20 Journal papers.

Contract-based Interlayer : a Two-way Approach to Integrate Call Center with J2EE Framework

Wu cen, Lin zuoquan, Zhao xinyu, Zhao chen
School of Mathematical Sciences, Peking University
Beijing, 100871, China
Email: {wucen, lz}@is.pku.edu.cn

ABSTRACT

The traditional software architecture of call center is based on C/S framework for the efficiency of communication among various hardware equipments. Thus, it is difficult to integrate call center with other J2EE applications, which are otherwise based on B/S framework. In this paper, we propose the contract-based interlayer, which we introduce into the call center software architecture. This new approach partitions the call center into a front-end and a back-end, and then achieves integration of them by building an infrastructure, which enables them to plug into a common protocol and thus to cooperate with each other. As a result, the front-end related to the hardware can be based on the C/S framework while the back-end related to the business can be based on the B/S framework and therefore be integrated with the company's other systems based on the J2EE framework, for example, the customer relationship management (CRM) system, so as to promote the values of the company's call center.

Keywords: Contract-based Interlayer, Call Center, J2EE, CRM.

1. INTRODUCTION

The importance of high-quality services can never be exaggerated as a component of a company's competitive advantage in today's market. Furthermore, since 70% of all the interactions between customers and companies take place at the call centers [1], to better the functions of call centers is crucial for a company's to survive and thrive. Attentions should be paid to strengthen functions of call centers, especially in following aspects:

1) Call centers should reconstruct the part of business-related functions to compose a back-end system within a B/S-based framework or, to be more accurately, a J2EE framework, so as to share pivotal commercial data with other J2EE-based enterprise applications, or even be able to integrate with them;
2) Call centers should enable the agents to accomplish their routines through browsers, such as IE, instead of using typical custom-built client applications. Therefore, the company can add not only new local agents but also remote IP agents easily as the agent workstation doesn't need additional installation of client application.
3) Call centers should provide a convergence of multiple forms of customer communication, such as phone, fax, e-mail, and the web. For example, while accessing the website of a company, customers can select the type of call (Internet voice call, text chat, or a call back) that he or she wants to initiate to contact with call center's agents.

As part of the call center's functions related with hardware (the telephone network) is in a lower level and demands efficiency, call centers should retain the C/S-based framework of the hardware-related part. Since B/S-based framework and C/S-based framework, which apply totally different techniques, should be compatible with each other within a same infrastructure, call center's old software architecture cannot survive such new conditions. Many efforts have been made by IT industry in order to share data among different applications, but these methods cannot succeed in an environment where real-time collaboration is the hinge. At the same time, some open standards which are developed in order to provide live deployments of integrated applications using service-oriented architectures, such as CORBA and DCOM, provide complex and rigid interfaces, depend on sophisticated wire protocols, and ties too closely those cooperating parts. As a result, all the mentioned approaches, which had tried to integrate the two different frameworks, were inevitably led to increased complexities of software architectures and difficulties of project development.

In this paper, we advance a contract-based interlayer, which can partition the hardware-related part and the business-related part of the call center's software system clearly. What is more, it can enable two different frameworks, the C/S and the B/S framework, to communicate with each other through a common protocol. Therefore, call center's agents can finish all the tasks with a browser while at the meantime customers can get the help of agents also with a browser when they visit the companies' websites. Besides, this approach minimizes the relevance between the front-end part and the back-end part applications, which enables the company to be less dependent on an exclusive call center solution vendor and the call center project to be accomplished with a higher quality.

The rest of the paper is organized as follows. In Section 2, we introduce the traditional software architecture of call center and discuss its limitations. In Section 3, we present possible approaches for integrating the two different frameworks and discuss their limitations. In section 4, we propose the novel contract-based interlayer and describe this new software architecture. In section 5, we present several advantages of this new architecture. We finally draw a conclusion in section 6.

2. LIMITATIONS OF TYPICAL CALL CENTER SOFTWARE ARCHITECTURE

Typical Software Architecture of Call Centers

In order to combine the telephone network and computer network, a typical call center is based on a two-tier architecture [Figure 1]:

1) Hardware-integration Tier;
2) Service-aggregation Tier

The hardware-integration tier consists of the CTI (computer telephone integration) server, IVR (interactive voice response) server and multimedia (such as SMS and email) gateways.

The CTI server provides the "glue" that links the call center’s information system with the Private Branch Exchange/ Automatic Call Distributor (PBX/ACD). The service-aggregation tier consists of a series of applications running on a stand-alone server. They communicate with hardware-integration tier and call the APIs, such as Telephony Services Applications Programming Interface (TSAPI), Telephony Applications Programming Interface (TAPI), and Java Telephony API (JTAPI) [2, 3, 4], which are provided by the hardware. As soon as the service-aggregation tier receives the events thrown out by the hardware-integration tier, it sends the call to one agent based on call routing logic. The customer's data sorted by the incoming call number or email address will be sought out and sent to the agent's desktop on the agent answering the call, called the "screen pop" procedure [5]). The service-aggregation tier has its own database to preserve the customers' information and the past operation data. The call center agents rely on the desktop client application to communicate with service-aggregation tier’s server and then to complete routine operations. In a word, this software architecture is completely based on the C/S model.

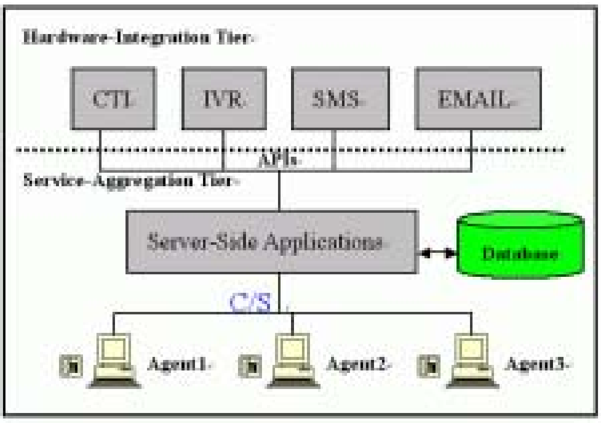


Figure1 call center's 2-tier architecture

Limitations of the 2-tier Architecture

The emergence and later prevalence of J2EE, the standard platform for enterprise solutions, makes the majority of corporations upgrade their enterprise applications based on J2EE infrastructure. Corporations hope call centers can share pivotal commercial data (such as customer information) with other J2EE-based systems, or can even be integrated with the company’s CRM system. Thus call centers must reconstruct the part of business-related functions to compose a back-end system within J2EE framework. And the agents’ workstations would not install client software---the agents work with JSP pages on browser to communicate with servlets on the back-end, then call the Enterprise Java Beans (EJBs), which encapsulate the business logics of call center application such as carrying out operations to insert pieces of records of electrical appliance maintenance into database. At the same time, the steep growth of e-commerce has made companies upgrade corporate Web sites from electronic versions of glossy brochures to full service storefronts. Consequently call centers have to be integrated with Web sites so the customer can contact an agent once he has a problem when visiting the corporation’s web pages [6].

While the part related with hardware (which is called front-end system corresponding to the back-end part) still has to retain the C/S framework for these applications are in a lower level and have to react quickly for real-time communication, in another word, B/S mode architecture is not fit for them. As a consequence, the old 2-tier architecture cannot fulfill this task for that these two different frameworks have to be separated from each other. Hence, the agents have to communicate with the service-aggregation tier within the C/S framework in order to cooperate with the front-end hardware to finish tasks such as replying the customer's incoming calls; meanwhile the agents communicate with the back-end sever within the B/S framework to finish tasks, for example, restoring a piece of maintenance record. Then these two different techniques have to be integrated at the agent's workstation while the browser, as a thin-client, cannot work as an integration application for it cannot interact with the front-end server within the C/S framework.

3. LIMITATIONS OF CURRENT SOLUTIONS

It is necessary for developers to integrate B/S-based framework and C/S-based framework, which means to realize the interactions between two different frameworks using different techniques. This type of business integration (that is, the task of tying many information systems together to support a business process) has historically been an extremely complex, costly, and time-consuming task and researchers who are interested in the study of the legacy system have proposed several solutions: screen scraping, database gateway, XML integration, database replication, functional (Logic) modernization like CGI integration, object-oriented wrapping, and componentization of legacy systems [7,8,9].

Among these many solutions, the screen scraping method cannot fulfill the needs of the call center because of its simple functions, and the data-based approaches (such as database gateway and the database replication) or XML integration lack the ideal efficiency for the swift reaction of call centers. It seems that functional modernization approaches, such as the CGI integration, object-oriented wrapping and the component wrapping, are appropriate solutions for that they can bring us adequate efficiency. But what follows next is that this approach makes too close a tie between the front-end and back-end, which made the development of the call center projects even more difficult.

A widely accepted eclectic solution using the component wrapping approach, which is used in CORBA and DCOM, is to rebuild the old desktop client application which retains its function such as answering the incoming call or initiating a three-part meeting and also runs a third-party component inside the application acting as a browser to communicate with the EJB tier. However, this solution has a vital weakness: due to the restriction of the limited functions of the third-party component, it cannot transfer smoothly the data necessary to the back-end such as incoming call's number. Moreover, when addressing some complicated functions such as three-part meeting (which means two agent working together for the same customer), this clumsy solution has to finish some complicated interactions of web pages that the back-end system have to expose the corresponding webpage logics even the corresponding EJB interfaces to the front-end. In all, too many interactions make the front-end and back-end parts unable to cooperate with each other smoothly and partition their responsibilities clearly, which further affect the schedule and quality of the project.

Last but not the least, since the typical developer of call centers has not got enough experience developing J2EE framework, the companies think it would be safer to let the leading J2EE solution venders take on the responsibility of the back-end information system. The cooperation of two infrastructures necessarily brings complexities and difficulties to separating the duties between them or finding out which one of them is responsible for the bug in the combination testing stage. Even the whole call center project is finished by one software company, the numerous complicated functions of the call center, such as three-part-meeting, will make the developer to translate and call different programming languages between two different frameworks, all of which increase the difficulties of the project significantly.

4. A NEW SOFTWARE ARCHITECTURE OF CALL CENTER

As we have discussed, the typical call center's software system framework cannot meet the new challenges and current solutions are not satisfying enough. Thus, a new system framework is imperative. We propose a totally new solution, to introduce into the typical call center software framework a new layer, a contract-based interlayer, the main part of which is a communication protocol followed by both the front-end and back-end project members. By following the same communication protocol, the front-end and back-end project members can work on their own part separately, and finally finish the integration of the front-end and back-end in a really short time. This method is successfully implemented in the call center project for the TCL Company, the biggest household electronic appliance manufacturer in China.

Our Novel Approach: The Contract-based Layer
In the new system framework, we partition the whole call center into four tiers [Figure 2]:

1) Hardware-integration Tier;
2) Service-aggregation Tier;
3) Contract-based Tier;
4) Business-process Tier.

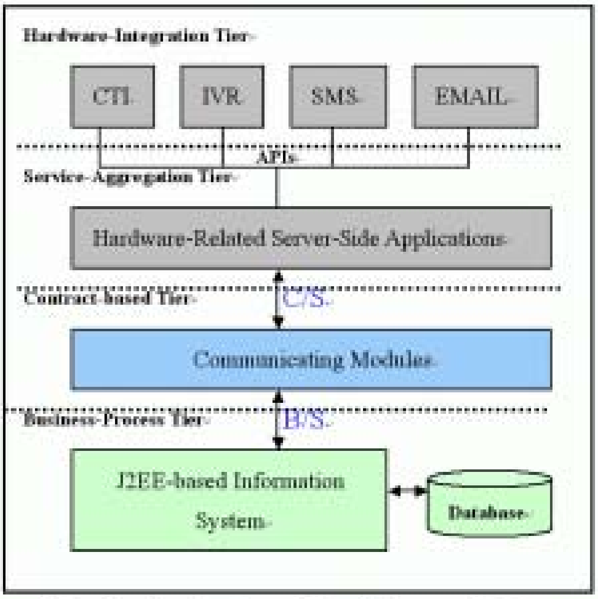


Figure2 call center's new 4-tier software architecture

Among the four tiers, the new service-aggregation tier is different from the old one for it renders the business-related functions; for a call center which mainly provides maintenance service for a company's products, its business-related functions include obtaining the information of the customer and its history of maintenance, committing the maintenance apply form, filling the dispatch form, collecting the customers' feedback and monitoring the process. Now the business-process tier realizes these parts of business-related functions and the agent takes care of whole business process by the webpage browser whose framework is based on J2EE.

In our project, the information system within this tier is called a Customer Service Management System. As the extension of the call center's information system, it provides the centralized management of the customers' files and helps the company's customer-related departments, such as the marketing, sales, service and technology support sections with their decision-making process, which means a perfect integration of functions of the old CRM system. Moreover, it integrates with the intranet information-publishing platform, the financial software platform and the integrative business platform to form a united company information service platform based on the J2EE framework.

In the rest part of our paper, the front-end will stand for the hardware-integration tier and the service-aggregation tier of which and the agent a typical C/S framework consists; the back-end will represent the business-process tier of which and the agent's browser a typical B/S framework consists.

The reason why the old system cannot meet the new demands is that it slides over the needs to partition the front-end and the back-end, of which the call-related functions integrate too closely. Thus, in order to integrate the two different frameworks (C/S and B/S), we have to find out all the information exchanged between these two frameworks and then define the possible exchange process, and ensure that there are clear partitions. We achieve it by introducing a totally original tier--the contract-based tier, which can make the service-aggregation tier and the business-process tier understand each other. In this contract-based tier, the front-end and the back-end both accept a common contract, which enables them to communicate in an understandable language. In a specific project, the literally contract-based layer is carried out by a function subset based on the company's requests. And both the front-end and back-end, according to the function subset, construct communication modules in their own programming languages which work together and build up the layer. Then these modules are belonging to their own framework, either the front-end or the back-end, so the relevance between the front-end and back-end becomes much less, which makes the developing work much easier.

The CC-CONTRACT Protocol
We decide to adopt the socket connection approach for the implementation of the contract-based interlayer according to the characters of the call center. The communication modules of the front-end and back-end send socket packages to each other's listening port to exchange their information. As the socket protocol is in a low level and don't need any other components, the agent workstation only has lightweight installation and desirable reaction efficiency. And the carrier of the contract-based layer followed by both part is a specific protocol that we call CC-CONTRACT.

Classified by the type of the exchanged data, there are four classes of package types in CC-CONTRACT protocol:

1) Information-exchange class, which enables the front-end server and the back-end server to exchange agents' information and password, as well as customers' importance ranks;
2) Voice-channel class, which enables the front-end and the agents to finish voice-related tasks such as answering an incoming call and start a three-part meeting.
3) Multimedia-channel class, which enables the agent to communicate with customers using E-mail and fax.
4) Error-caution class, which defines the error types.

They can be further divided into twenty-eight subclasses according to their specific functions, which together cover every possible information exchange type and every possible interaction process between the front-end and the back-end. The back-end part applies the certificated Java applet which is embed in JSP pages to open up the local agent's listening port for the socket package sent by the front-end and meanwhile sends socket package to the server at the front-end. The front-end can choose to install client application on the same agent workstations to receive socket package from applet (figure3a, mode 1) or assemble a given sever to receive the socket packages from all applets on agents' browsers and then realize interactions (figure3b, mode2). Under the latter condition, there is no need for agents to install any client application on their workstations, which affords unprecedented convenience for the agents' layout.

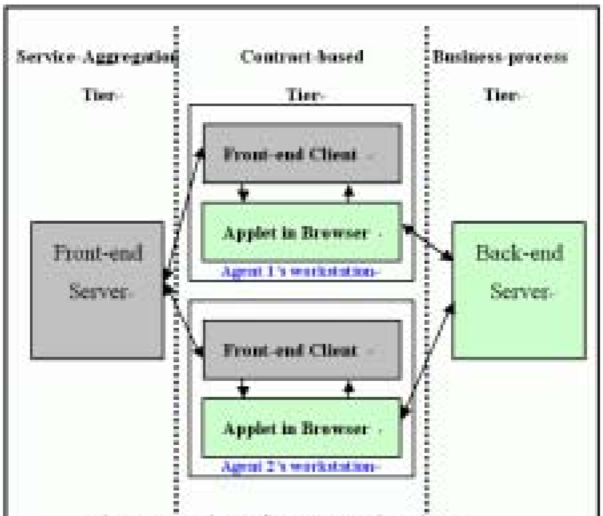


Figure3a mode 1 of agent's workstation

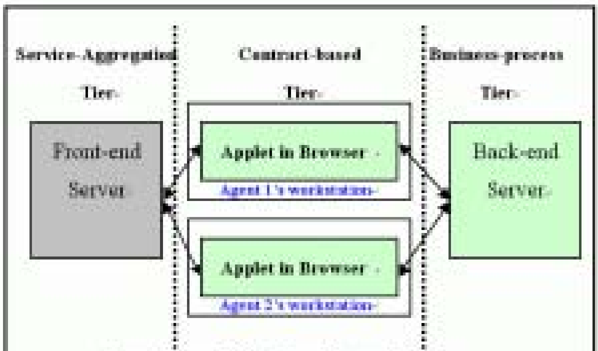


Figure3b mode 2 of agent's workstation

Typical Interaction Procedures Between the Front-end and Back-end

A typical interaction procedure of answering an incoming call (the "screen pop" procedure) using CC-CONTRACT protocol can be like the followings:

1) Based on call routing logic, the front-end server sends the request of answering an incoming call to an agent. Then the applet on the webpage displays the incoming call number and prompts the agent to answer the call.
2) After the agent click the button for answering the call, the applet sends the incoming call number to the servlet on back-end and the servlet will display the customer's data sorted by the incoming call number on the agent's browser.
3) The applet informs the front-end server that the agent has answered the call, and then the front-end server enables voice conversation between the agent and the customer.

A more complicated interaction procedure, three-part meeting, can be like the followings:

1) When agent A feels that it cannot manage to tackle the problem of customer C, it invites the agent B to a three-part meeting. After agent A click the button for three-part meeting, the present webpage's content of agent A will be send to the servlet on back-end and the present webpage's applet send the request of three-part meeting to the front-end server.
2) After the front-end sever redirect the request to agent B and agent B accept the invitation, the webpage's applet of agent B send the socket package of acceptance of the request to the front-end sever and request the back-end for agent A's webpage content.
3) When the sever knows Agent B have accepted the request, it will add Agent B to the line between Agent A and Agent B and then these three part can start their voice conversation.
4) After agent A receive the socket from the front-end sever indicating the success of the three-part meeting, it begins to send the content of its webpage consistently to back-end sever from which Agent B synchronously gets the content of Agent A's webpage. Now, Agent A and Agent B can share the same webpage renewed every second at the same time.

From the above example, we can draw the conclusion that because of the introduction of the contract-based layer we partition the front-end and back-end so clearly that the information which has to be exchanged between the front-end and back-end (such as incoming call number and three-part meeting's invitation and the acceptance), the tasks which are taken on only by the front-end (such as carrying out the voice conversation) and the tasks which is the responsibilities of the back-end (such as display customers' information on the webpage, and the synchronously renew of the webpage content) are quite separated from each other.

5. ANALYSIS OF THE NEW ARCHITECTURE

Minimal Relevance between the Front-end and The Back-end

The introduction of the contract-based layer to the old framework surely adds a loose coupling tier between the front-end and back-end, thus provides a clear partition of responsibilities between them. By doing so, the front-end and back-end developers can work on their own responsibilities separately after reaching an agreement on mutual communication while sparing no more effort to unnecessary interaction with each other. When both the front-end and back-end projects have gone through the unit-test, our mode will ensure least time and effort for the integration of them. At the same time, once there is a bug, our mode can find out who is responsible for it and correct it swiftly; as in the implementation of our project, it only takes examining the data in the socket package. Not only fully utilizing the speedy reaction of C/S mode and the remarkable expansibilities of the B/S mode, but this mode also enables the company to be less dependent on the exclusive developer since the company only have to make sure the new front-end or back-end system correctly understand and then carry out the agreement on mutual communication when upgrading the call center, which gives the company the most feasibility.

Favorable Expansibility
The introduction of contact-based layer makes the setup of agent workstation most convenient. In the mode two, the agent workstation even doesn’t need additional client application installation, which enables the company to add both local agent and remote IP agent easily while the company needn't face the complicated application installation procedure. At the meantime, the back-end J2EE framework ensures the stability and the efficiency of the system after enlarging the number of agents.

Besides, this mode enables those customers who used to visit the company's website for help to get help also from the call center's agent. The lack of real-time customer service on the Web site requires company to provide a mechanism by which real-time interaction can take place between the Web surfer and the corporation [10]. Since most companies have got calls centers, either as an internal operation or as an outsourced function, merging a Web site with a call center is an effective solution [6]. The customers used to have to provide their phone numbers to wait for the agent to call back or to install Internet phone software such as NETMEETING and then click the 'contact-us' button on the webpage. While in the new framework for the introduction of contract-based layer, the real-time communication between the customer and the agent can be achieved by the applet on the webpage. What's more, for the call center uses the same B/S mode, the agent can visit the same webpage as the user and keep the synchronous renew with the user's browser through the similar function as three-part-meeting in order to help the user to accomplish the webpage operations. The introduction of contact-based layer makes the company's website much more user-friendly, which means the customer desktop need no more installation other than a web-browser to finish all the functions. This was unthinkable in the old pattern of software architecture.

Extraordinary Performance
Contract-based interlayer achieved by socket has little affection on the system's efficiency. In the new mode, the call center's front-end server's C/S mode connection with the agent’s desktop client application has been replaced by its socket connection with the Java applet on the webpage. This substitution causes no lose of efficiency, for the reaction time only depends on the performance of the front-end hardware and the back-end J2EE framework. The success of our project has proved it: In spite of the fact that 6 branches of TCL Company’s call center scatter in 6 cities of China, which altogether have 500 agents and the annual volume of dealing records can achieve up to 10,000,000 pieces, however, a bout of interaction between the front-end and back-end, even in rush hours, can be finished in milliseconds.

6. CONCLUSION

Call center plays such a significant role in the company's strategy that a successfully implemented call center can bring immeasurable profit to the company. While the emergence of customer's needs enlarge the range of call center's functions and the J2EE and other company system frameworks mature, the old software architecture cannot survive the new conditions. By the partition of the call center's software system and the introduction of contract-based layer, we combine the C/S and B/S mode well in the call center project which thus has got high performance, favorable expansibility, more feasibility and stability .In the call center project for the TCL corporation, we implemented this contract-based layer for the first time and have achieved extraordinary outcome, which means it may become a successful standard for the industry in the near future.

7. REFERENCES

[1] AT&T. As reported on http://callcenternews.com/resources/statistics.shtml

[2] John Silling, “CTI Piece by Piece,” Byte Magazine, February 1997.

[3] Tom Keating, “An API Refresher Course,” CTI Magazine, August 1998.

[4] Richard Grigonis, “TAPI 3.0 Tutorial” Computer Telephony, p. 64, February 1999.

[5] Greg Borton, “CTI Beyond the Screen-POP”, Business Communications Review Volume 27, Number 6, June 1997

[6] Howard G. Bernett, Areg Gharakhanian. “Call Center Evolution: Computer Telephone Integration and Web Integration”, The Telecommunications Review, MitreTek Systems, pages 107--114, 1999.

[7] S Comella-Dorda. K Wallnau, R. Seacord, and J Robert. “A Survey of Legacy System Modernization Approaches”. SEI Technical Note CMU/SEI-00-TN003. Software Engineering Institute, Carnegie Mellon University, Apr. 2000.

[8] Bergy J.K, Northrop L.M., Smith D.B., “Enterprise Framework for the Disciplined Evolution of Legacy Systems” , Software Engineering Institute Carnegie Mellon University, CMU/SEI-97-TR-007 ESC-TR-97-007, October 1997

[9] Weiderman, N., Tilley, S. and Smith, D. "Approaches to Legacy System Evolution", Technical Report CMU/SEI-97-TR-014, 1997

[10] “Web Customers Abandon Shopping Carts,” Call Center Magazine, August 1999.

Wu Cen is a Master candidate in School of Mathematical Sciences, Peking University. His research interests include legacy system modernization and agent-based supply chain management.

Design and Implementation of a General Secure Extensible Payment Gateway Architecture

Bo Meng[1] Qianxing Xiong[2] Huanguo Zhang[1]
[1]College of Computer Science, Wuhan University Wuhan, Hubei 430072 P. R. China
[2]College of Computer Science and Technology, Wuhan University of TechnologyWuhan, Hubei 430063 P. R. China
Email: mengbo@263.net.cn qxxi@ public.wh.hb.cn liss@whu.edu.cn
Tel: 027-87885922 027-86551711 87885922-2494

ABSTRACT

With the development of electronic business, more and more enterprises do electronic business. The last work of doing electronic business is the electronic payment. The bank and financial organization generally use the secure financial network to deal with inter financial transactions. They generally provide the financial service with payment gateway, thus we can execute our payment transactions through Internet. In general the enterprises developing electronic business systems have several financial accounts in different bank or financial organization. At the present time every bank has itself payment gateway and these payment gateways are not compatible each other, which increases the difficulty and complexity of design and implementation of the enterprise's payment system. So in this paper we present a general secure extensible payment gateway architecture. In addition we give its prototype implementation applying web service technology and XML technology. The enterprises use it to develop its payment system instead of using several different payment gateways to implement its payment system. Thus the difficulty and complexity of the design and implementation of the payment system is decreased.

Keywords: payment gateway, web service, DataCash, payment gateway manager

1. INTRUDUCTION

With the development of electronic business, more and more enterprises want to implement electronic business. The last work of doing electronic business is the electronic payment. The bank and financial organization generally use the secure financial network to deal with inter financial transactions. They generally provide the financial service with payment gateway, thus we can execute our payment transaction through Internet. In general the enterprise developing electronic business system have several financial accounts in different bank or financial organization. At the present time every bank has itself payment gateway such as ASSIST Internet payment gateway [1], DataCash bank payment gateway [2], CyberCash payment gateway [3] etc, has itself payment gateway and these payment gateways are not compatible each other, which increase the difficulty and complexity of design and implementation of the enterprise's payment system. Hence in this paper we present general secure payment gateway architecture and give its implementation applying web service [4] technology and XML technology. The enterprises can use it to develop its payment system instead of using several different payment gateways to implement its payment system. Thus the difficulty and complexity of the design and implementation of the payment system is decreased.

2. A GENERAL SECURE EXTENSIBLE PAYMENT GATEWAY ARCHITECTURE

The general secure payment gateway architecture in Figure 1 is composed of the enterprise payment system, payment gateway manager, payment gateway interface and bank payment gateway. The payment gateway interface and the payment gateway manager are the core parts of the general secure payment gateway architecture. The payment gateway interface is the bridge of the bank payment gateway and the payment gateway manager. The payment gateway manager interacts with the given bank payment gateway through the given payment gateway interface.

First the enterprise payment system interacts with the payment gateway manager, then the payment gateway manager calls the corresponding payment gateway interface. The payment gateway interface communicates with the relevant bank payment gateway and gets the results and returns it to the enterprise payment system through the payment gateway manager.

We develop the payment gateway manager and the payment gateway interface as web services. So we need not change its architecture when we add a new payment gateway interface In a addition the developer of the enterprises payment system can use it expediently. In the bellowing we describe the development of web service of the DataCash payment gateway interface and the payment gateway manager.

3. WEB SERVICE TECHNOLOGY

A Web service [4] 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 [5,6,7]). Other systems interact with the Web service in a manner prescribed by its description using SOAP [8,9] messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.

A Web service is an abstract notion that must be implemented by a concrete agent. The agent is the concrete piece of software or hardware that sends and receives messages, while the service is the resource characterized by the abstract set of functionality that is provided. To illustrate this distinction, you might implement a particular Web service using one agent one day, and a different agent the next day with the same functions. Although the agent may have changed, the Web service remains the same.

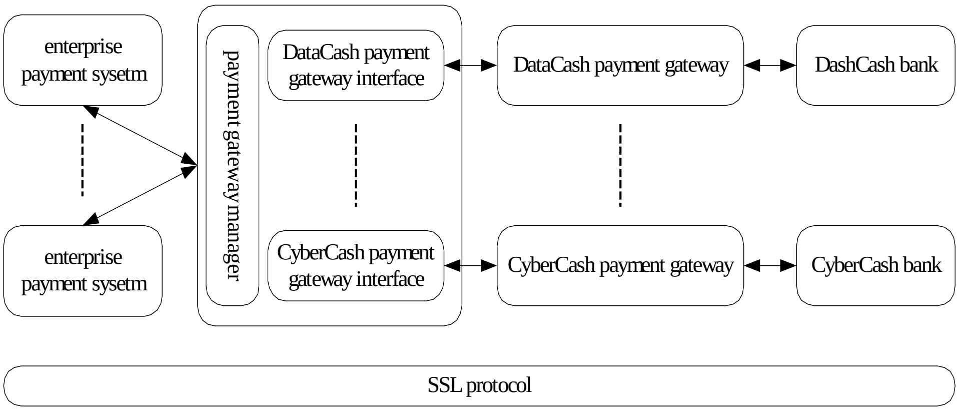

enterprise payment sysetm
enterprise payment sysetm
payment gateway manager
DataCash payment gateway interface
CyberCash payment gateway interface
DataCash payment gateway
CyberCash payment gateway
DashCash bank
CyberCash bank
SSL protocol

Figure 1 The general secure extensible payment gateway architecture

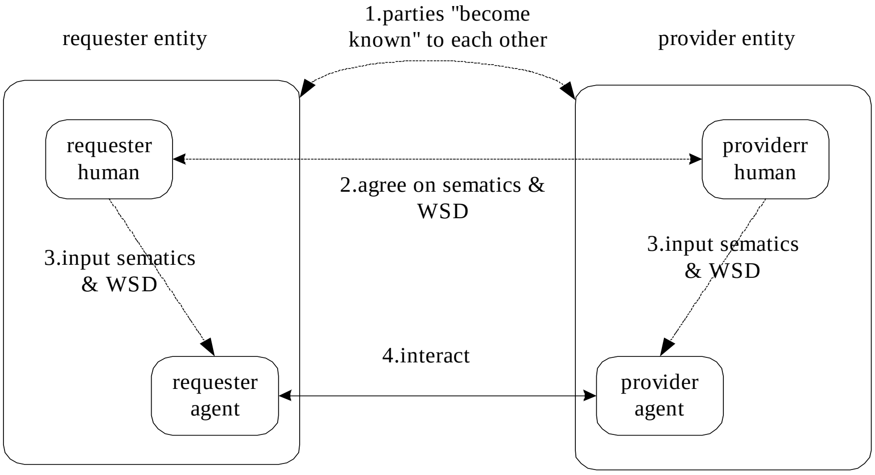


Figure 2 The general process of engaging a web Service

There are many ways that a requester entity might engage and use a Web service. In general, the following broad steps are required, as illustrated in Figure 2[4]: (1) the requester and provider entities become known to each other; (2) the requester and provider entities somehow agree on the service description and semantics that will govern the interaction between the requester and provider agents; (3) the service description and semantics are realized by the requester and provider agents; and (4) the requester and provider agents exchange messages, thus performing some task on behalf of the requester and provider entities.

4. PROTOTYPE IMPLEMENTATION

The prototype implementation of the general secure extensible payment gateway applies XML technology and web service. The operating system is Windows 2000. The web server is IIS 5.0. The develop tool is visual Studio.NET. The Transaction records of the general secure extensible payment gateway are described in Figure 3.

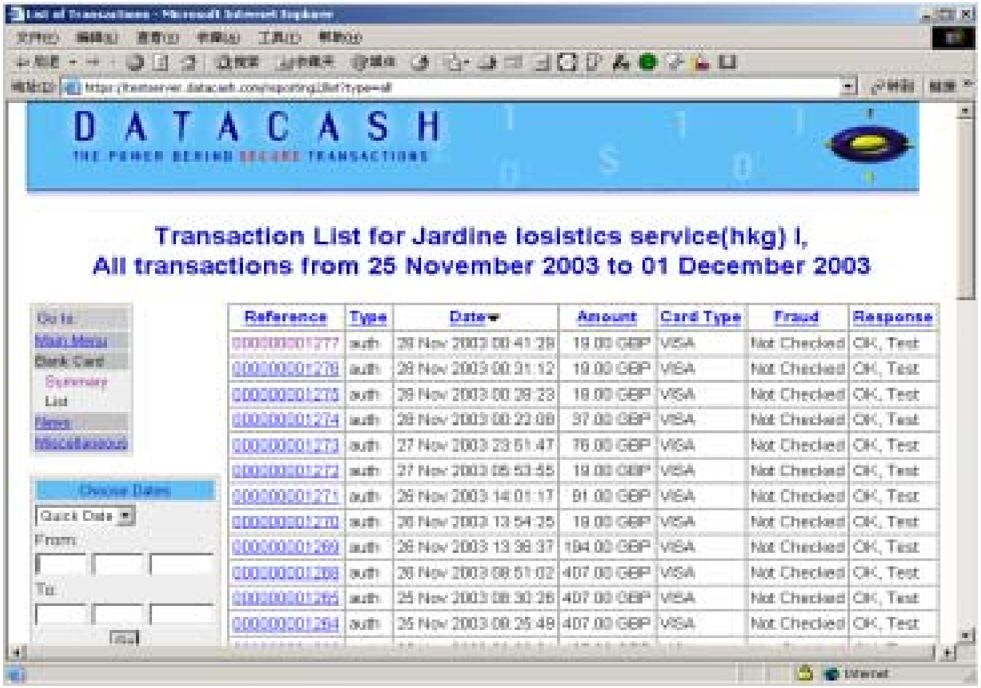
4.1 Payment Gateway Manager Web Service

The payment gateway manager web service in Figure 4 is the key part of the general secure extensible payment gateway. It is responsible for calling the corresponding bank payment gateway interface web service according to the request of the enterprise payment system and return it's the response to enterprise payment system.

4.2 DataCash Payment Gateway Interface Web Service.

DataCash payment gateway interface web service receives the information from the payment gateway manager, creates the information including DataCash username, password, unique transaction reference number, amount, amount currency, Transaction mode, credit card number and credit expiry date, Then sends the information to the DataCash payment gateway, receives the response information including status code, reason, transaction time, transaction mode from the DataCash payment gateway, sends it to the payment manager. We use the DataCash API [10] in the development of it.

The following is the core code of the DataCash payment gateway interface web service.

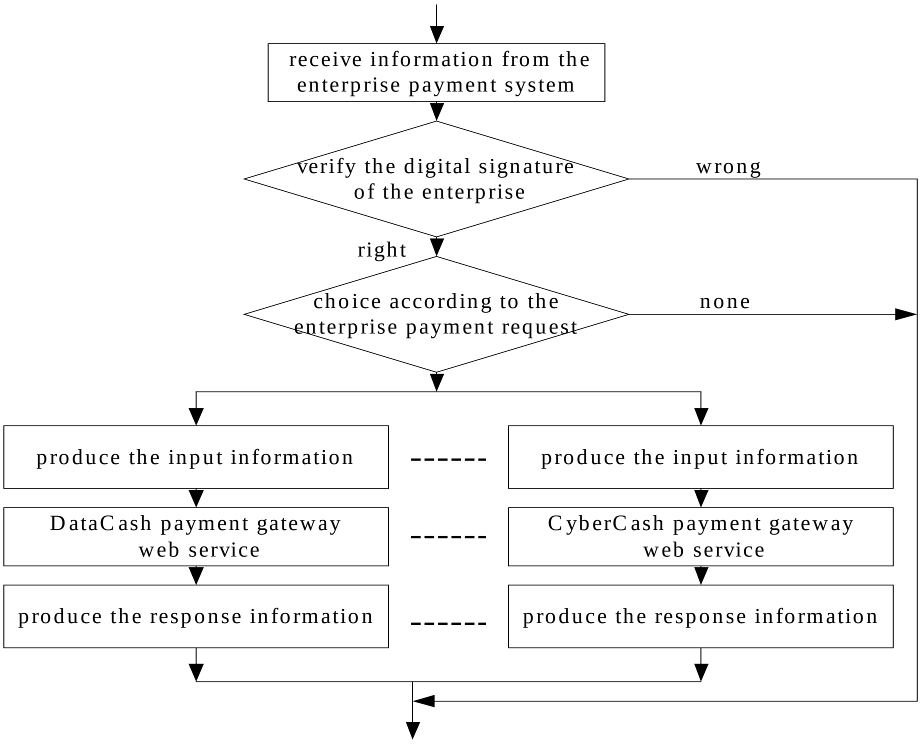


List of Transactions - Microsoft Internet Explorer
https://testserver.datacash.com/reporting2/list?type=all
DATACASH
THE POWER BEHIND SECURE TRANSACTIONS
Go to:
Main Menu
Bank Card
Summary
List
News
Miscellaneous
Choose Dates
Quick Date
From:
To:
Go
Internet

Transaction List for Jardine losistics service(hkg) l,
All transactions from 25 November 2003 to 01 December 2003

Reference	Type	Date	Amount	Card Type	Fraud	Response
000000001277	auth	28 Nov 2003 00:41:28	19.00 GBP	VISA	Not Checked	OK, Test
000000001276	auth	28 Nov 2003 00:31:12	19.00 GBP	VISA	Not Checked	OK, Test
000000001275	auth	28 Nov 2003 00:28:23	18.00 GBP	VISA	Not Checked	OK, Test
000000001274	auth	28 Nov 2003 00:22:08	37.00 GBP	VISA	Not Checked	OK, Test
000000001273	auth	27 Nov 2003 23:51:47	76.00 GBP	VISA	Not Checked	OK, Test
000000001272	auth	27 Nov 2003 05:53:55	19.00 GBP	VISA	Not Checked	OK, Test
000000001271	auth	26 Nov 2003 14:01:17	91.00 GBP	VISA	Not Checked	OK, Test
000000001270	auth	26 Nov 2003 13:54:25	18.00 GBP	VISA	Not Checked	OK, Test
000000001269	auth	26 Nov 2003 13:38:37	194.00 GBP	VISA	Not Checked	OK, Test
000000001268	auth	26 Nov 2003 08:51:02	407.00 GBP	VISA	Not Checked	OK, Test
000000001265	auth	25 Nov 2003 08:30:26	407.00 GBP	VISA	Not Checked	OK, Test
000000001264	auth	25 Nov 2003 08:25:48	407.00 GBP	VISA	Not Checked	OK, Test

Figure 3 Transaction records of the general secure extensible payment gateway



receive information from the enterprise payment system
verify the digital signature of the enterprise
wrong
right
choice according to the enterprise payment request
none
produce the input information
------
produce the input information
DataCash payment gateway web service

CyberCash payment gateway web service
produce the response information
------
produce the response information

Figure 4 The flowchart

```
Sub ReqRes(ByVal DataCashClient As String, ByVal DataCashPassWord As String, ByVal MerchantReference As String, ByVal Amount As String, ByVal Currency As String, ByVal CardNumber As String, ByVal ExPiryDate As String, ByVal DataCashUrl As String)
        Dim request As DataCashRequest
        Dim response As DataCashResponse
        request = New DataCashRequest()

        request.Authentication.Client = DataCashClient
request.Authentication.Password = DataCashPassWord

        request.Transaction.TxnDetails.MerchantReference = MerchantReference
        request.Transaction.TxnDetails.Amount.Amount = Amount
        request.Transaction.TxnDetails.Amount.Currency    =
"GBP"
        request.Transaction.CardTxn.Method = "auth"
        request.Transaction.CardTxn.Card.CardNumber       =
CardNumber
        request.Transaction.CardTxn.Card.ExpiryDate       =
ExPiryDate
        response = request.GetResponse(DataCashUrl)
        status = response.Status
        reason = response.Reason
End Sub
```

5. CONCLUSIONS

With the development of electronic business, more and more enterprises do electronic business. The last work of doing electronic business is the electronic payment. The bank and financial organization generally use the secure financial network to deal with inter financial transactions. They generally provide the financial service with payment gateway, thus we can execute our payment transaction through Internet. In general the enterprises developing electronic business systems have several financial accounts in different bank or financial organization. At the present time every bank has itself payment gateway and these payment gateways are not compatible each other, which increases the difficulty and complexity of design and implementation of the enterprise's payment system. So in this paper we present a general secure extensible payment gateway architecture. In addition we give its prototype implementation applying web service technology and XML technology. The enterprises use it to develop its payment system instead of using several different payment gateways to implement its payment system. Thus the difficulty and complexity of the design and implementation of the payment system is decreased.

6. REFERENCES

[1]. http://www1.assist.ru/eng/find_tranz.htm

[2]. https://reporting.datacash.com/reporting2/login

[3]. http://www.cybercash.com

[4]. Web Services Architecture. W3C Working Draft, 8 August 2003. http://www.w3.org/TR/ws-arch/

[5]. Web Services Description Language (WSDL) Version 1.2: Part 1: Core Language, http://www.w3.org/TR/wsdl12/, W3C Working Draft, 11 June 2003

[6]. Web Services Description Language (WSDL) Version 1.2 Part 2: Message Patterns, http://www.w3.org/TR/wsdl12-patterns/, W3C Working Draft 11, June 2003

[7]. Web Services Description Language (WSDL) Version 1.2 Part 2:Bindings, http://www.w3.org/TR/wsdl12-bindings/, W3C Working Draft, 11 June 2003

[8]. SOAP Version 1.2 Part 0: Primer, http://www.w3.org/TR/soap12-part0/

[9]. SOAP Version 1.2 Part 1: Messaging Framework, http://www.w3.org/TR/soap12-part1/

[10]. https://testserver.datacash.com/software/datacash_overview.html

Meng Bo is a postdoctor of college of computer of Wuhan University. He gets his PH.D in Wuhan University of Technology in 2003.

Now his research interests are in electronic commerce, information security and electronic government.

The Study on Exchange Platform of Network Manufacturing Products Based on Digital Watermarking Techniques and Fair Exchange Protocol

Zude Zhou[1], Zhiyang Wang[1], Quan Liu[2]
[1]School of Electron-mechanical Engineering, Wuhan University of Technology
[2]School of Informational Engineering, Wuhan University of Technology
Wuhan, Hubei, China
[1] zudezhou@mail.whut.edu.cn **Tel:** 027-87651445
[1] wzhiyang.student@sina.com **Tel:** 13554241021
[2] qliu@public.wh.hb.cn **Tel:** 13908659571

ABSTRACT

An exchange Platform of network manufacturing products is presented on the basis of study of digital watermarking techniques and fair exchange protocol. Many safety measures are adopted in the platform on copyright, fair exchanges, safety of data transmission, access control and thereby it can better meet the need of network manufacturing products.

Keywords: Digital Watermarking; Fair Exchange; Electronic Commerce; Network Safety; PKI

1. INTRODUCTION

With the development of internet techniques electronic commerce has already gradually become a new access to business activity. More and more people take part in electronic commerce through internet. Internet changes enterprises' model of production, management and sales as well as people's living styles. Network manufacturing products exchange is an important branch of electronic commerce and network manufacturing products include generally multimedia products which are exchanged by enterprises, customers, providers or copartners. Security of the exchange Platform of network manufacturing products must be solved on practicality and commerce as that of the electronic commerce must do. It is a key problem for customers and enterprises how to establish a convenient and safe environment.

Nowadays, the way of copyright protection of digital multimedia is encryption algorithms and watermarking techniques. Only in the process of data transmission does encryption system go into effects. Data are accepted and decrypted the protection disappear. While digital watermarking techniques go into effects all the time [1]. So the digital watermarking techniques are effective supplementary ways to solve the copyright protection.

Fair exchange protocol is the foundation of establishing safe and fair network exchange platform. An exchange platform of network manufacturing products is presented on the basis of study of digital watermarking techniques and fair exchange protocol.

2. THE REQUESTS TO THE EXCHANGE PLATFORM OF NETWORK MANUFACTURING PRODUCTS

An exchange platform of network manufacturing products provides not only conveniences and quickness of exchanges but also securities. In details, it should meet following requests [2]:

①Justice. Fair exchanges are foundations of any of electronic exchanges. Any electronic exchanges will not exit without justice. Justice is known as both participants should exchanges honestly under normal conditions. Once one side performs deceits or can not perform promises or obligations, the other is able to defect and obtain the testimony of his dishonest operations or failures to perform the promises or obligations.

② Secrecy. It should be guaranteed that a lot of data transmitted on network aren't betrayed. Internet is a n open system interconnection, and if safe protections are not applied, the data transmitted on internet including e-mail and so on are likely to be detected and read by the third side, even modified on hostility. So how to guarantee data secret on network is a basic problem of establishing an exchange platform of network manufacturing products.

③Integrality. It should be guaranteed that the exchanged data transmitted on open network aren't be modified and the cheated exchanges by repeating sending are detected and aren't carried on. Internet is open and the persons with special knowledge and tools are likely to get and modify the data transmitted on open network. So it is necessary that the original formats and content of the data should be saved. The data acquired by the receiver are consistent with those send by sender.

④Identity authentication and authorization. In the process of electronic exchanges, how to identify both sides and to authorize right authorization is an important question. In this process, it is necessary to confirm the current operations legal. The identification of computer system on internet is made sure by IP address. Hacker might use other's IP to acquire data. It is very easy to send anonymous e-mails or use dishonest user's name in the normal e-mails. So establishing strictly identity system makes sure the identifications of all the partners legal and valid. That is, receiver can confirm data from not the fake third party but right sender. Sender must confirm receiver legal and valid in order not to send the data to improper receiver. Correlative authorization should be established properly.

⑤ Resistance. After the trade has been finished, it is important that any bargainer can not deny his/her preceding operations and records. They include original record and sent record aren't denied; it should be confirmed that the data have been sent and received; it should be forbidden that receiver denied acquired data to attempt delay next work. For the sake of security of exchange operations, the system can stand attaches from hackers and improper operations from the customers. The system can prevent and correct underling threats from network malfunctions, improper operations, hardware malfunctions, software malfunctions and computer viruses to make exchange data valid in the right moment and place

3. DIGITAL WATERMARKING TECHNIQUES AND FAIR EXCHANGE PROTOCOL

3.1 Digital Watermarking Techniques

Digital watermarking techniques provide measures to solve exchange platform of network manufacturing on justice, secrecy, integrality, identity authentication, and resistance. It offers supplementary safe measures to the exchange platform of network manufacturing.

Digital watermarking techniques are defined as that hidden mark is embedded in multimedia data in way of signals manipulation, and the hidden mark in the multimedia can not been “seen” by eye unless by special inspecting device or professional reading device.

Qualified watermarking need to meet following features:

①Directness. The watermarking should be embedded not in the header or end but in the data of the multimedia, for the information of the watermarking should be irrelevant to formats of the carrier. Once the watermarking is embedded in the header or end, the hacker may acquire the data from the header or end and the data can be saved as familiar format by the hacker. The information of the watermarking will be lost.

②Invisibility. Once the watermarking is embedded in the digital images, it should not be perceived by receiver by vision system and the qualities of the carrier such as the visual effect and reality of the image are not damaged. In theory, it is a just dream to not be perceived completely. Now the watermarking by use of the popular damaged compression algorithms based perception model is deleted easily and can’t reach the purpose of mark. Also, JPEG is the transmitted format of a great deal of images and JPEG is a typical damaged compression algorithm. Accordingly, it’s

practical that the embedded information which doesn’t damage carriers to a degree is not perceived by eyes

③Robustness. Digital marking must have much robustness to all kinds of signal manipulations. That is, the availability of the watermarking and the accuracy of identity retain fine after the multimedia suffered purpose or purposeless signal manipulations. In theory, the watermarking may be removed on condition of the authorization. A good watermarking can’t be removed at all or very hardly. For example, it’s not cost-effective to take too many calculations to removal the watermarking. Different application has different requests to robustness and watermarking normally resists to common image manipulation. Such as filter, the averaging column figures etc.

This paper presents a uniform embedded watermarking algorithm based wavelet packet transform and spread spectrum technique [3].

This algorithm has good robustness to mean value filter, crop, rotation subsampling, and JPEG compression. What’s more, it’s suitable with characteristic of human vision and meets newly international compression standard. It overcomes basically the susceptibility to image translations and scale transforms caused by wavelet transform.

In view of the good characteristic of spatial and spectrum domain wavelet packet transform is adopted. Image can be decomposed in different space domain and frequency band and processed image is still suitable with human vision system. The algorithm follows bellow in detail:①To uniform different watermarking with 64 dimensions symbol aggregates.②To analyze wavelet decomposing number and embedded watermarking result in order to acquire the best solution. To decide embedded watermarking position.③To analyze the maximal information amount which can be embedded in the carrier with band amount theorem. ④To embed watermarking in the wavelet packet domain according to spread spectrum communication.
The process of watermarking being embedded follows below (Fig.1):

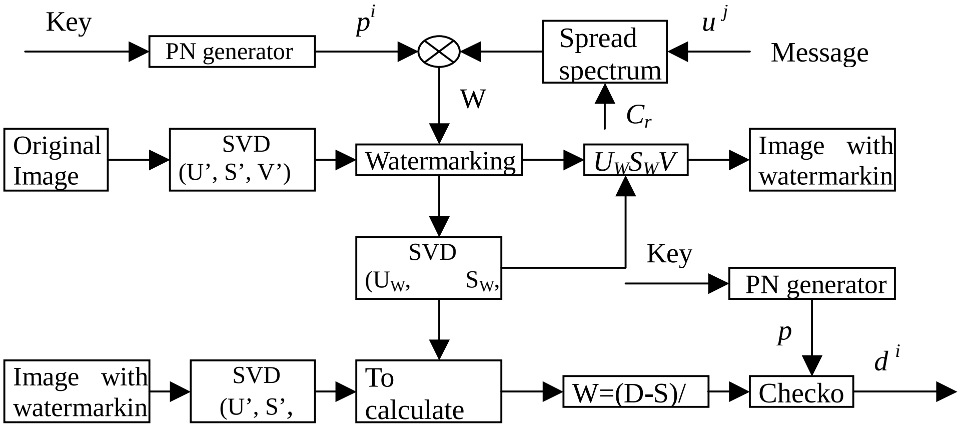


Fig.1 The digital watermarking process illustration

3.2 A Fair Exchange Protocol

A fair exchange protocol with off-line semi-trusted third party is adopted in the paper. In this protocol, the watermarking data are produced and transmitted by asymmetric key. The public or private watermarking is embedded in the images. The copyright centre can detects public watermarking, while private watermarking only can be detected by producer who embedded the watermarking. After the buyer acquired confirmation from the copyright authentication center, the exchange can be finished through the protocol [4] [5]. Considering five participants with the name H for the digital product owner, B for buyer, T for PKI (public key infrastructure), C for copyright authentication center, T for public directory service, the process of the exchange follows below in detail (fig.2):

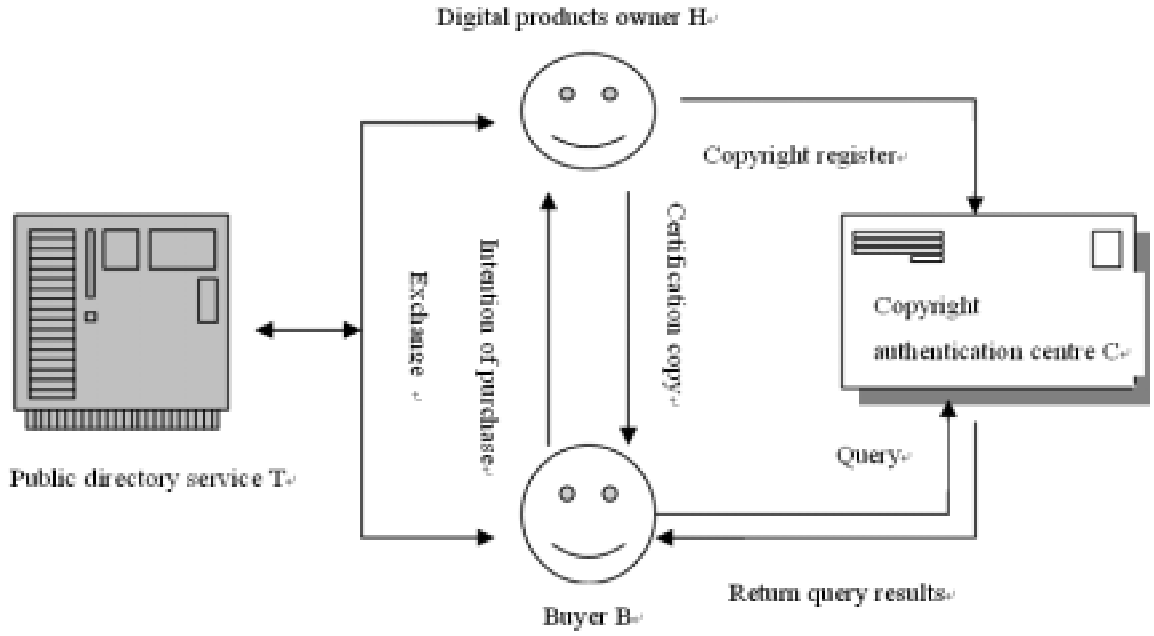


Fig. 2 a fair exchange illustration

Step 1: H register with C;
Step 2: B send the intention of the purchase to H;
Step 3: H send certificate copy to B;
Step 4: B query from C;
Step 5: C send query result to B;
Step 6: B exchange with C under the public directory service.

The watermarking carried image ownership information as a mark of owner copyright, it is produced through auto-adaptive spread spectrum. In order to embed and pick up watermarking, the seed value of the pseudo-random sequence is necessary. Different watermarking such as public watermarking, check watermarking and private watermarking are embedded the same image. Owner can easily found out right position of his/her image in internet with help of check watermarking carried with random seed. Public watermarking provides the copyright of the image. The secrecy depends on private watermarking with private key. After the image owner registers with copyright authentication

center with public key, the copyright authentication centre proves that he/she is holder of copyright of the image with private key and transmits the product to the buyer. Thus, once the illegal third party should claim that he/she holds copyright of the image, the copyright authentication centre can find out right holder with help of private key. The problem of the copyright ownership can be solved.

In the project, the third party is off-line. The commutative information between exchangers and the third party is little. Transmitting the need commutative information to two exchangers reduces probability of the third becoming bottleneck. The weakness is that we increase element and more data in public directory service.

4. AN EXCHANGE PLATFORM OF NETWORK MANUFACTURING PRODUCTS BASED ON DIGITAL WATERMARKING TECHNIQUES AND FAIR EXCHANGE PROTOCOL

The platform adopts browser/serve pattern, which is easily maintained and expanded compared with client/server pattern. B/S pattern can provide real integrated system service based different platforms and reduce greatly requests of network brand. The double arrow stands for intercommunicated information follows (fig.3)

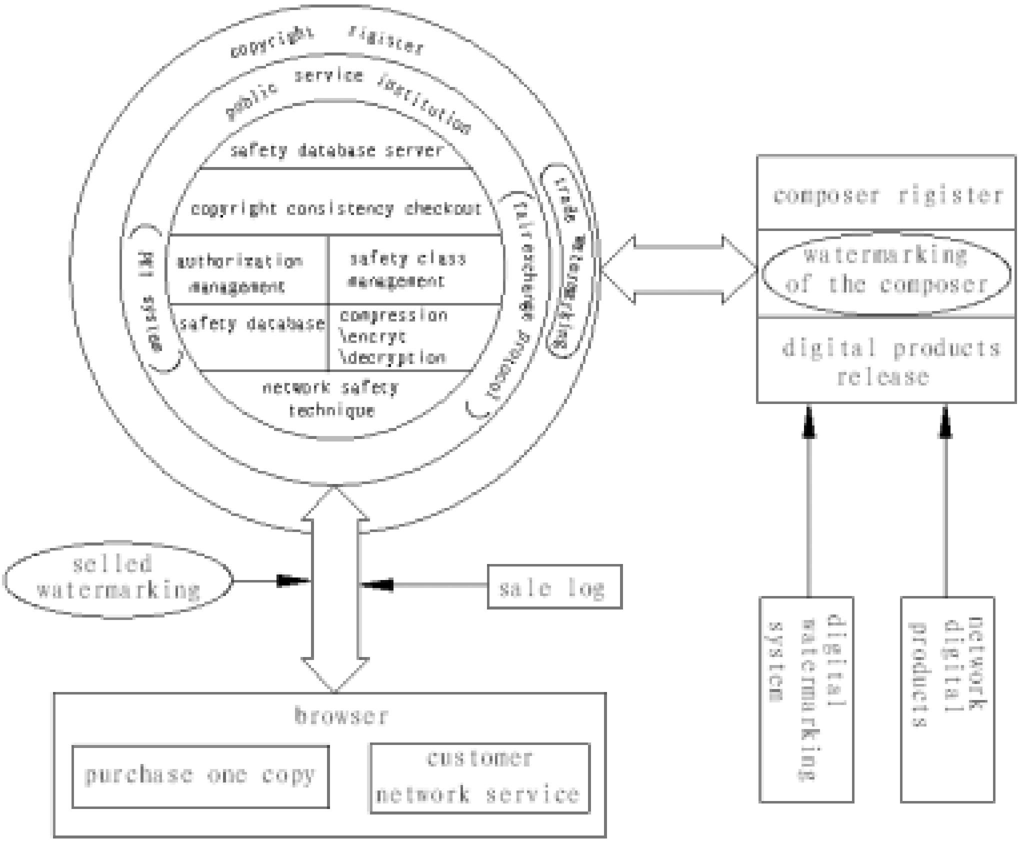


Fig. 3 the exchange platform of the network manufacturing products

The platform consists of the composer of digital product, publishing institution, trader, public service institution, and customer. The customer can purchase products/copies and acquire free or paid network service by means of the browser. The off-line semi-trusted fair exchange protocol is implemented by the public institution, and the platform adopts public key infrastructure [6] [7].

PKI is a secrecy system which provides online identity by use of public key theory and technique. It bases on uniform safety authentication standards and rules. PKI provides valid measures in identity authentication and authorization, integrality, resistance, etc. on network safety.

There is a crisis in public key. The sender of the information needs to identify the information by public key. If other used the third party's public key to encrypt the data, he wishes that the receiver don't understand the information, while the third party with corresponding private key can do that easily. In fact, it involves the key problem of PKI: how to confirm right person with right public key and private key. In PKI, in order to ensure the user's ID and keep user with private key, the public keys are managed by an independent and trusted third authentication center. In this platform, the public service institution performs this function. Having a safe and trusted authentication center, the users can enjoy safe service from PKI techniques.

In the safe database serve, the copyright data consistency check guarantees the consistency of the data of all related customers, media composers, and press businessmen. The server manages the authorization of composers, press businessmen. For example, a composer can only embed his/her own watermarking into the works. The password verification adopts cookies technique. The passwords of advanced class users adopt auto-generation technique. Not all the data need to be in strict secrecy. For example, some service can be acquired by all the customers, but others can be acquired by paid customers. Therefore scientific safety tactic is that the data should be managed in different class. For the sake of save and transmission, the transmitted data should be compressed, encrypted, decrypted. The access control list model is adopted in the platform, and access control can be class into coarse grained access control, medium grained access control, and fine grained access control according to the classification of the customers. The firewall techniques prevent illegal or unwanted information from outer network into the inner system.

5. CONCLUSION

This paper analyzes the requests to the exchange platform of network manufacturing products. A uniform embedded watermarking algorithm based wavelet packet transform and spread spectrum technique and a fair exchange protocol with off-line semi-trusted third party are used in the platform. It adopts many safety measures to meet network manufacturing products exchange such as distant access control, exchange justice, copyright protection, and database safety.

6. REFERENCES

[1] Stefan Katzenbeisser, Fabie A.P.Petitcolas, editors. QiuxinWu, interpret .Information hiding techniques for steganography and digital watermarking. Beijing .Posts & telecom press. 2001. pp72-73.

[2] Safety trading platform. Knowledge and Technology of Computer. June 2001. pp 59-60.

[3]Quan Liu, X.M.Jiang. Research on Adaptive Digital Watermarking Based on Spread Spectrum. the Third International Symposium on Multi-spectral Image Processing and Pattern Recognition. 2003.10.

[4] Asokan N, Janson P, State of the art in the electronic payment systems. IEEE Computer, 1997. 30(9):pp28-35.

[5] Xujun, Scheme for digital image e-trading based on fair exchange protocol. Journal of Computer-aided Design & Computer graphics. Vol.14, N02. February 2002. pp 153-157.

[6] Andrew Nash, William Duane, Celia Joseph, Derek brink, Editors. Yuqing Zhang, etc. interpret. PKI Implementing and Managing E-Security. Beijing. Tjinghua university press. Dec.2002

[7]Tang Z, Long Y, Chen Y, Zhou z. Information Secure Strategy Based on Digital Certificate of Virtual Enterprise. China Mechanical Engineering, Vol.14, No.3, pp.234-237, 10 Feb.2003

Zude Zhou, male, professor, tutor of doctor, is the president of Wuhan University of Technology, whose research interests are CNC Theory and technology, Intelligent Control, Digital Manufacturing, Reliability and Fault Diagnosis of the Modern Manufacturing Systems and etc.

Zhiyang Wang, male, is a postgraduate in School of electron-mechanical engineering, Wuhan University of Technology, his research interests are Watermarking and Digital Manufacture.

Data Mining System Based on Web Services for E-commerce: Architectonics and Algorithm

Luo Zhong, Qiwei Tong, Bin Fan, Chengming Zou, Qiong Jiang
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: qwtong@mail.whut.edu.cn **Tel.:** 86-27-87642927

ABSTRACT

As the Internet advances and with the great development of E-Commerce, companies involving in online shopping and services are eager for ways to find out what their customers want. Companies now with the help of data mining, they can have a much more active role in promoting their online shops, merchandise and services compares to the conventional method of waiting for customer to look at their sites. This paper discusses about the new development of E-Commerce that differs a lot from the old conventional commences. It also goes in depth of a new architecture of data mining system based on E-commerce web services. The conventional “single data-mining system” system is replaced by the new “distributed data mining system”. The new distributed data mining system is a system that provides mining services. In elaboration of the system, the E-Commerce Server request services in the network, Data Mining Server provides concurrent mining services for serial E-Commerce Servers (Serial E-Commerce Companies). Data Mining Server will summarize all the data in different E-Commerce Server for mining process. It will produce more comprehensive and informative knowledge than conventional methods. A new effective preprocessing algorithm of distributed data mining for E-Commerce is provided in this paper. According to the architecture given in this paper, a distributed data-mining algorithm for finding association rules is provided.

Keywords: e-commerce; distributed; web services; web mining; preprocessing algorithm

1. INTRODUCTION

Data mining is becoming a key technique in discovering meaningful patterns and rules for large amounts of data (Berry and Linoff, 1997). It is often used in the fields of business such as marketing and customer support operations.

Currently, most Web-based commercial applications are used in B2C e-commerce (i.e. online retailing transactions with individual shoppers), in which users can surf a company's Website to conduct transactions. With the great development of e-commerce, companies such as online retailing stores are eager for a method to find out what their customers want and much other related information. It is very important to make customers' interest active and dig their purchasing potential.
Techniques of web mining have recently been requested and developed to achieve this purpose. Cooley divided web mining into two types: web-content mining and web-usage mining. Web-content mining focuses on information discovery from sources in World Wide Web. On the other hand, web-usage mining focuses on the automatic discovery of user access patterns from web servers [6]. In the past, several web-mining approaches for finding sequential patterns and user interesting information from the World Wide Web were proposed [2,3,4,5].

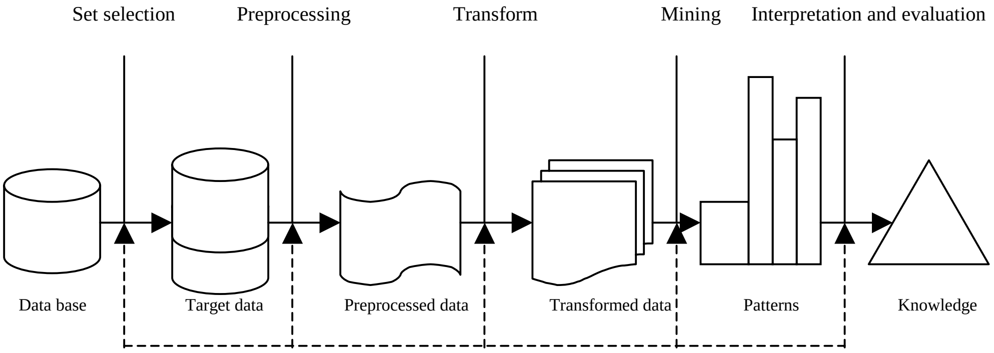


Fig. 1 An overview of the data mining process

2. BACKGROUND

Years of effort in data mining have produced a variety of efficient techniques. Depending on the types of databases to be processed, mining approaches may be classified as working on transactional databases, temporal databases, relational databases, and multimedia databases, among others [8].

Depending on the classes of knowledge sought, mining approaches may be classified as finding association rules, classification rules, clustering rules, and sequential patterns.
Data mining is a complex process involving multiple iterative steps. Fig.1 shows an overview of this process. [9]

Conventional preprocessing in data mining system for e-commerce is a complex process. Fig 2 shows an overview of those steps. According to the new character of electronic commerce, this paper gives a more efficient preprocessing algorithm of data mining for e-commerce.

3. THE ARCHITECTONICS OF A DATA MINING SYSTEM BASED ON WEB SERVICES FOR ELECTRONIC COMMERCE

Because of new character of e-commerce, which is different from conventional commerce, data mining system for e-commerce has distinct character:

1) The data mining system for e-commerce should deal with more data types and more complicated data models.
2) As far as computing environment of system is concerned, distributed and isomerous are going to be the model for future data mining system.

Considering the first character, optimize conventional algorithm is needed in data mining system. Considering the second one, a distributed, oriented network data mining system is needed.

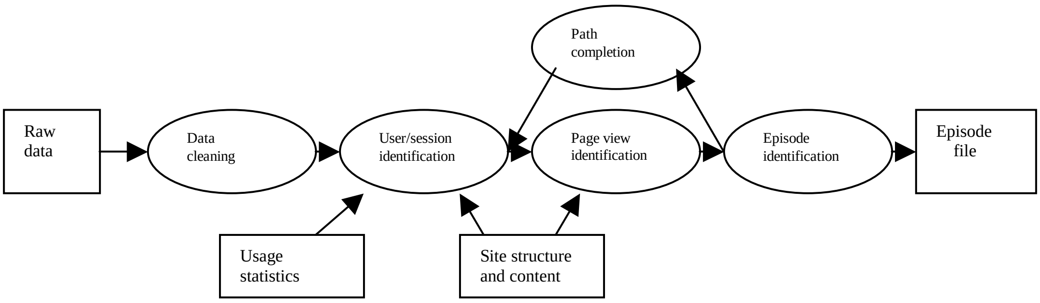


Fig. 2 An overview of the preprocessing in data mining system for e-commerce

As shown in Fig. 3, mining system is an individual system to provide mining services. E-commerce servers request services by network, data mining server gives mining services to serial e-commerce servers (serial e-commerce companies) at the same time.

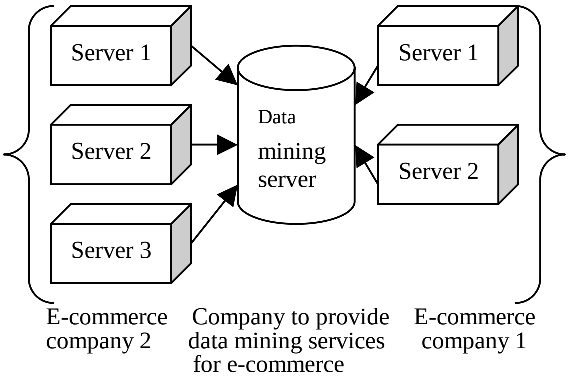


Fig.3 location of data mining system for e-commerce

Mining server can summarize all data in different e-commerce servers to mine, and then it will get more comprehensive knowledge. For example, in web-usage mining, web-usage knowledge for a user from one e-commerce server is simplex, but the knowledge for this user from a lot of e-commerce server will provide real information about this customer, then help the company to make better stratagem.

Conventional network application usually adopts C/S (Client/Server) structure, which is called two levels architecture. System like this is easy to be designed, but has many disadvantages. Client is restricted under this mode and this mode is unfit for the extended functions. With the development of technology, three layers mode structure was developed. Divides the system is divided into three different layers in three layers mode: client layer, operational logical layer and data accessing layer (application service providing layer). Client layer deals with UI; data accessing layer acts as data source, usually it refers to database; operational logical layer, which is added recently, is to make intelligent decision. In early years, the function isn't complex and often is applied in client layer and others are put in data accessing layer by means of memory procedure or trigger. However, with the development of software, software becomes more and more complex and the functions of software are extended most in this layer. If all extended functions are applied in client layer, the client layer may not be able to handle them. Therefore, it's suggested that operational logical should be separated from the system and become an individual layer. It becomes three layers structure. As shown in Fig 4, the system in the paper is three levels mode.

Client layer provides the function of data preprocessing. Considering the algorithm of data preprocessing is steady, it's put it into client layer so that the system running speed is higher. The data, which has been preprocessed, is put in the local database, so there is a problem: data mining system must access e-commerce database from a long distance, which delays the response. One of the solutions is to connect e-commerce and data mining services by high-speed networks. And data-mining algorithm is optimized.

Data integration in the operational logical layer is based on the result of the data preprocessing. For example, in web mining, the data of a user in different e-commerce website need to be integrate. Therefore, the other problem arises: the user who has the same name maybe isn't the same person in the different e-commerce website. One of the solutions is that apply the authentication by real name. At present, there are some e-commerce websites have realized it in china, such as Eachnet. Another solution is that the user authentication of e-commerce is authenticated by attestation services. If the customer set provided by authentication services is R, while the customer set provided by data mining services is M, then the system would make sure that $M \subseteq R$.

Mining module of the operational logical layer is an open aggregate of mining tools, which can be extended with new modules and new functions easily to satisfy user demands. The result, which produced by mining tools, directly returns to the client layer to output.

Architectonics of this Data Mining System for Electronic Commerce is shown in Fig.4.

4. OPTIMIZATION FOR DISTRIBUTED DATA MINING ALGORITHM FOR E-COMMERCE

4.1 Optimization for Preprocessing Algorithm

Data of e-commerce is abundant and complicated, and the data mining system in this paper is distributed, a more efficient preprocessing algorithm is needed, which is optimized for distributed e-commerce. The cost of this algorithm and conventional preprocessing algorithm are the same, but only one step is needed in this algorithm.

Identification of user is ignored in this algorithm. As user names are needed in almost all e-commerce websites, it's supposed that there are user names in web logs. In this algorithm all links are checked orderly in web logs, when a new user name arises, a new session begin. When that user name arises again, the link in web logs will be transmitted to the same session process. Detailed flow is shown below.

Session begins

1. Next log
2. If $(t-t_{last})>T$ and $t_{last}=t$, begin a new session, exit
3. Is this web page accessed from last page? Yes: this page joins stack; this page joins the session. Go to step.1
4. Backdate in the stack, does exist any page from which it can access the current page? Yes: the page in the stack joins the session; current page joins the session; adjust the stack. Go to step.1
5. Begin a new session, exit

In this algorithm, there is a stack. This stack stores page ID in the session. When user surfing the website, these page will be stored in the local buffer, and the stack helps data mining server to get surfing full path. T in the algorithm refers to the limit interval time between two sessions. Generally, it is 25 minutes.

4.2 Optimization for Distributed Data Mining Algorithm for E-Commerce (searching association rules)

As the architectonics of this Data Mining System provided in this paper, if conventional data mining algorithm is applied, time in net transmission will bevery long. A new distributed mining algorithm is developed to meet the need of e-commerce in this paper. There are five servers in Fig 1, every server has a web log, and we can use an optimized Apriori algorithm [1]: every server mines in local place, and the data mining server synthesize all results to get the integrated, correct conclusion. If the minimum support value for whole system is X, minimum support value for local place is X too, this can be proved easily.

Here is a simple example: there are 3 servers, in which data has been preprocessed. The minimum support is 40%, only run one step.

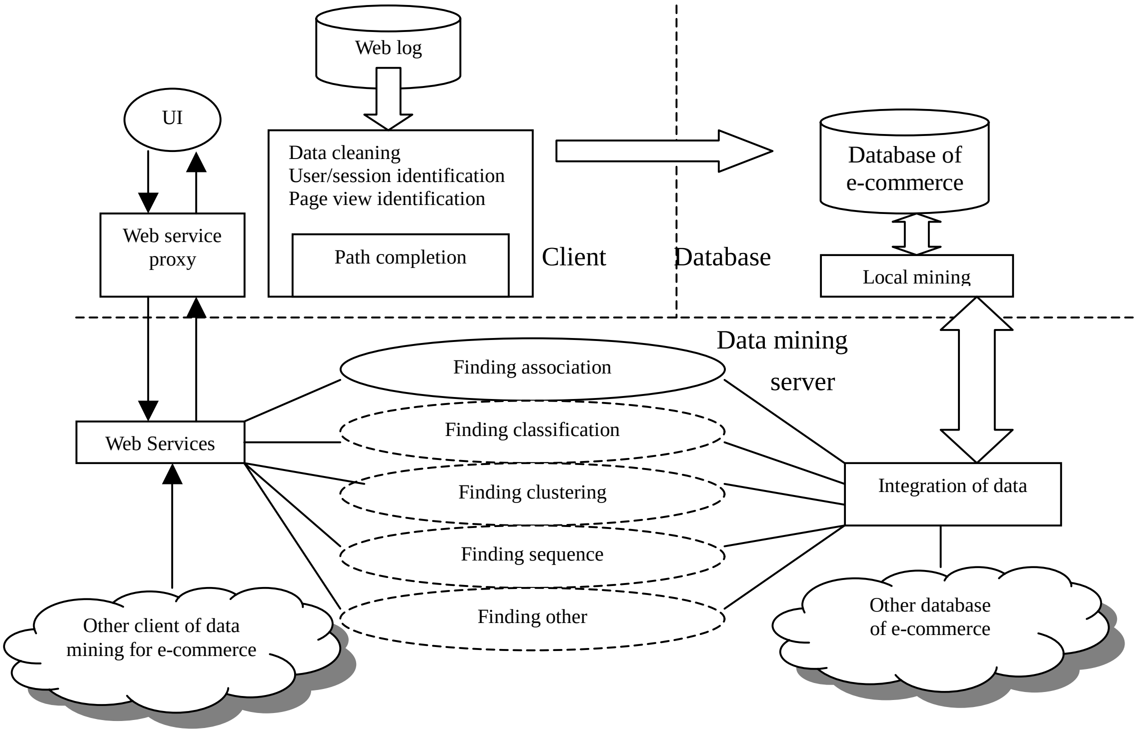


Fig.4 Architectonics of this Data Mining System for Electronic Commerce

Server1

User	Session
1	1,2,5
2	2,4
3	2,3
4	1,5
5	2,3,5

Server2

User	Session
6	1,2,4
7	1,3
8	2,3
9	2,4
10	1,5

Server3

User	Session
11	1,3
12	1,2,3,5
13	1,2,3
14	2,3
15	2,3

Every server mines for one step, and frequent 1-item set is obtained as below:
Server1: 1(40%), 2(80%), 3(40%), 5(60%); 4(20%)
Server2: 1(60%), 2(60%), 3(40%), 4(40%); 5(20%)
Server3: 1(60%), 2(80%), 3(100%); 4(0%), 5(20%)

As shown above, the non-frequent 1-item set from 3 servers is made up of session 4 and session 5. All result is synthesized and the integrated minimum support value of session 4 and session 5 for whole system is 20% and 33.3%. Thus global non-frequent 1-item set is made up of session 4 and 5.

Those sets are obtained and synthesize by data mining server, the global frequent 1-item set is calculated:
1(53.3%), 2(73.3%), 3(60%)

One round in the whole algorithm is shown in this example, and other rounds work similarly. In this example, after mining all frequent item sets in local e-commerce server, all results are transmitted to data mining server to synthesize all together. There is another way: the frequent item set got in every step is transmitted to data mining server to synthesize. The previous way needs powerful local compute ability; the second way needs high-speed networks. Powerful local compute is demanded in the previous method; high-speed network is demanded in the later method.

5. CONCLUSIONS

In this paper architecture of a data mining system based on web services is provided with optimized algorithm. Furthermore, Integration of data among different e-commerce servers is discussed. In the further research, an integrated solution should be the focus. A protocol for integration of e-commerce data should be built including data format, website structure, ID name of merchandise and so on.

6. REFERENCES

[1] D. W. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In Proc. 1996 Int. Conf. Parallel and Distributed Information Systems, Miami Beach, Florida, Dec. 1996, 31~44.

[2] Chen MS, Park JS, Yu PS Efficient data mining for path traversal patterns, IEEE Trans Knowledge and Data Eng. 10, 1998, 209¨C221.

[3] Chen L, Sycara K WebMate: a personal agent for browsing and searching, The Second International Conference on Autonomous Agents, ACM, 1998.

[4] Cohen E, Krishnamurthy B, Rexford J Efficient algorithms for predicting requests to web servers, The Eighteenth IEEE Annual Joint Conference on Computer and Communications Societies 1, 1999, 284¨C293

[5] Cooley R, Mobasher B, Srivastava J Grouping web page references into transactions for mining world wide web browsing patterns, Knowledge and Data Engineering Exchange Workshop, 1997, 2¨C9

[6] Cooley R, Mobasher B, Srivastava J Web mining: information and pattern discovery on the world wide web, Ninth IEEE International Conference on Tools with Artificial Intelligence, 1997, 558¨C567

[7] Zhong Luo, Xia Hongxia, Yuan Jingling. Study and Improvement on Hierarchical Algorithm of Association Rule, Data Mining and Knowledge Discovery: Theory, Tools, and Technology Ⅳ (Proceedings of SPIE). Beijing: SPIE, 2002, 88～93.

[8] Chen MS, Han J, Yu PS Data mining: an overview from a database perspective, IEEE Trans Knowledge and Data Eng 8(6), 1996, 866¨C883

[9] U.Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine, 1996, 37-53

Luo Zhong is a Full Professor .He graduated from Wuhan University in 1982; His research interests are in intelligent technology, software engineering, and image graphic.

Qiwei Tong is a Member of Intelligent Technology and System Lab, Postgraduate Student of School of Computer Science and Technology, Wuhan University of Technology. She graduated from Wuhan University of Technology in 2002 with specialty of civil engineering. Her research interests are in Artificial Intelligence, Intelligent Calculation, Expert System, E-commence and distributed parallel processing.

Building Robust J2EE Web Applications with Integration of Struts and JavaServer Faces

Yang Hao, Guo Qingping
School of Computer Science and Technology, Wuhan University of Technology
WuHan, HuBei 430063, China
Email: glrema@sohu.com; qpguo@public.wh.hb.cn **Tel.:** 027-86556339

ABSTRACT

It is likely that the front-end power of JavaServer Faces (JSF), the content-formatting strengths of Tiles, and the flexibility of the Struts controller tier can be all wrapped up in a J2EE Web application. This article specifies the integration of the features of all three. It also demonstrates the customization of the classes in the Struts-Faces integration library to make them work with Tiles and JSF, and explains the rationale behind doing this. By using Struts, Tiles, and JSF together, developers can ensure a robust, well-presented Web application that is easy to manage and reuse.

Keywords: Struts, JSF, Tiles, Integration, J2EE.

1. INTRODUCTION

The Struts framework has been around for quite some time and has become the actual standard that developers turn to when developing a J2EE Web application. The Tiles framework, which was part of Struts, established its position by offering developers the ability to assemble presentation pages using component parts. JSF, the newest kid on the block of Web application framework, provides mechanisms for validating user input and handling user events; most importantly, it is a protocol-independent way of rendering user interface components.

To run a J2EE Web application with these three technologies, you will need Struts 1.1, JavaServer Faces Reference Implementation (JSF-RI) Early Access Release 4.0, and Struts-Faces 0.4. Struts and Tiles come bundled in the Struts 1.1 release from the Jakarta project. The Struts-Faces integration library can also be obtained from the Jakarta project. The JSF-RI is part of the Web Services Developer Pack from Sun.

Although some of the functionalities in Struts and JSF overlap, they are complementary in other ways. The combination of these three technologies can provide an efficient way to develop a Web application, organize its presentation, and render custom user interface (UI) components independent of protocol.

And now, there are some news about integrating the three technologies. First the bad news: as of the writing of this article, the three technologies do not interoperate out of the box. And the good news: in this article, we specify the integration of Struts, Tiles, and JSF.

2. JAVASERVER FACES

2.1 Roles in Web Applications

JSF is a standard UI framework for Java Web applications. It is being led by Sun Microsystems under the Java Community Process (JCP). It promotes rapid Web application development by easily assembling UI components, plumbing them to the back end business logic components and wiring UI component generated events to server-side event handlers.

A JSF application typically runs on a web server and renders the UI back to the client. Its clean separation of UI components from their presentation allows the components to be rendered in different ways on different devices.

2.2 JavaServer Faces Architecture

The JSF architecture is comprised of six different layers.

UI Component Model: The UI component model primarily defines the functionality of the components. It consists of JavaBean components representing various modes of input, selection, and grouping capabilities. Every component has a type, an identifier, local values and some generic attributes.

Rendering Model: The rendering model defines the presentation aspects of a UI component. The same UI component can be rendered in different ways by creating multiple renderers.

Event Model: The JSF event model is similar to JavaBeans event design. It defines Listener and Event classes that a Web application can use to handle events generated by UI components. To be notified of an event, an application must provide an implementation of the Listener class and register it on the component that generates the event.

Validation Framework: The validation framework provides a mechanism by which validators can be registered with UI components. Validators are simple classes which can define data type validation, range validation, or required field validation. In its simplest form, for a class to act as a validator it has to implement the Validator interface.

Page Navigation Framework: The page navigation framework provides an easy declarative mechanism to specify the sequence of pages to be loaded without requiring any special code in the application. The navigation can be completely defined in the application resource file, a simple XML file, which contains navigation rules that define possible outcomes.

Internationalization Framework: This framework provides an easy mechanism for localizing static data, dynamic data, and messages in applications. Static data can be localized using the standard tag library internationalization tags and providing resource bundles and associating specific data in JSP pages with keys.

The JSF architecture is summarized in figure 1 below.

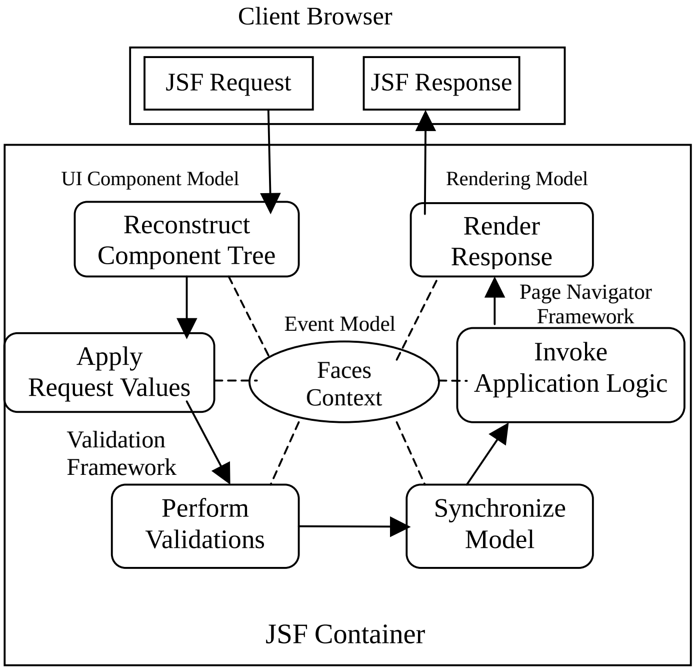


Figure 1 JSF Architecture

2.3 JavaServer Faces Page Lifecycle

At a high level, the lifecycle involves a request for a specific page submitted to a HTTP web server followed by a response generated for a specific device. The standard lifecycle for a typical JSF page is expressed in the following phases:

Reconstruct Component Tree: This phase is initiated when a page request is submitted to the server. During this phase the server creates a hierarchical tree of UI components which represent the elements constituting the page. The JSF engine parses this component tree and wires all the declared event handlers and validators to the specific components.

Apply Request Values: In this phase the JSF engine applies the values from the request object to the respective components in the component tree created in the previous phase. These values are stored locally in the UI components.

Perform Validations: After the request values have been applied, the JSF page enters the validation phase, in which all the registered validators are applied to relevant UI components.

Invoke Application Logic: After the model has been successfully validated, the page enters this phase. In this phase the JSF engine handles all application level events. When processing an application level event, a default action listener determines the outcome of the action and passes the outcome to the navigation handler.

Render Response: In this phase the JSF engine renders the UI components in the component tree persisted in the class *FacesContext*. This component tree is persisted so subsequent requests to this page can access it and it is readily available to the reconstruct component tree phase.

3. REASONS FOR THE INTEGRATION

As the JSP and the related specifications mature, new standards like JSF and the JSP Standard Tag Library (JSTL) which uses simple tags to encapsulate the core functionality common to many JSP applications are emerging. Following are some of the advantages to using the new technologies as an integrated whole:

1) Cleaner separation of behaviors and presentation. With the separation of tag, renderer, and component, the roles of page authors and application developers in the development cycle become better defined.
2) Changing the presentation for a component does not have an avalanche effect. Now we can easily just change the renderer. In the traditional MVC model, since this separation did not exist, any change in tags needed changes to the business logic as well.
3) Renderer independence. Or restated, protocol independence by reusing component logic for multiple presentation devices with multiple renderers. The ability to use different renderers eliminates the need to code the entire presentation tier for specific devices.
4) A standard for assembling and reusing custom components. JSF thinks beyond "forms and fields" and provides a rich component model for rendering custom GUI components. Using JSF we can customize the way each component looks and behaves in a page. Developers also gain the ability to create their own GUI components, which can easily be included in any JSP page with simple custom tags. Just like the Java front-end GUI components provided by AWT and Swing, we can have custom components for our web pages that use their own event handlers and have customizable appearances.

Struts is a framework that already possesses a large customer base. Many IT departments have recognized the value of this MVC framework and have been using it for quite a while. JSF doesn't possess the equivalent of Struts's powerful controller architecture, as well as its standardized *ActionForm* and *Action* classes (with their declarative capabilities). When we integrate Tiles into the mix, we have the ability to reuse and change corporate layouts in a seamless manner.

The challenges of migrating JSF-enabled Struts applications are two-fold. First, Struts tags are not JSF-compliant. In other words, they do not extend the *UIComponentTag* as mandated by the JSF specification, therefore, JSF cannot interpret and associate UI component and renderer with them. Second, there is no link between the *FacesServlet* and Struts *RequestProcessor*. In a Struts application, *RequestProcessor* manages the show with the callback methods into *ActionForm* and *Action* classes. Unless the *RequestProcessor* gets invoked, the callback methods in Struts *ActionForm* and *Action* classes do not get a chance to invoke the business logic.

4. INTEGRATION OF STRUTS AND JSF

4.1 Components of Struts-Faces

Struts-Faces is an early access release of the Struts JSF integration library which makes it easy to migrate the existing Struts applications to JSF. Struts-Faces also strives for a clean integration with JSF so that JSF can be used on the front end while the back end will still have the familiar Struts components.

The following are the major components of Struts-Faces:

1) The *FacesRequestProcessor* class. This class subclasses the regular Struts *RequestProcessor* and handles the faces requests. Non-faces requests are delegated to its parent, *RequestProcessor*.
2) The *ActionListenerImpl* class. This class is used instead of the default *ActionListener* implementation which handles action events such as submitting a form or clicking on a link.
3) The *FormComponent* class. This class extends from the JSF form component but is invoked within the Struts life cycle.
4) The *LifeCycleListener* class. This class is an implementation of the *ServletContextListener*, which is used to register the appropriate *RequestProcessor* during initialization.

4.2 Migrating Struts to JSF

In order to integrate the Struts Web application with JSF, we should modify the JSP pages to use the JSF and Struts-Faces tags instead of the Struts tags, and for each JSP page that uses JSF tags, modify the Struts configuration file to include the particular prefix in the global forwards and the local forwards in the action mappings pointing to that JSP page.

JSF form has its own model class and gets the values in update model values phase. In Struts however, the form action points to the action mapping in the Struts configuration file. In other words, the *ActionForm* itself is the model for the HTML form in Struts-Faces and its action indirectly points to this model. As the Struts-Faces uses the form action itself as the form name, the familiar *.do* is missing in the action attribute.

We should notice that no *ActionListener* is explicitly associated with the submit request. This is because the *ActionListener* is already provided in Struts-Faces and always forwards the faces requests with action events to the *FacesRequestProcessor*, from which the requests are dispatched to appropriate *Action* based on the Struts configuration file.

5. RICHER PRESENTATION WITH TILES

The Struts-Faces library provides an efficient bridge between Struts and JSF, making rich presentation layers a reality in J2EE Web applications. We can make the presentation layers even richer by adding Tiles to the combination, so that we not only get the benefit of the Struts and JSF combination, but we can also efficiently reuse the various JSP pages because they will be made up of component parts or tiles that can be added or removed as required.

Since JSF is still in the early stages, the Struts-Faces integration library is being developed iteratively to accommodate the various features of JSF and does not yet support Tiles. So we will encounter roadblocks when using the Struts-Faces integration library in conjunction with Tiles. For each of these problems, we can put forward a solution by modifying the Struts-Faces classes.

The first problem we will see is that the response has already been committed when we try to access a JSF form. This problem lies in the class *ViewHandlerImpl* which is a JSF-RI class implementing the *ViewHandler*. *ViewHandler* is the class that forwards the request to the next page. When forwarding the request, it does not check the status of the response before forwarding it. This happens only when we use Tiles, because Tiles internally includes the JSP pages in the response and JSF-RI commits the response after the first forward and then tries to forward again to the next Tiles including JSP. To fix this problem, we will have to create a custom *ViewHandler* implementation that will check the status of the response to determine whether it has been committed. If the response has not been committed, then the request is forwarded to the next page; otherwise, the request is included and the appropriate JSP is displayed.

The committed response problem has already been fixed. But if we access any Tiles-specific link or a URL that would render a JSF response, we will get an error of 404 (Resource Not Found). *FacesRequestProcessor* does not have the capability of processing Tiles requests because it extends *RequestProcessor* instead of *TilesRequestProcessor*. And therefore, ends in an error. To debug this problem, we can create a new request processor class which subclasses *TilesRequestProcessor* instead of subclassing the regular *RequestProcessor*. This new request processor can handle Tiles requests.

Thanks to the new request processor above, at this point we can navigate and see all the JSP pages. However, as soon as we submit a regular form, we get the same form in return, and there are no validation errors. If we view the HTML source from the browser, we notice that the form action is pointing to a JSP page instead of a suffix of *.do*. That’s the problem. It is the default behavior of Struts-Faces to have the same JSP name as the form action. Because there is no action mapping for suffix of *.do*, Struts cannot find the action form. The solution to this problem is to use class *RequestUtils* in Struts framework. This class is intelligent enough to resolve the path or suffix of *.do* mapping into appropriate action mapping.

As the form action in JSP page has changed, we have to make a modification to the action method of form renderer which renders the Struts-Faces form in HTML format. We do this by subclassing *FormRenderer* and overriding the method to change the form action written out to the HTML. As we already know, when the component and renderers change, the tag has to change, too. We can create new tags by subclassing *FormTag* in Struts-Faces.

6. CONCLUSION

We have covered the customization of Struts classes to enable a tightly integrated working relationship with both JavaServer Faces and the Tiles framework and have provided an overview of the component technologies used in Web applications. More importantly, we've offered a cogent roadmap to combine Struts, Tiles, and JavaServer Faces into a powerful, flexible mechanism for building J2EE Web applications.

Because of the independence of JavaServer Faces and Struts, however, we have to spend some time to integrate them instead of using a single powerful framework including the function of both JavaServer Faces and Struts. We believe that this multi-functional framework will be released in the near future with the dedication of some open source organizations. At that time, we can build presentation and logic layers of a J2EE Web application in one framework instead of using the integration.

7. REFERENCES

[1] Craig McClanahan. JavaServer Faces Specification. Sun Microsystems Inc, 2003.

[2] Chuck Cavaness. Programming Jakarta Struts, 2nd Edition. Sebastopol, CA: O'Reilly & Associates Inc, 2004.

[3] Hans Bergsten. JavaServer Faces. Sebastopol, CA: O'Reilly & Associates Inc, 2004.

[4] Craig McClanahan. About Struts and JavaServer Faces. http://jakarta.apache.org/struts/proposals/struts-faces.html, 2002.

[5] Ed Burns. An Introduction to JavaServer Faces. http://java.sun.com/j2ee/javaserverfaces/jsfintro.html, 2003.

[6] He Chengwan, Yu Qiuhui. Study on MVC Model 2 and Struts Framework. Computer Engineering, 2002, 6: 274-275.

Yang Hao is a master degree candidate of School of Computer Science and Technology, Wuhan University of Technology. He graduated from Chongqing Technology and Business University and got bachelor degree in 2000. His research interests are in Electronic Commerce and Distributed Databases.

An XML Web Service Application Architecture Based on Microsoft BizTalk Server

Zhou Ying, Liu Quan
School of Information Engineering, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: going-zy@163.com **Tel:** +86 13971591850

ABSTRACT

Using Microsoft BizTalk Server, We can develop a kind of XML Web Service application architecture, which is robust, stable, asynchronous and transactional. This architecture takes full advantage of protocol standard feature of web service, combined with the message integration and automatic e-business process of the BizTalk, it acts as a long run-time, couple and distributed e-business process, so that it meets the requirement of the application which needs good stabilization, compatibility and fault tolerance. In this paper, firstly we give the detailed introduction on the concept concerned with XML Web Service and some problems existing in XML Web Service application. Then we describe the concept and function of BizTalk architecture, and at last put forward a solution on how to develop the XML Web Service using BizTalk to settle the application problems of XML Web Service.

Keywords: BizTalk, XML Web Service, SOAP, WSDL, UDDI, XLANG, Orchestration.

1. XML WEB SERVICE AND THE APPLICATION PROBLEMS

1.1 Summary of XML Web Service

An XML Web Service is a programmable application logic, which combines the best aspects of the World Wide Web and component-based development. As viewed from Internet, XML Web Service is accessible using standard Internet protocols; As viewed from components, XML Web Services represent a black-box functionality that can be reused without concerning about how the service is implemented.[1] But unlike previous component technologies, XML Web Services are not accessed through some object model-specific protocols, such as the distributed Component Object Model(DCOM), Remote Method Invocation(RMI), or Internet Inter-ORB Protocol(IIOP).

Instead, XML Web Services are accessed through a lot of existing Web protocols and data formats, such as HTTP, XML, SOAP. Furthermore, an XML Web Service interface is defined strictly according to the messages the Web Service accepts and generates. If the users can create and consume the messages defined for the XML Web Service interface, then they can implement an XML Web Service on any platform in any programming language.

XML is the obvious choice to represent data related to XML Web services in a standard way. That is, XML is good at representing the format of the data transmitted to and from the Web service, and the various Web service-related specifications all use XML for data representation.

XML Web services require a messaging protocol that can invoke the Web services and exchange data with them. SOAP is a lightweight, XML-based protocol for exchanging information in a distributed environment.[4] SOAP is also a network protocol, with no programming model at the bottom. Because SOAP does not appoint the technology used to implement the client or server applications, it requires no application programming interface (API) or object model. So, SOAP provides XML Web services as a solution for application-to-application communication.

1.2 Challenges in XML Web Service application

XML Web Services are undoubtedly a key technology, if we want to provide great software, any time, any place, and on any device. XML Web services enable a code reuse pattern for delivering highly distributed applications, so services are available to an application without being physically connected. However, building applications in a distributed, loosely coupled environment introduces some significant challenges:

1) Interactions among XML Web Services: XML Web Services provide convenient access to both local and remote e-business logic. However, if an application is composed of many XML Web Services, then there will be a management challenge: How the interactions between the XML Web Services are managed in a flexible way, so that new XML Web Services can be added to an application easily.

2) Transaction management and exception handling: XML Web Services provide access to remote e-business logic, so there are problems that how the transactions can be managed across XML Web Services, and how the exception can be processed if different XML Web Services are required to be called.

3) Concurrency: Application should be able to call XML Web Services with no dependencies with one another. Our aim is to achieve this in a parallel manner without complex threads.

4) Interactions with non-XML applications: Nowadays application development must make interactions not only with XML Web Services but also non-XML Web Services. We should find a similar manner to manage the interactions across XML Web Services and other systems.

In many loosely coupled XML Web Services, we can find lots of requirements for these technologies above. In many cases, it is not possible for Web Services to use a kind of models for simple synchronous applications as COM does. If the application developer wants to develop highly extensible and available Web Services, BizTalk Orchestration provides some special characteristics. In this article, we'll bring forward a solution with working stream integration based on Microsoft BizTalk Server, which is the kernel of the XML Web Services application architecture.

2. XML WEB SERVICE BASED ON BIZTALK SERVER

2.1 summary of BizTalk Server

Generally speaking, it is difficult to realize the application integrations in an enterprise. It will be more difficulty to integrate the applications across the enterprises. BizTalk is a system for information exchanging and applications integrations that realizes enterprise application integrations and Business-to-Business integrations. BizTalk Server provides significant support for the development of applications that are widely distributed in space (EAI and B2B) and time (long-running business processes).[2]

A BizTalk Server Orchestration is a process created in the BizTalk Orchestration Designer, serialized in XML, and executed under the control of COM+ services (called XLANG Scheduler). BizTalk Orchestration provides a long-running, loosely coupled business process to enable the application designer to create robust business processes.

BizTalk Server mainly provides two core functions:

- Message-level integration, from the enterprise (enterprise application integration, or EAI) to the Internet (business-to-business, or B2B), through BizTalk Server Messaging.
- Business process automation using BizTalk Server Orchestration services, which can be able to implement a long-running, loosely coupled business processes.

Because of the support to Internet standard technologies such as XML, HTTP, HTTPS, SMTP, SSL, S/MIME and x509v3 certificate, BizTalk server is a powerful tool in application integrations, and BizTalk Orchestration provides significant benefits to the application designer building highly distributed, long-running processes. But Microsoft BizTalk Server also has the limitation that the services connect with windows operation system closely, so the development will be complex if the application integrated is based on the other operation system platform.

2.2 BizTalk Orchestration and XML Web Services

We introduce a kind of working-stream integration tool as the kernel of Web Services application architecture. This integration tool communicates with XML Web Service directly, and controls the running of all the application. The implementation principle is that the calling XML Web Service from clients can be divided into calling proxy from clients and calling XML Web Service from proxy. That is to say, clients can asynchronously issue a request for service (from a proxy), and the proxy ensures that the request is eventually submitted to the Web service for processing, including exception handling and transactional integrality. Of course, the system must also be able to eventually return a response to the proxy, and then give back to clients.

BizTalk Orchestration, a key part of BizTalk Server, applies directly to this problem. It can connect every XML Web Service all together so that to communicate with every Service, just like the proxy mentioned above. BizTalk Orchestration was built to solve the problems associated with managing long-running, loosely coupled business processes that are distributed across organizational boundaries. BizTalk Orchestration provides services, such as timed and long-running transactions, exception handling, and transaction compensation, to allow the application designer to design robust business processes that are capable of recovering from failure. These are the facilities required to build robust XML Web services.

The following sections discuss two ways of integrating every XML Web Service: Calling XML Web Services from an Orchestration Schedule, and implementing an XML Web Service Using Orchestration.

2.2.1 Calling XML Web Services from an Orchestration Schedule

Calling XML Web Services from an Orchestration Schedule discusses how XML Web services can be combined with BizTalk Orchestration.

As a proxy and integration tool, BizTalk orchestrations should be enabled to call XML Web services to implement specific actions within those orchestrations. XML Web services, however, are invoked by sending SOAP-formatted requests to the Web service and receiving SOAP-formatted responses back to the client.[3] Because BizTalk Orchestration Designer cannot call the Web service, the easiest way to implement this from a BizTalk Orchestration schedule is by calling a COM or .NET component, which invokes the Web service using SOAP. In effect, a COM object that BizTalk Orchestration can call acts as a client proxy object for the XML Web service, as shown in figure1.

The COM component proxy can be created using either SOAP Toolkit 2.0, or Visual Studio .NET. Both methods require the WSDL file that provides a description for the methods exposed by the Web service. If the Web service was created using the SOAP Toolkit, the WSDL file will have been generated using the SOAP Toolkit WSDL Generator utility. If the Web service was created using Visual Studio .NET, then the characteristics of the Web service can be showed in the service's .asmx file. Both the WSDL and .asmx file describe the details about the format of the SOAP messages required to invoke the various methods for the Web service.

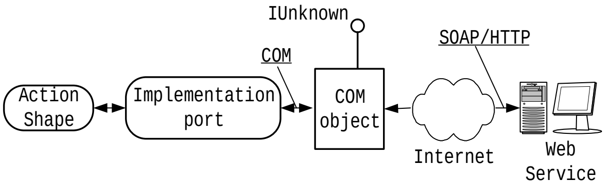


Figure1 The COM object serves as a proxy for the Web service

This particular proxy object combines several methods from the Web service and creates a more complex business process. When this object is compiled, it generates a COM component linked into the orchestration.

2.2.2 Implementing an XML Web Service using Orchestration

As a tool for integrating every XML Web Service, BizTalk orchestration should be exposed as an XML Web service used to be integrated by other orchestrations, as shown in figure2. It can be exposed in three ways: Programmatically exposed, using SOAP Toolkit, and Using Visual Studio .NET.

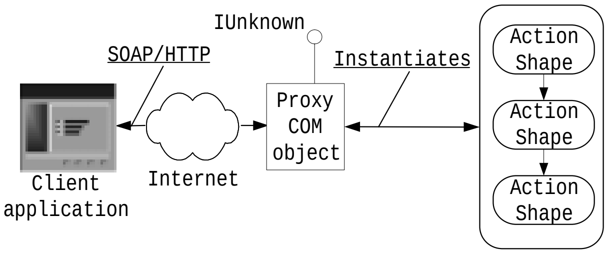


Figure2 The COM object exposes the BizTalk Orchestration as an XML Web service

By default, when BizTalk Server is installed, it creates the single COM+ package to run all schedules on that server, and adds an XLANG tab to the COM+ Applications Properties. An XLANG schedule is a process created in the BizTalk Orchestration, serialized in XML, and executed under the control of COM+ services (called XLANG Scheduler). It is possible to run an entire orchestration as a COM component under the control of COM+ services. Implemented with COM+ components, the orchestration engine provides a powerful mechanism for business process automation.

When a new COM+ application is created, the XLANG tab can be used to specify that the COM+ application is also a host for XLANG schedule instance. A client application can use COM to invoke these XLANG schedule instances by specifying the path to the XML file and the orchestration port. Also, once the Web service has been created, it can be invoked by the client application either using the SOAP Client COM object supplied by SOAP Toolkit 2.0, or using Visual Studio .NET and the common language runtime.

We should implement the same code that can instantiates an instance of the schedule firstly when using SOAP Toolkit and using Visual Studio .NET, and package the interfaces and functions of this code, then package the code again to be a Web Service at last.

3. CONCLUSIONS

XML Web Service is an excellent solution, but it also has some problems. If the application architecture is composed of many XML Web Services, then the problems about the service management, exception handling and transaction must be solved. It is complicated for every Web Service to handle exception and transactions by itself, and sometimes it is incapable to. There have been a lot of solutions; however, each of them is more or less deficient in requirement of the system running safety and good stabilization.[5] As a platform of information exchange and application integration, BizTalk Service can make schedules of the XML Web Service easily and ensure the application running with security and reliability, by using exception handling and transaction. Therefore, the XML Web Service based on BizTalk is robust, extensible and convenient in dispose, providing better solution for business applications.

4. REFERENCES

[1] Brian E.Travis, XML and SOAP Programming for BizTalk Servers [M], Seattle: Microsoft Press, 2003

[2] Peishu Li, BizTalk Server Developer’s Guide [M], Seattle: Publishing of Tsinghua University, 2003

[3] Carlos C. Tapang, Web Service Description Language (WSDL) Explained [M], Seattle: Microsoft, MSDN, 2002

[4] Scott Seely, An XML Overview Towards Understanding SOAP [M], Seattle: Microsoft, MSDN, 2002

[5] Roger Wolter, XML Web Service Basics [M], Seattle: Microsoft, MSDN, 2002

A Model of the 3D Virtual Shopping that Has the Intelligent and Cooperative Purchasing Functionalities *

Zhao Yiming
Department of Computer Science & Technology NingBo Uinversity
Ningbo, Zhejiang 315211, China
Email: zhaoym@mail.nbptt.zj.cn **Tel.:** +86(0574)87604357

ABSTRACT

This paper represents a model of the 3D virtual shopping realized by technologies such as multi-agent, VRML, JAVA2, XML and the computer network. In this model, there is a symbol with mood for every customer, which can communicate freely with symbol of other customers, commodities in the shopping mall and the intelligent agent. In the meantime, a virtual purchasing guider is designed to communicate with different symbols of customers in different manners. This virtual guider is able to patrol in this virtual 3D environment along the path planned by itself. Customers may have dynamic communication with other purchasers through the virtual guider or operate interactively on all kinds of commodities. This system simulates the whole purchasing process including window-shopping, choosing and paying for the items. Such comprehensive simulation is a new – brand E-business system that makes the customer feels like they are personally on the scene.

Keywords: E-business, 3D Virtual Shopping, intelligent, Cooperative

1. INTRODUCTION

A 3D virtual shopping mall based on Web is introduced in this paper [1]. It is a virtual shopping environment based on Internet with the B2C mode applied so that the customers can cruise in and interact with a 3D virtual environment. All processes of viewing, selection and payment are virtualized and managed with digital technology. The advantage of E-commerce is the unlimited time and space of purchase, i.e., the deal can be carried out at any time in any place. However, it also faces some problems.[1].Based on paper [1], paper [2] studies the virtual shopping mall with capacity of sense and presents a prototype system called EasyMall. This system provides customers with the ability of viewing and manipulating the commodities using 3D technology so that they can view the items under their favorable environment. At the same time, the agent technology is applied to greatly enrich the customer's shopping process by simulating the body language of the customer's symbol such as their pose and the controller's response in the virtual environment. Furthermore, customers will be more ready to understand and accept the virtual environment.

This paper represents a model of the 3D virtual shopping realized by technologies such as multi-agent, VRML, JAVA2, XML and the computer network. . In this environment, there is a symbol with mood for every customer, which can communicate freely with symbols of other customers, commodities in the shopping mall and the intelligent agent. In the meantime, a 3D virtual purchasing guide with emotion is designed to communicate with different symbols of customers in different manners. This virtual guide is able to patrol in this virtual 3D environment along the path planned by itself. Customers may have dynamic communication with other purchasers through the virtual guide or operate interactively on all kinds of commodities. This system simulates the whole purchasing process including viewing, choosing and paying for the items. Such comprehensive simulation is a new generation of E-Commerce system that makes the customer feels like they are personally on the scene.

2. SYSTEM MODLE

2.1 Basic Idea

This system is a distributed multi-user virtual environment made up of multiple servers with the client/server structure. The VRML objective files, the CGI programs and the VRML browser embedded with Java programs are integrated at the client. The 3dmax technology is applied to present the scene of people shopping in the mall and to design the user symbol and the virtual shopping guide, all of which are exported as VRML files. All VRML models are divided into four kinds: the scene, the product, the client symbol and the virtual guide. The scene, the client symbol and the virtual guide are stored in the table list at the client end.

The product model is the information about goods exported from the server like comments about the items and the VRML model of products. The client end is responsible for simulating the roving in the virtual environment and the pre-order of the client as well as the submission of the bill of order. The sever program is responsible for updating the data about the models and creating dynamic web page at the client, accepting user's bill of order from the client and interacting with the database. Both ends should communicate with the third party to achieve authentication to guarantee the integrality, validity, confidentiality and reliability of the deal.

2.2 System Structure of 3D Web Virtual Shopping Mall

The system includes user interface layer, business logic layer (function layer) and data-visit layer. It is divided into two parts: the server end and the client end with a protective firewall between them. The client end and the server end can be linked together by Internet at any place. The host sever includes the file server, the database server, the WWW server, the producer for dynamic web page and 3D object as well as data server for the user symbols and the agents . It also supplies a functional CGI for a general course. The whole system is given in Fig.1.

* A Project Supported by Scientific Research Fund of Zhejiang Provincial Education Department

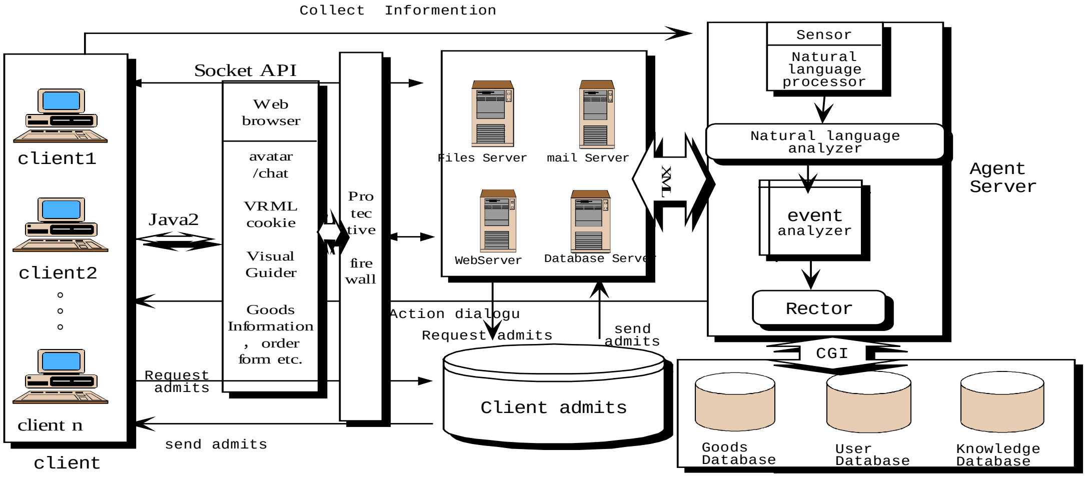


Figure 1 The Structure of the System

2.3 Key Technology

Design the Virtual Guider

We apply role cartoon technology to design the virtual guider: first, the role model is set up as the motive system; then the system will move like alive. The role cartoon is designed with Character Studio 3.X role cartoon system that provides methods of setting up and modifying roles including tools to create group cartoon. Character Studio provides various unique tool groups for motion-catching, the free cartoon style as well as the step-trace cartoon. In the virtual shopping environment, we use virtual bones to control the points on the limbs and these bones make up the skeleton of the role. We make the role to put up various poses by adjusting the skeleton. It should be noted that the points on some places often have problems of dragging the skin and the muscle, that is, they may be dragged by other parts of the body or stay at the same place uncontrolled by the skeleton. This is caused by the unreasonable distribution of the weight of the skeleton.

We get rid of the phenomenon by adjusting and controlling the weight of points on the model. In Fig. 2, the deformable nodes of the joint place between the left arm and armpit are showed as red points; the green nodes are rigid points, and the blue points near the chest should not be linked to the arms. The virtual guide designed by us is given in Fig. 3

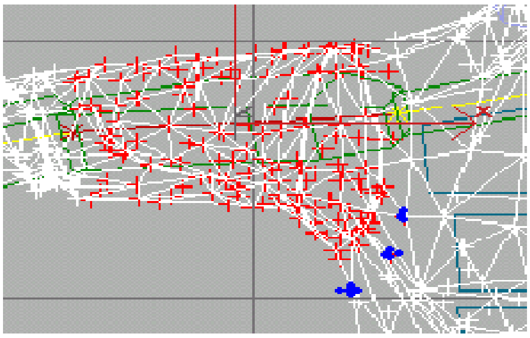


Figure 2 The animated points on the limbs

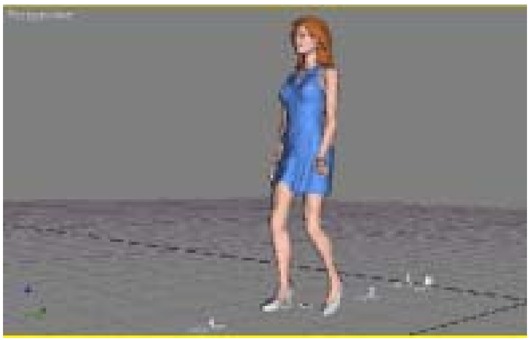


Figure 3 Virtual guide in move

Path Planning [3]

To combine the path of the virtual guide and the inquiry of the goods, we introduce the 3Dmax model into VRML, and then implement the algorithm optimizing the whole path by programming in JAVA and XML. This algorithm presents the virtual environment using the girding method. The data structure of the octo-tree is used to represent the discrete environment. Every node of the tree represents a cell and includes a table of eight pointers pointing to the other nodes. The null pointer represents a free cell. The 3D environment can be divided into the borderline, the free zone and the blocking zone according to their accessibility.

Design the Multi-agent Model

This is a multi-agent system including five kinds of different agent models to process the text in natural language and the emotion involved in the interaction between the customers and the virtual guide (Fig. 4). Those agents are the user interface agent [4], the interactive agent for language processing [5,6,7], the emotional agent, the intelligent search agent and the database agent. The customer requests for some items using text in natural language. The customer interface agent accepts this request, transfers it to the event analyzer, and sends back the information after inquiring the database. At the same time, the emotional agent will transfer the movement of the customer symbol and the virtual guide to the thought module. Meanwhile the thought module also accepts the semantics of the natural language to define the body behavior, i.e., introduces the goods and leads the customer to the exhibition zone.

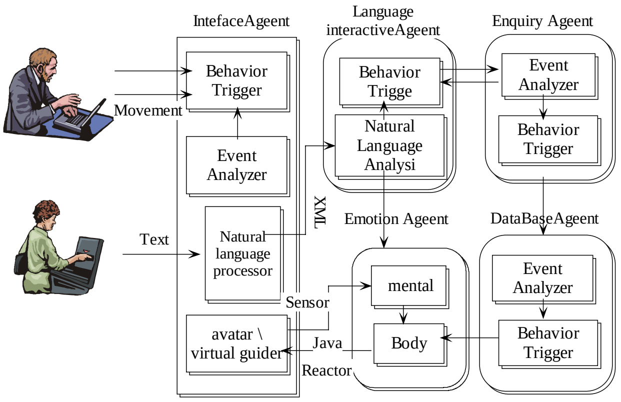


Figure 4 The structure of the multi agent model system

3. DESIGN AND INNER STRUCTURE OF CLIENT SYMBOL AND VIRTUAL GUIDE [8,9,10]

In our feel-like-true shopping environment, the emotion and the movement of the user symbol and the virtual guide are realized by a series of emotional Agents. Every Agent makes different responses to the outside changes by the sensors and the reactors. Here is our introduction of the Agent environment (Fig. 5)

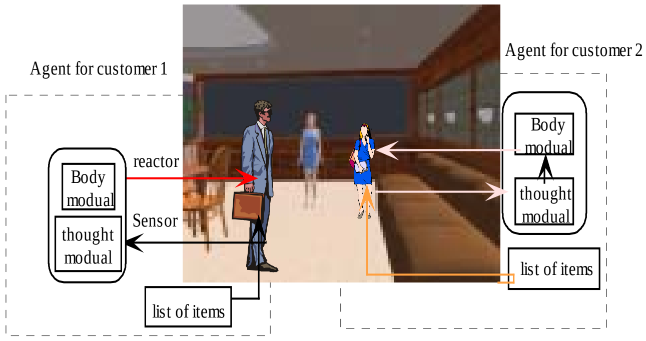
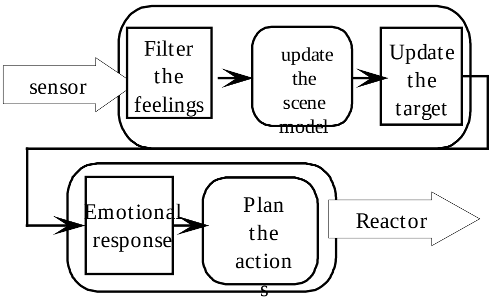


Figure 5 The inner structure of visual person

The sensor is used to collect information from outside and transform it into the cognition, which in turn is submitted to the thought module. There are four kinds of sensors in the system. The behavioral sensor is to "feel" all behaviors in the current environment including that of itself. The scope sensor is generally defined on an object in the 3D world and will be triggered when an agent enters or leaves the scope of this object. The new-entity sensor is attached to an agent so that when other agents enter or leave the environment in which this host agent exists it will be triggered and send the related news to the host agent. The new-environment sensor is also on an agent and when this agent enters into a new environment it will send a message to this agent. The reactor is for the allowed actions of the body on the scene by the agent. Such actions include walking, holding objects, placing objects, using objects and talking, etc.. The reactor contains the information needed when taking these actions. Every action is divided into three stages. First its feasibility is examined by checking to see whether all the preconditions are satisfied. After the check the action is implemented mainly by changing the body. When the action finishes it goes into the third stage and its influence on the environment is implemented.



Figuer 6 . Thought construt and flow chart

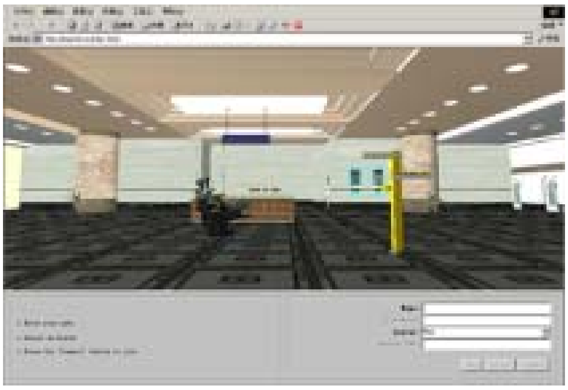


Figure 7 The entrance of the virtual emporium

The body module represents the person agent in the shopping environment. It not only displays the appearance of the agent, but also reflects its physical state including its stature, weight and position in the environment, etc. In order to reduce the cost of computation, the body will be displayed in 2D form in this 3D environment except that the virtual guide will be modeled in 3D form. The agent's action is implemented by continuous images.

The thought module is the key part. It defines most behaviors of the agent including its characteristics and the planned target. The thought module makes choice by receiving message from the sensor. It plans the coming movement of the agent and stores all information that is required during the planning process. Fig. 6 is the map for decomposing the thought module.

We can learn from the map that the thought module first stores emotional information previously transferred from the sensor including the current scene model, some currently executable movement, the target and the emotional state, etc. When the information is transferred from the outside to the thought module, this module will check the essentiality of the feelings by filtering and then decide whether to accept it. When the thought module receives the apperception that the scene model has been updated, the scene model within it will be changed accordingly. The target of the agent will be changed consequent to the changes in the emotional state, e.g., the agent may stop communicating with the other customers if its emotion turns sad. And the target of the agent may change too if some items on the scene are changed. For example, the target of the agent is to get an item, so the sub-target is to get closer in order to fetch it. But if some body has taken it away, this target will be deleted because it is not realizable any more.

All the changes of these targets are implemented by the part of updating the target. The part of emotional response will change the emotional state according to the changes in the scene and the current target. As long as the target changes, the part of planning the actions should re-design the actions to achieve the target. The design algorithm comes from the comprehensive analysis of the current scene model, the emotional state as well as what actions the agent knows can be taken. After the design, it should decide on the actions to take and the corresponding reactor will begin to implement them.

4. CONCLUSION AND EXPECTATION

In this paper, the model and construction technology of a virtual shopping mall system with virtual guider is introduced Fig 7). This system offers different functions to client, businessman and system administrator respectively. The user symbol with 3D multi user environment in the system roves in the virtual scene. A multi agent system model integrated with natural language and emotion construction is put forward. The technology that multi person shop simultaneity at the virtual shopping mall is realized including walk, greeting each other, chat and the expression of corresponding body language.

As a new generation of E-business system, the 3D user symbol and emotion interaction still need to be improved; there are still a large amount study works to 3D expression and drawing of complicated shopping environment to be done[12,13,14,15]. Especially the foundation of the on-line agent-driving, the realization of open, automatic and extendible platform of E-business, the negotiation to price and other goods factors between client agent and businessman agent, the shopping guiding of the virtual shopping mall to the shopping process of the customer still need to be researched.

5. REFERENCES

[1] Zhao Yiming, Wu Shuwen, Pan Zhigeng, the design and construction of 3D web virtual shopping mall, Transaction of System Simulation: pp980-982 (7) 2003

[2] Pan Zhigeng, Xu Bin, Chen Tian, the design and construction of interactive 3D virtual shopping mall EasyMall system, Transaction of System Simulation, (Delivered)

[3] He Huaiqing, Liu Haohan, a path planning arithmetic of virtual environment and virtual person, Computer Engineering, 31-33 (12) 2002

[4] Weidong Geng, Wolfgang Strauss, Monika Fleischmann, Vladimir Elistratov, Thomas Kulessa, Marina Kolesnik: Perceptual User Interface in Virtual Shopping Environment. VRAI'2002, pp.220-226. Hangzhou, P.R.China, April 9-12 2002

[5] C.Greehalgh,and S.Benford, Virtual Reality Teleconferencing:Implementation and Experience, European Conference on SCW, Stockholm North-Holland, Sep.1995

[6] Worlds Chat Homepage: http://www.worlds.net/

[7]. Kamyab, F. Guerin, P. Goulev and E. Mamdani (2001) Designing Agents for a Virtual Marketplace. Workshop on Information Agents in E-commerce; Agents and Cognition, AISB Convention, York

[8].R.H.Guttman, A.G. Moukas, and P.Maes, Agent- mediated Electronic Commerce: A Survey. http://ecommerce.media. mit .edu /papers/ker98.pdf(30.09.1999).

[9] J.Cassell and H.Vilhjlmsson.Fully embodied converstional autonomous. In Autonomous agents and Multi-Agent Systems, volume 2(1),1999

[10] A. Chavez and P. Maes. Kasbah: An Agent Marketplace for Buying and Selling Goods. In Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology, London, UK, April 1996.

[11] Paiva, A., Prada R. & Machado, I."Heroes, villains, magicians..: Dramatis Personae in a virtual story creation environment" In "Intelligent User Interfaces" ACM Press, 2001

[12] J.C.Oliveira, X.Shen and N.D.Georganas , "Collaborative Virtual Environment for Industrial Training and e-Commerce", Proc. Workshop on Application of Virtual Reality Technologies for Future Telecommunication Systems, IEEE Globecom'2000 Conference, Nov.-Dec.2000, San Francisco.

Zhao Yiming is a Associate Professor, Dpet of Computer Sci &Eng,NingBo University ,Zhejiang . He graduated from XIDIAN University in 1982; He was a visiting scholar of State key Laboratory of CAD & CG, Zhejiang University (2000~2001), He has published eight books, over 20 Journal papers, His research interests are in distributed Virtual Reality, Computing Intelligence, network security and e-commence.

A Server Electronic Wallet Architecture Supported Multi-payment Protocols and Instruments

Bo Meng[1], Zhang Huanguo [2], Xiong Qianxing[3]
[1]College of Computer Science, Wuhan University Wuhan, Hubei 430072 P. R. China
Email: mengbo@263.net.cn **Tel:** 027-87885922
[2]College of Computer Science, Wuhan University Wuhan, Hubei 430072 P. R. China
Email: liss@whu.edu.cn **Tel:** 87885922-2494
[3]College of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei 430063 P. R. China
Email: qxxi@public.wh.hb.cn **Tel:** 027-86551711

ABSTRACT

With the development of electronic commerce, more and more people and enterprises do electronic business on the Internet. So the electronic payment tools, which should be secure and extensible, are needed very much. Electronic wallet is one of the most important electronic payment tools. An electronic wallet is a collection of confidential data of a personal nature or relating to a business role carried out by an individual, managed to conventions agreed with the owner, to facilitate completion of electronic transactions. Electronic wallets can be classified into two classes with whether they are based on fat electronic wallet model or server electronic wallet model. At the present time several electronic wallets, some are fat electronic wallet, others are server electronic wallet, have been developed. But they don't support the multi-payment protocols and instruments together, which is not convenient to the people and enterprise. In this paper first we research on the fat electronic wallet model and the server electronic wallet model and analyses the difference between them. Secondly we present a new server electronic wallet architecture, which supported multi-payment protocols and instruments, based on sever electronic wallet model and is secure and extensible. In the end we give its prototype implementation based on the architecture presented by us with the extensible markup language, web service and security technologies.

Keywords: server electronic wallet, fat electronic wallet, multi-payment protocols, multi-payment instruments.

1. INTRODUCTION

With the development of electronic commerce, more and more people and enterprises do electronic business on the Internet. So the electronic payment tools, which should be secure and extensible, are needed very much. Electronic wallet is one of the most important electronic payment tools. An electronic wallet [1] is a collection of confidential data of a personal nature or relating to a business role carried out by an individual, managed to conventions agreed with the owner, to facilitate completion of electronic transactions.

At the present time many electronic wallets, some are fat electronic wallet, such as SET electronic wallet of bank-of –china [2], others are server electronic wallet, such as 51 QB [3], Shenzhen financial organization [4], NetPay server electronic wallet [5], have been developed. But they don't support the multi-payment protocols and instruments together. When the people and enterprise do electronic business with different merchant and pay the bill with the electronic wallet that only supported one payment protocol and instrument, the people and enterprise must have to install several electronic wallets to pay different bill. This is not convenient to people and enterprise.

In order to solve this problem we present a new server electronic wallet architecture in this paper, which supported multi-payment protocols and instruments, based on sever electronic wallet model and is be secure and extensible. In addition we research on the fat electronic wallet model and the server electronic wallet model and analyses the difference between them. In the end we give its prototype implementation based on the architecture presented by us with the extensible markup language, web service and security technologies.

2. ELECTRONIC WALLET MODEL

Electronic wallets can be classified into two classes with whether they are based the fat electronic wallet model or the server electronic wallet model. Electronic wallet based on fat electronic wallet model was called fat electronic wallet, such as SET electronic wallet of bank-of –china. Electronic wallet based on the server electronic wallet model called server electronic wallet, such as 51QB, Shenzhen financial organization, NetPay server electronic wallet.

Fat electronic wallets model has the general functions: Access/modification of electronic wallet information, Authentication to initiate a transaction, Reliable conveying of authentication status to merchants and other service providers, Making available required personal data to merchants to complete transactions, provision of payment/card details and making payment, Active management by electronic wallet manager to agreed contractual standards.

But it has a few limitations as following:

1. At the present time the fat electronic wallet is about 3~7M. Its installation, downloading, maintenance, updating are done by the user of the fat electronic wallet user, which need time and fee. So it is not good for the user of fat electronic wallet.
2. The fat electronic wallet has not the platform compatibility.
3. It is not safe owing to the key data of the user stored in the local computer.

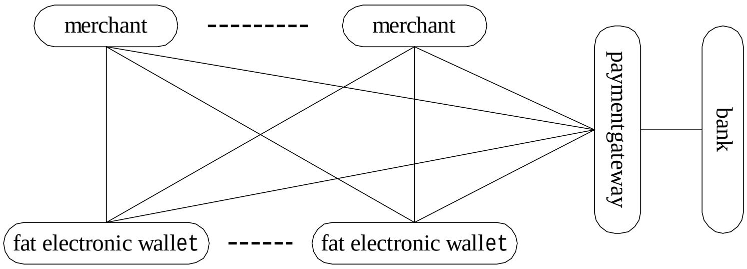


Figure 1 The fat electronic wallet model

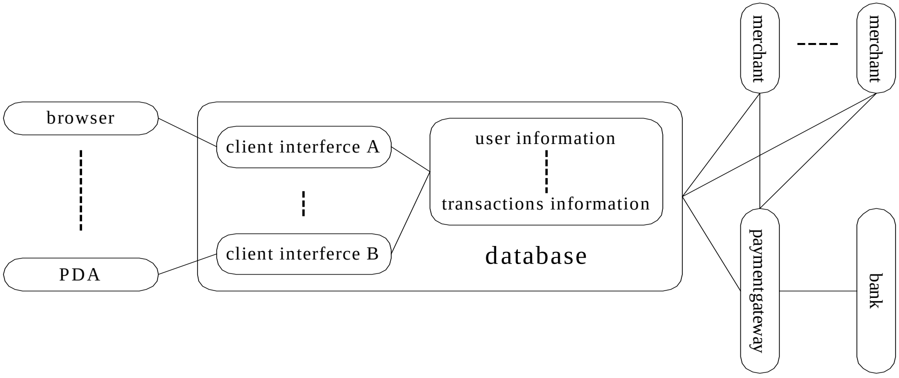


Figure 2 The server electronic wallet model

In order to deal with these limitations of the fat electronic wallet model people present the server wallet model. The server electronic wallet model is composed of two parts. One is the client side, which is generally browser. The other is server electronic wallet. The user can communicate with the merchant by browser and explore the information in the server. The server electronic wallet is responsible for the communication with the other part joined the electronic business transaction. At the Same time it manages the digital certificate and digital key. It also stores and maintains account information and transaction records and finishes the electronic payment. In a word it has the fat electronic wallet functions. At the same time it has the following advantage comparing to the fat electronic wallet model:

1. It's updating is easy owing to using the central managed electronic wallet supported the multiinterferce electronic wallet instead of the distributed the electronic wallet.
2. The server electronic wallet is operated by the bank or financial organizations, which make the data security and improve the ability of calamity restoring. In addition it decreases the requirements of bandwidth and computing ability of the user' computer.
3. The information of the user is stored in the server electronic wallet side. It is not stored in the personal computer. That improves the security of the data of user.
4. The user can pay the bill by the server electronic wallet in anywhere without taking with the special personal computer installed the fat electronic wallet.
5. It is platform compatibility because it generally applies the Browse/Server model.

Hence our proposed electronic wallet architecture supported multi-payment protocols and instruments is based on the server electronic wallet model.

3. A SERVER ELECTRONIC WALLET ARCHITECTURE SUPPORTED MULTI-PAYMENT PROTOCOLS AND INSTRUMENTS

The server electronic wallet is composed of login module, user information module, register module, payment instruments module, transaction evidence module, payment protocol module, transaction module, help module and logout module.

The user uses the SSL protocol to communicate with the server electronic wallet. Secure basic web services provide the encryption, decryption, digital signature and verification digital signature functions. The server electronic wallet architecture is secure and extensible. Its extensibility is achieved by the payment protocol module and payment instruments module. We can add a new payment instrument by modifying payment instrument module and add a new payment protocol by modifying payment protocol module. Its security is achieved by SSL protocol and secures basic web services.

The login module is responsible for the management of the login of the user, such as authentication.

The user information module is responsible for the management of the information of user, such as modifying the information of the user.

The register model is deal with the user register application.

The payment instrument module is one of the most important part of the server electronic wallet and is responsible for the choice of the payment instruments and manages the payment instrument, such as adding, deleting, modifying, displaying the payment instrument and so on.

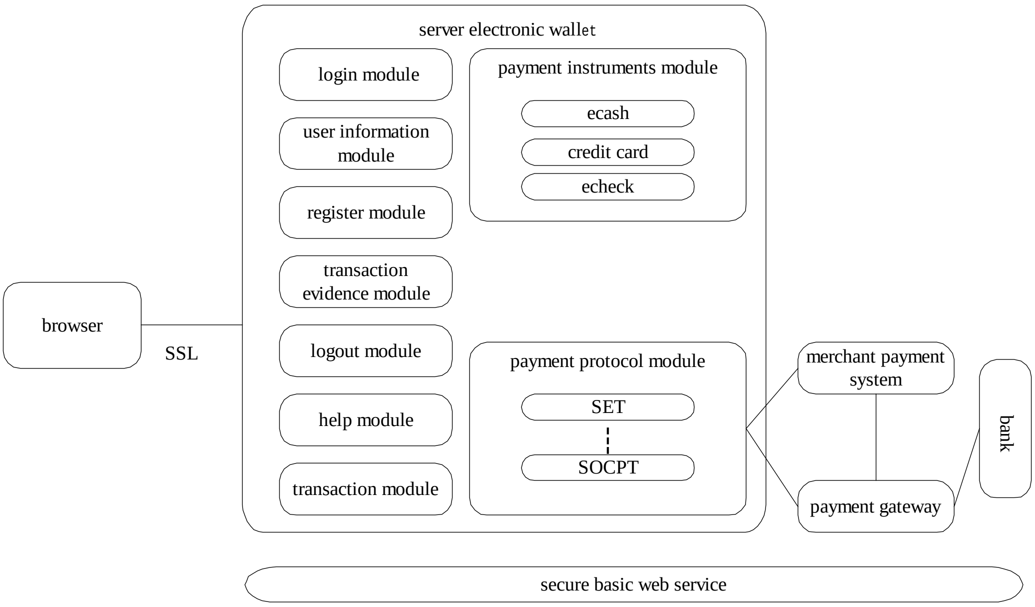


Figure 3 A server electronic wallet architecture

The transaction evidence module takes charge the management of transaction evidence information.

The payment protocol module is the core of the server electronic wallet and is responsible for which payment protocol is chose by the user and the merchant as the payment is doing.

The transaction module is responsible for displaying, deleting, inquiring, downloading the each transaction information. In addition it can display the status of the transaction.

The help module is help system.

The logout module is responsible for the logout of the user.

4. PROTOTYPE IMPLEMENTATION

The prototype implementation of server electronic wallet architecture supported multi-payment protocols and instruments applies XML technology, web service and CAPICOM. The operating system is Windows 2000. The database is Microsoft SQL 2000.The web server is IIS 5.0. The develop tool is visual Studio.NET. At the present time the development of Chinese version have been finished. The English version is developing.

In order to test our server electronic wallet we develop a merchant system including merchant payment system and merchant online shop and a payment gateway based DataCash bank. Own to the space limitation and its irrelative to the topic of this paper, we don’t describe design and implementation of merchant system and payment gateway based on DataCash bank.

Thereinafter we discuss the XML technology, web service and CAPICOM.

4.1 XML Technology

XML [6] was developed by an XML Working Group formed under the auspices of the World Wide Web Consortium (W3C) in 1996. In 1998 W3C made it as the recommendation standard. XML is a subset of SGML (the Standard Generalized Markup Language). XML is an application profile or restricted form of SGML. Not only has it the powerful function of SGML but also the simplicity of HTML. XML describes a class of data objects called XML documents and partially describes the behavior of computer programs which process them. XML documents are made up of storage units called entities, which contain either parsed or unparsed data. Parsed data is made up of characters, some of which form character data, and some of which form markup. Markup encodes a description of the document's storage layout and logical structure. XML provides a mechanism to impose constraints on the storage layout and logical structure. At the present time XML has been the standard of Internet data exchange of electronic business. So In our server electronic wallet prototype system we use XML as the data exchange format.

4.2 Web Service

A Web service [7] 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 [8,9,10]). Other systems interact with the Web service in a manner prescribed by its description using SOAP [11,12] messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.

A Web service is an abstract notion that must be implemented by a concrete agent. The agent is the concrete piece of software or hardware that sends and receives messages, while the service is the resource characterized by the abstract set of functionality that is provided. To illustrate this distinction, you might implement a particular Web service using one agent one day, and a different agent the next day with the same functionality. Although the agent may have changed, the Web service remains the same.

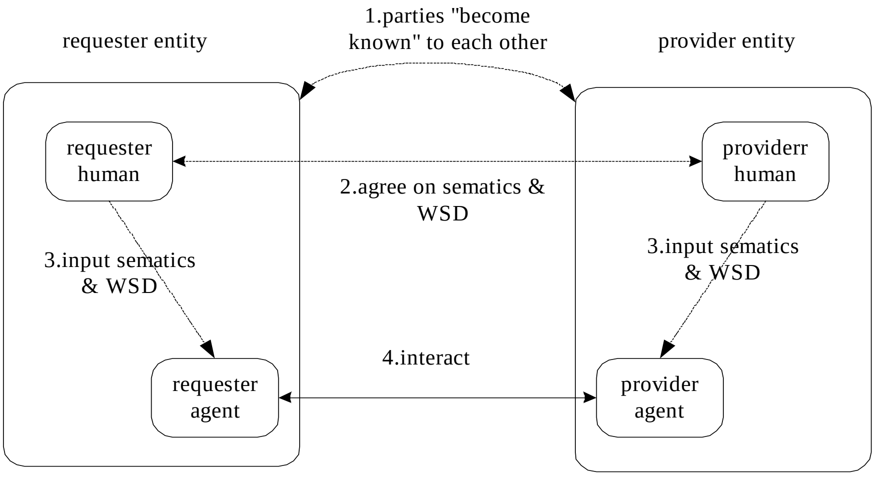


Figure 4 The general process of engaging a web service

Table 1 Web services developed by us

Web service	Description
PinitReq web service	Send to merchant
Opreq web service	Send to merchant
PinitRes web service	Send to server electronic wallet
AuthCapReq web service	Send to payment gateway
OPres web service	Send to server electronic wallet
signedata web service	Public service
Enc web service	Public service
EncB web service	Public service
Decrpt web service	Public service
Verify web service	Public service
Ex web service	Public service
GateWay web service	Send to merchant
DataCash web service	Send to payment gateway

Table 2 CAPICOM description

Category	Description
Certificate Store objects	Objects available for using certificate stores and the certificates in those stores
Digital Signature Objects	Objects used to digitally sign data and to verify digital signatures.
Enveloped Data Objects	Objects used to create enveloped data messages for privacy and to decrypt data in enveloped messages.
Data Encryption Objects	Objects used to encrypt data and to decrypt encrypted data.
Auxiliary Objects	Objects used to change default behaviors and to manage certificates, certificate stores, and user interface (UI) messages.
Interoperability Interfaces	Interfaces that allow derivations of CryptoAPI to work together with CAPICOM 2.0.
Enumeration Types	Enumeration types used with CAPICOM.

There are many ways that a requester entity might engage and use a Web service. In general, the following broad steps are required, as illustrated in Figure 4[7]: (1) the requester and provider entities become known to each other; (2) the requester and provider entities somehow agree on the service description and semantics that will govern the interaction between the requester and provider agents; (3) the service description and semantics are realized by the requester and provider agents; and (4) the requester and provider agents exchange messages, thus performing some task on behalf of the requester and provider entities.

We have developed thirteen web services described in Table 1 in our server electronic wallet prototype system.

4.3 CAPICOM

The CAPICOM [13] provides services that enable application developers to add security based on cryptography to applications. CryptoAPI includes functionality for authentication using digital signatures, for enveloping messages, and for encrypting and decrypting data. The CAPICOM is described in Table 2.

5. CONCLUSIONS

With the development of electronic commerce, more and more people and enterprises do electronic business on the Internet. So the electronic payment tools, which should be secure and extensible, are needed very much. Electronic wallet is one of the most important electronic payment tools. Electronic wallets can be classified into two classes with whether they are based on fat electronic wallet model or server electronic wallet model. At the present time many electronic wallets, some are fat electronic wallet, others are server electronic wallet, have been developed. But they don't support the multi-payment protocols and instruments together, which is not convenient to the people and enterprise when they have electronic business. In this paper we research on the fat electronic wallet model and the server electronic wallet model and analyses the difference between them. Secondly we present a new server electronic wallet architecture, which supported multi-payment protocols, and instruments, based on sever electronic wallet model and is be secure and extensible. In the end we give its prototype implementation based on the architecture presented by us with the extensible markup language, web service and CAPICOM technologies.

6. REFERENCES

[1] Andrew Hinchley ， Authentication and transaction support-the role of electronic wallet ， http://www.radicchio.org/member_center/download/t2r/9th-tb5-andrew-Hinchley.ppt

[2] http://supermart.stockstar.com/supermarket/chinabank/aqzs.htm

[3] https://paygo.51qb.com.cn/index.jsp

[4] https://www.szpos.com/webhelp/faq.jsp#2

[5] Xiaoling Dai, John Grundy, "Architecture of a Micro-payment System for Thinclient Web Applications", In proceedings of the 2002 International Conference on Internet Computing, Las Vegas, June 24-27 2002.

[6] http://www.w3.org/XML/

[7] Web Services Architecture. W3C Working Draft, 8 August 2003. http://www.w3.org/TR/ws-arch/

[8] Web Services Description Language (WSDL) Version 1.2: Part 1: Core Language, http://www.w3.org/TR/wsdl12/, W3C Working Draft, 11 June 2003

[9] Web Services Description Language (WSDL) Version 1.2 Part 2: Message Patterns, http://www.w3.org/TR/wsdl12-patterns/, W3C Working Draft 11, June 2003

[10] Web Services Description Language (WSDL) Version 1.2 Part 2:Bindings, http://www.w3.org/TR/wsdl12-bindings/, W3C Working Draft ,11 June 2003

[11] SOAP Version 1.2 Part 0: Primer, http://www.w3.org/TR/soap12-part0/

[12] SOAP Version 1.2 Part 1: Messaging Framework, http://www.w3.org/TR/soap12-part1/

[13] http://msdn.microsoft.com/library/default.asp?url=/library/en-us/security/security/capicom_reference.asp

[14] Michael Peirce, "Multi-Party Electronic Payments for Mobile Communications", PH.D thesis, University of Dublin, 2000.

[15] Steven P. Ketchpel, Hector Garcia-Molina, and Andreas Paepcke. "Shopping Models: A Flexible Architecture for Information Commerce", Proceedings of the Fourth Annual Conference on the Theory and Practice of Digital Libraries, 1997, http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDLWP-1996-0052.

[16] SOAP Version 1.2 Part 0: Primer, http://www.w3.org/TR/soap12-part0/

[17] SOAP Version 1.2 Part 1: Messaging Framework, http://www.w3.org/TR/soap12-part1/

Meng Bo is a postdoctor of school of computer of Wuhan University. He gets his PH.D in Wuhan University of Technology in 2003.

Now his research interests are in electronic commerce, information security and electronic government.

ebXML: the Global Standard for Electronic Business

Chen Caixian, Ran Chunyu, Hu Hengying
College of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China
Email: chy7627@163.com **Tel:** 027-87297484

ABSTRACT

Today the solution of global electronic commerce has been the focus of much attention by the electronic commerce community, the analysts and the press. The article addresses the drawbacks of traditional standards aiming at realizing global e-business system framework and putting forward an ideal standard which gracefully reaches the global-electronic-commerce goal---ebXML. Following of this the article elaborates on the ebXML framework in details, including the concepts of ebXML, the principle of ebXML technology and the overview of ebXML standard.

Keywords: ebXML, EDI, electronic business

1. INTRODUCTION

The Internet and the World Wide Web have dramatically changed the way companies can do business. E-business has seen a surprising increase in just a few years. Today, the world is progressively engaging itself in creating technological advancements for the purpose of expanding international business trade and to increase efficiency throughout their B2B (business to business) value chains. The goal has been attempted in the past few years, but with less-than-stellar results.

2. EDI AND XML

EDI (Electronic Data Interchange) is a standard designed to provide business forms for transmission of electronic data from one trading partner to another. Organizations conduct EDI transactions using a wide range of communication infrastructures, including VANs (Value Added Networks), VPNs (Virtual Private Networks). EDI has been proved to be expensive due to the high cost in network infrastructure setup/running and System Integration. As a result, it is only practical for large enterprises. Many companies (especially small business or organizations) have tired to build their electronic businesses and collaboration network with their trading partners in an ad-hoc manner. So, EDI is not a practical solution to encourage all organizations to conduct electronic business over the Internet.

Another standard, XML (extensible markup language) provides a collaborative methodology to facilitate electronic data interchanging over a global scale. One of XML's design goals is to enable B2B transmissions via a flexible exchange of data. XML meta data format makes it easy to create both XML documents and the applications that process them. In addition, XML passes through corporate firewalls as a common language from computer to computer, supporting real-time transactions over the internet. XML can be customized and integrated with other processes, applications and systems, including EDI- based legacy systems. With regard to XML functionality and flexibility, this standard is poised to become the SME (small to medium enterprise) complimentary of EDI. XML is a powerful, flexible tool. However, as with any technology by itself, XML is not enough to encompass the business process requirements. Furthermore, XML does not solve interoperability issues (but is an important tool for doing so). Also, using XML alone does not provide semantics and partnership agreements.

What is a set of ideal standards to make the global electronic commerce a reality is the vision of ebXML (electronic business XML). It provides uniform guidelines to boost effective practices, from streamlining inventory and shipment procedures, to automating how one communicates a purchase order or invoice using the internet while offering easy implementation and management to (SME).

3. EBXML: THE SOLUTION OF GLOBAL ELECTRONIC BUSSINESS

Concepts of ebXML

ebXML is a global electronic business standard that is sponsored by UN/CEFACT (United Nations Center For Trade Facilitation And Electronic Business) and OASIS (Organization for the Advancement of Structural Information Standards). It is a set of specifications that together enable a modular electronic business framework. The vision of ebXML is to enable a global electronic marketplace where enterprises of any size and in any geographical location can meet and conduct business with each other through the exchange of XML-based messages. ebXML has opened a whole new era of enterprise alignment, because it provides the only globally developed open XML-based standard built on a rich heritage of electronic business experience. It enables parties to complement and extend current EC/EDI investment and expand electronic business to new and existing partners, while facilitating the convergence of current and emerging XML efforts. The ebXML framework is based upon non-proprietary technology that encourages interoperability. Key areas addressed with ebXML are:

1. Open Standard
   (1) All industries are welcome to join and submit specifications and suggestions
   (2) A general vote approves all specifications
   (3) ebXML specifications are openly and freely available

2. Global Framework
   (1) Opens up business to many more potential trading partners throughout the world
   (2) Provides a single framework to exchange business data with anyone who has access to these networks

3. Convergence Support
   (1) Enables interoperability with other industries due to common elements and an open architecture
   (2) Ensures the success of common business processes

and message exchanges by reusing existing standards such as

ebXML Technology

Before looking at the whole "vision" of ebXML interactions it is necessary to digest an alphabet soup of new acronyms and other special terms, which make a good starting point. With the new vocabulary in mind you can make sense of how all of the differing processes in ebXML hold together.

Core Components:
A core component is a common, or general, building block that basically can be used across several business sectors. Components can be built together into aggregates; aggregates and components can be gathered into "document parts." The ebXML Core components are like a catalogue of XML schemas for business information entities. These are like building blocks that you can assemble for the right message for a business process.

Collaboration Protocol Profile (CPP):
A Collaboration Protocol Profile (CPP) describes the message-exchange capabilities between business partners that engage in business transactions. The CPP describes both sides of B2B transactions. CPP is like Web Services Description Language (WSDL) on steroids. CPP includes four parts: partyInfo, packaging, ds:signature and comments and every part also includes many subparts, thus the whole CPP document form a tree structure. In order to understand the format of CPP document, I provide an example of following:

```
<?xml version="1.0" encoding="UTF-8"?>
<CollaborationProtocolProfile
      xmlns:xsi="http://www.w3.org/2000/10/
XMLSchema-instance"
        xmlns:xlink="http://www.w3.org/1999/x
link"
        xmlns:ds="http://www.w3.org/2000/09/x
mldsig#"
       xsi:schemaLocation="http://www.ebxml.o
rg/namespaces/tradePartner version="1.1">
       <PartyInfo>
         <PartyId
         type="RCBF">1234567</PartyId>
         <PartyRef
         xlink:href="http://random.com/about.
         html"/>
               ...
       <Transport transportId="N03">
        <SendingProtocol version="1.1">HTTP</
SendingProtocol>
        <ReceivingProtocol
version="1.1">HTTP</
ReceivingProtocol>
          <TransportSecurity>
             <Protocol
version="3.0">SSL</Protocol>
           <CertificateRef certId="N07"/>
          </TransportSecurity>
       </Transport>
          ...
     </PartyInfo>
      ...
<Packaging> …
<ds:signature> …
<Comment> …
</CollaborationProtocolProfile>
```

Collaboration Protocol Agreement (CPA):
The message-exchange agreement between two business partners is described by a Collaboration Protocol Agreement (CPA). The CPA is a technical, binding description: It specifies the exact requirements and mechanisms for the transactions through which two companies exchange messages electronically. The CPA references the CPP business process schema definition. The CPA can define delivery channels, abstract configuration of how messages are exchanged and characteristics of message exchange on a message by message basis. By comparison, WSDL is just a common, shared endpoint.

Business Processes:
Activities that a business can engage in (and for which it would generally want one or more partners). A Business Process is formally described by the Business Process Specification Schema (a W3C XML Schema and also a DTD), but may also be modeled in UML.

Business Process Specification Schema:
The ebXML Business Process Specification Schema supports the specification of business transactions and the choreography of business transactions into Business Collaborations using the Unified Modeling Methodology (UMM). Using the ebXML Business Process Specification Schema, the user may create a Business Process Specification. A business process describes in detail how trading partners take on shared roles, relationships and responsibilities to facilitate interaction with other Trading Partners. The interaction between roles takes place as a choreographed set of Business Transactions. Each business transaction is expressed as an exchange of electronic Business Documents. The sequence of the exchange is determined by the business process, and by messaging and security considerations.

ebXML registry/repository:
The ebXML registry/repository provides a set of services that enable sharing of information between interested parties for the purpose of enabling business process integration. The shared information is maintained as objects in a repository. Submitted content may be XML schema and documents, process descriptions, core components, context descriptions, UML models, information about parties, and even software components. Registries index these items, but they are actually stored in corresponding repositories. Such information is used to facilitate ebXML-based B2B partnerships and transactions. With the ebXML registry/repository, you register your business process and are able to *store* the process. The focus of the repository is for the developer, at design time, to be able to download the schema for a business process. There is some overlap between UDDI and the ebXML registry/repository, but they are really complimentary.

ebXML Message Service:
The ebXML Message Service is provided for environments requiring a robust, yet low-cost, solution to enable electronic business. The mechanism provides a standard method for exchanging business messages among ebXML trading partners.

It is defined as a set of layered extensions to the base Simple Object Access Protocol (SOAP) and SOAP Messages with Attachments (SOAPATTACH) specifications, which have broad industry acceptance and serve as the foundation of the work of the W3C XML Protocol Core working group. The ebXML Message Service provides the security and reliability features that are not available in SOAP or SOAPATTACH but necessary to support international electronic business.

ebXML standard overview

Below provides the high level step by step of the ebXML interaction between 2 companies (Fig.1).

1. Company A will first review the contents of an ebXML Registry, especially the Core Library, which may be downloaded or viewed there. The Core Library (and maybe other registered Business Processes) will allow Company A to determine the requirements for their own implementation of ebXML (and whether ebXML is appropriate for their business needs).
2. Based on a review of the information available from an ebXML Registry, Company A can build or buy an ebXML implementation suitable for its anticipated ebXML transactions.

3. Company A creates and registers a CPP into the Registry. Company A might wish to contribute new Business Processes to the Registry, or simply reference available ones. The CPP will contain the information necessary for a potential partner to determine the business roles in which Company A is interested, and the type of protocols it is willing to engage in for these roles.

4. Once Company A is registered, Company B can look at Company A's CPP to determine that compatibility exists between the two companies.

5. Company B is able to negotiate a CPA (agreement) automatically with Company A, based on the conformance of the CPPs, plus agreement protocols, given as ebXML standards or recommendations.

6. The two companies begin actual transactions.

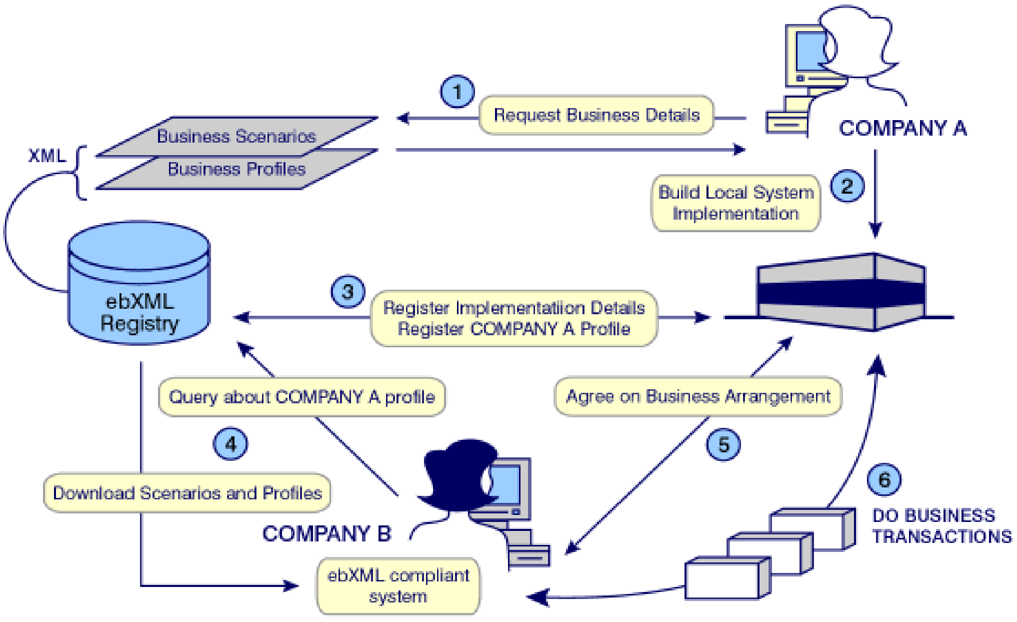

1
Request Business Details
COMPANY A
XML
Business Scenarios
Business Profiles
Build Local System Implementation
2
ebXML Registry
3
Register Implementatiion Details
Register COMPANY A Profile
Query about COMPANY A profile
Agree on Business Arrangement
4
5
Download Scenarios and Profiles
COMPANY B
6
DO BUSINESS TRANSACTIONS
ebXML compliant system

Fig.1 ebXML Interaction between 2 Companies

4. CONCLUSION

In the adoption of the ebXML standard, industries of all sizes and locations now have a standard method to exchange business messages, conduct trading relationships, communicate data in common terms, and define and register business processes. With ebXML, companies may discover alternative, potentially lower cost methods for their supply chain. There is an abundance of ways to reduce financial overhead, save expenses, and increase efficiencies. For example, ebXML provides a practical and affordable solution to SMEs by trading XML-type documents electronically, without investing in EDI systems. Alternatively, large enterprises that have invested in EDI, ebXML preserves and extends EDI functionality. Thus, ebXML can adapt to shifting industries, or transcend industry boundaries, exchanging data in a common format with existing or new trading partners.

5. REFERENCES:

[1] Benoit Marchal. An Introduction to the ebXML CPP. http://www.developer.com/xml/article.php/2247851. 8/2003

[2] Dacid R.R. Webbfer. EbXML:The New Global Standard for Doing Business. New Riders. September 26 2002

[3] Brian Gibb, Suresh Damodaran. ebXML: Concepts and Application. John Wiley & Sons, October 21,2002.

[4] Alan Kotok, David R. R. Webber. ebXML:The New Global Standard for Doing Business on the Internet. New Riders, 1st edition, August 23,2001.

[5] Joseph Chiusano. UDDI and ebXML Registry: A Co-Existence Paradigm. 4/2003

Ran Chunyu is a Full Professor and a head of computer system institute in computer science and technology department, Wuhan University of Technology. He graduated from Beijing architectural material college in 1969 with specialty of automatic control in engineering department. He once studied in Nanjing University and Huazhong University for computer courses, and then went to the Rumanian International Training Center for further study. He have undertaken some research items of “ninth five-year national key science and technology project”. Now, he is carrying on some major researches of “key scientific research projects at provincial bureau level”. More than 30 articles of him have been published in native or foreign journals and 5 teaching materials are compiled with his attendance. His research areas are computer applications like network database, communication security, graphics and images, and so on.

Chen Caixian is a graduate student of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. She graduated from Qufu Normal University in 2002 with specialty of computer technology in Computer Science and Technology Department. Her research areas are network database, e-commerce applications etc.

The Design and Implementation of Distributed Database In Trust Investment Synthesized Business System

Kaiying Yang, Fahong Yu
Department of Computer Science & Technology, Wuhan University of Technology
Wuhan City, Hubei Province, 430070
Email: fhyu520@hotmail.com **Tel:** +86 (0)27-87371513

ABSTRACT

This paper analyzed the characteristics and requests of trust investment synthesized business system, and then designed and implemented a distributed database based on them. The design and implementation were the embodiment of the distributed database technology, which assured the security, reliability and flexibility of the database.

Keywords: Distributed Database, Trust Investment Synthesized Business, Transaction, Data security, Data Fragmentation

1. INTRODUCTION

Distributed Database (DDB) is a product mutually infiltrated and organically integrated by network technology and database technology. Compared with traditional database, it has surpassing merits: Good compatibility and usability, reliability, flexibility, and expansibility and so on. Therefore, the application of distributed database technology become more and more widespread. The domain name system (DNS) is a typical instance for the application of DDB.

In a trust investment companies, the decentralization of operational organizations result in the scatter of operational data. As a company and its branches locate in different cities or different regions in the same city, it needs not only the local control and the decentralized operation but also the whole organizational control and the high level coordination management. These are the just questions solved by Distributed Database System (DDBS). The following is the design scheme and application of DDB in trust investment business synthesized system.

2. THE THEORY AND TECHNOLOGY OF DDB APPLIED TO TRUST INVESTMENT SYNTHESIZED BUSINESS SYSTEM

DDB is composed by a group of databases which are distributed physically in different nodes and spots of network, but it is a integral system in logical. By using of the network, those logical units are dispersed to different positions to manage and control in different degree centralization and connected into a united database system. It has two major characteristics, one is the data distribution, another is the cooperation among databases. The emphasis of DDB is knot's autonomy instead of the system is centralized control. The transparency of data distribution should be maintained in system, which enable the application programmer not to consider the data distribution completely when they are programming. It is necessary for many large and middle enterprises to use DDB safety and convenience to operate when they carry out informalization. The technologies that are used to satisfy the needs of the design and application mainly include:

Data Fragmentation
According to demands of enterprises and system, datum is divided into many fragments that are no intersection. Generally, the units of data storage are not relations but fragments. So it is helpful to organize the distribution of data according to the needs of users, and it is also helpful to control the redundancy of data. The basic calculations of relational algebra are used in the realization.

Data Assignment
In DDBS, data assignment means that data fragments are allocated to different physical database on network according to a certain strategy. The ways of assignment should base on the actual demands.

Data Storage
Different from central database system, data redundancy is regarded as the needed characteristic in DDBS. The reasons lie in: Firstly, the local application will be improved if data are copied on the needed nodes. Secondly, the copies on the other nodes can be operated when a node is wrong, which can enhance the validity of system. Of course, the appraisal for the best redundancy in DDBS is very complex.

In DDB, data storage should be carried out in three different three ways:
(1) Fragmentation: A relation is divided into several fragments, and each fragment is saved on the different node.
(2) Duplication: The several quite same copies involved in system maintenance are saved on different node.
(3) Duplication + Fragmentation: A relation is divided into several fragments, and several copies of each fragment are maintained in system.

3. THE DESIGN OF DDBS IN TRUST INVESTMENT SYNTHESIZED BUSINESS SYSTEM

Trust investment as a financial field, there are many features, such as vast data, frequent transaction, high demand of data security, and distribution of different areas. A high performance DDB conformed to demands in this field has been designed. According to design's demands of trust investment synthesized business system, the system's structure of DDB should be considered and designed as following:

Distributed Scheme
Considered entire application, the top-down method is adopted to constitute DDB and to realize data integrity and

uniformity. Only the local data are stored in various branches, while all business data are stored in parent company where the data integrity and uniformity are examined. The Top structure of servers of distributed database is shown in chart1. Three main branches are respectively disposed a database server which is used to store the local data, simultaneously, those local data must be identically store to the central server in parent company's (a total database server and a backup server). Although there are some data redundance, the method that several copies of the same data are stored in different position can improve the reliability and the usability of system, improve the efficiency of local application, and decrease the communication cost. This system of database can be extended in the condition of lowest effect to the current organization, at the same time, it can decrease the interference to the lowest between different organization. Only a node is added, a new business organization will be produced.

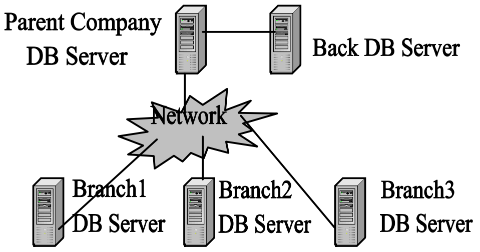


Chart1 the Top structure of database servers

Design Of Data Fragmentation

In DDBS, relation fragmentation is useful to organize the distribution of data according to the user’s demand. At present, there are four ways about fragmentation: Horizontal Fragmentation, Vertical Fragmentation, Educing Fragmentation and Mixture Fragmentation. According to the relation between different data, the different fragmentation method is selected as following:

(1) In the relation of data between parent company’s database and branches’, since the data in branch is the subset of parent company, Horizontal Fragmentation method is selected, and the reconstruction of relation is realized by calculation.

(2) In the relation of data between parent company’s database servers and Web server’s, since the datum is divided according to its function, Vertical Fragmentation method is selected.

Assignment Of Fragments

It contains confirming the assignment of fragments on different areas and the fragments that should be copied.

In order to reduce the communication and to guarantee the autonomy in various stations, the data in standard databases and in tables about user’s information must exist the same copies in different server and the consistency of data should be kept by merging and duplicating. Simultaneously, in order to reduce data redundance, the content of operational information, clerks’ information, clerks’ rights and so on in different branch should be different. That is to say each branch only contain its relative contents.

Design Of Physics Database

It refers to a access method which is used to map a conception pattern to the area of storage and determine proper data.

The design of physical database is based on demands of current software system, and on this basis, DDB should tell apart the source of data. Therefore, some fields are added in some tables which need to be merged and duplicated so as to mark the source of data.

Design Of Data Storage

DDBS can store data through three different ways: copying, fragmentation and copying + fragmentation. Since there are some redundancies and differences among distributed databases, the way of copy + fragmentation is adopted to store data in this design.

4. THE TECHNOLOGIES AND METHODS OF DDB USED IN TRUST INVESTMENT SYNTHESIZED BUSINESS SYSTEM

The Embodiment Of Data Fragmentation Technology

In the course of data fragmentation, some tables in special region are fragmented by horizontal relationship according to the actual situation of trust investment synthesized business system. For example, for the tables about business information, the definition of the data of the table01 about business information in branch1 is described by SQL language.

DEFINE FRAGMENT SF1 AS SELECT * FROM table01 WHERE company = branch1

Using Duplication To Maintain The Consistency OF Data

According to the design’s idea as above, the method of duplication based on publisher/subscriber is adopted. Chart2 has embodied Top structure of subscription and release among parent company and branches, which aim to implement the consistency and integrality of data.

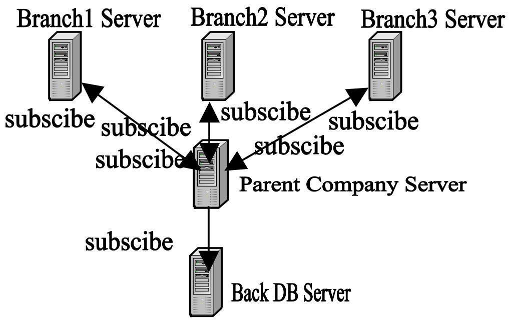


Chart2 duplication system structural based on publisher/subscriber

As shown in chart2, server in parent company subscribes some data from branches servers, and also issues some recent data to branches. Similarly, when various branches subscribe the data from parent company, the branches issue data to parent company. But the data of subscription and release are different, which can be explained as table1.

Table1 subscription and publishingin the duplication system structure

Data Table	subscriber	publisher
Customer Information Table	Parent Company Server, Branch Server	Parent Company Server, Branch Server
Standard Data Table	Branch Server	Parent Company Server
Assistance Information Table	Branch Server	Parent Company Server
Business Information Table	Parent Company Server	Branch Server
……	……	……

In order to reduce data redundance and transmission of information, all contents subscribed by branch should be correlative with itself. So the contents of subscription should be designed in horizon establishment. Obviously, when the table is designed, the field used to distinguish information of different company must be added in order to mark various company, however, when branches subscribe data, such field is unnecessary.

Accessing Technology Of DDB In Subsystems

Power Builder as a front-end developing tool of branches' system is adopted to develop DDB in Trust Investment Synthesized Business System. Transaction object is used to access the database. The mode of connection between the front-end program and the sub-database of the system is a direct connection through a special interface. The following script segment described how to connect with various database when users of subsystem enter to access the database.

```
//Read parameters from the initialized file and connect database directly by a special interface.
disconnect;
SQLCA.DBMS=
ProfileString(“.\jsjprofile.ini”,“Database”,“DBMS”,“ ”)
SQLCA.Database=
ProfileString(“.\jsjprofile.ini”,“Database”,“Database”, “ ”)
SQLCA.LogID=
ProfileString(“.\jsjprofile.ini”,“Database”,“LogID”,“ ”)
SQLCA.LogPass=
ProfileString(“.\jsjprofile.ini”,“Database”,“LogPassword”,“ ”)
SQLCA.ServerName=
ProfileString(“.\jsjprofile.ini”,“Database”,“ServerName”,“ ”)
SQLCA.UserID=
ProfileString(“.\jsjprofile.ini”,“Database”,“UserID”,“ ”)
SQLCA.DBPass=
ProfileString(“.\jsjprofile.ini”,“Database”,“DatabasePassword
”,“ ”)
SQLCA.DBParm=
ProfileString(“.\jsjprofile.ini”,“Database”,“DbParm”,“ ”)
SQLCA.AutoCommit = true connect;

// Connect with the database which has been assigned by above parameter. …
```

In the script segment, jsjprofile.ini is an initialized file, which is a structured text as INI file in Windows. Such file is composed with several sections and various sections contains many assigned statements of variable. With the parameter ProfileString provided by PowerBuilder, each attribute value of transaction object --- SQLCA can be acquired from the structured text.

Distributed Transaction Processing

The distribution of data in DDB results in the distribution of transaction. A execution of entire transaction can be divided into the execution of many sub-transactions on many spots.

To make it is possible that the distributed transaction is executed on many servers, MS DTC as a transaction manager are used to cooperate the processing operation of every server. In order to reduce the influence of network's obstruction to the distributed disposal of transaction and avoid the inconsistency of data among different servers which caused by distributed transaction. According to X/Open XA, the process of the distributed transaction is prescribed to two phase, that is the phase of preparation and the phase of submission.

When the distributed transaction is processing, the first thing is to start up a distributed transaction by word “BEGIN DISTRIBUTED TRANSACTION” in the script program of Transact SQL in the server which will be regarded as distributed transaction operational server. Then script procedure carries out the storage process or the distributed inquiry on remote server, distributed operational server will call MS DTC automatically and make remote server take part in the process of distributed transaction. When the script procedure “COMMIT TRANSACTION, COMMIT WORK, ROLLBACK TRANSACTION” or “ROLLBACK WORK” is implemented, the distributed operational server will recall MS DTC again, which is used to manage the two-phase process of submission and make connective server and remote server submit or roll back.

Distributed Inquiry

The distribution of data in DDB results in the distribution of inquiry. Distributed inquiry may aim at the heterogeneous OLE DB or ODBC data source. SQL Server supports distributed inquiry, which includes data inquiry from two or more servers. It also supports retrieval, renew and cursor and use Microsoft Distributed Transaction Coordinator (MS DTC) to ensure the meanings between nodes and uphold the safety of servers.

In the process of systematic design, to reduce the communicational capacity of network, data relations have been fragmented and stored in various database according to the function of application, so the majority of applications are faced with operation of local database, but entire inquiry still needs the support from many database. In the administrative module of clerks, due to the direct management of branches to clerks and difference in operational regulations, the information of clerks is stored in database of branches, and all clerks are registered by united distributed inquiry. In the operational module of customers, customers information is respectively stored into traditional customer's table of the server of business database and customer's table of Web server database according to the origination of respectively customer, so all clerks and customers are registered only by united distributed inquiry. The following take clerks' yearly inquiry for an example to introduce the method of united distributed inquiry.

```
SELECT emp.emp_name, emp.emp_id, emp.emp_gender…
FROM DBServer1.business.dbo.employee AS emp WHERE
date between '01/01/2003' and '12/31/2003'
UNION
SELECT emp.emp_name, emp.emp_id, emp.emp_gender…
FROM DBServer2.business.dbo.employee AS emp WHERE
date between '01/01/2003' and '12/31/2003'
```

The Data Security Of DDBS

Safety and secrecy are important in financial system. As a database system that connects with vocational safety and secrets of company, so the basic means of safe guard is not enough, especially guarding against Hacker to bypass the safety mechanism of database and directly access database by real-time monitoring the message in data transmitting or using the security hole of operational system or protocol of network. Such is extremely dangerous. Therefore, the following two security means should be adopted.

(1) The secrecy of data communication.

After succeeding in checking ID between client and server or server and server, the data can be transmitted. In order to resist massage could be attacked and intercepted, a secret passage which is used to encrypt data to transmit can be built between both sides of communication. In DDBS, the quantity of transmission is very large, so the speed of algorithm about encryption and decryption may bring influence to the performance of system. Asymmetrical password system is very complicated and its rate is very slow, so the algorithm of symmetrical password is adopted to encrypt. In fact, the process of creating secret passage is the process of dialogizing key by promise of both sides. Usually, it is connected with ID authenticating. Secret passage can be realized by DDB and also can adopt safety regulation that offered by the first floor network protocol, such as SSL (Secure Socket Layer).

(2) Database encryption.

To prevent Hacker, it is necessary to encrypt database. There are two methods of encryption. One is DES which is a encrypting standard issued by ANSI. DES includes 64 secret bits and its algorithm is implemented on a chip, which work at 1Mb/S. Another method is called secrecy system with public key, its idea is that each user has two codes, one is used to encrypt, another is de encrypt. The code of user’s encryption is public to everyone that just like telephone number, but only the corresponding key can decrypt it. The key can never be educed from the encrypting code because the encrypting system is asymmetrical, namely the process of encrypting is irreversible (This algorithm cannot be proved at present, and there is no method to prove that it is reversible.

5. CONCLUSIONS

Since the technology of DDB was introduced to trust investment synthesized business system, the conflict between the data decentralization and centralized management has solved effectively, and the sharing and exchange of data has realized. As the speeding up of the informational construction, DDB will be a bright prospect. The paper has elaborated the establishment of financial system based on DDBS, and realized its autonomy of various branch in different regions and parent company’s central management. It also improved the speed of accessing and the efficiency of the system.

6. REFERENCES

[1] Shixuan Sa, shan Wang, Introduction of Database System(third edition), [M] Beijing: Higher Education Press, 2000.

[2] (A) Jie Wu, The Design of Distributed System, Machinery Industry Press, 2000.

[3] Wei Liu, Distributed Database and Its technology, Newspaper of Changchun University Journal, February, 2000.

[4] Weiyuan Cao, Overlapping MIS Platform of C/S Model and B/S Model, Micro Computer Application, 1999

[5] Sarah Brown, Object Design of a Distributed Client/Server System, IEEE, 1998.

[6] Elmasri, Ramez and Navathe, Shamkant B,Fundamentals of Database Systems,2nd Ed.Benjamin, 1994.

[7] Yusha Liu, The Safety System Scheme Based On Client/Server, Micro Computer And Its Application. 2001.

The Realization of CAI on Campus Net

Wang Jing
Art and Information Design, Department of Electronics and Information Engineering,
Kyushu Institute of Design, Kyushu Kyoritsu University, Japan
jingjingww2003@yahoo.com

ABSTRACT

In this paper, the author introduces the distributed processing information investigation, real-time multimedia information reciprocation and streaming media technology application when we make the CAI system programming on campus net.

Key words: distributed processing, CAI, streaming media.

1. INTRODUCTION

In recent years, multimedia computers are more and more popular. The abilities of processing the sounds, words, pictures, cartoons, video information synthetically changes the pattern people using computers. The interface more similar to human nature and the more convenient reciprocal methods compel us to process the better CAI system. With the rapid development of computer and network technologies, we can not only transfer the data mainly on test files and hypertext files on-line, but also transfer those streaming media data characteristic of real-time and reciprocation on the wide band network. Therefore, the CAI system on campus net that consist of more abundant and practical contents, especially it has great influence on the reform of teaching means and methods in the art universities and colleges in reality.

2. DISTRIBUTED AND LAYERED ARCHITECTURE SYSTEM

The distributed service system, which is constructed in the form of client/server, has been one of the main service models on Internet. According to our realistic situation, the model of CAI system we adopt is multi-grade of software architecture with two layers. The basic architecture is client/server.

The whole CAI system is programmed as a sub-system on the campus net.

3. THE FUNCTION REALIZED ON CLIENT

The users come into the CAI system by the programming particular (software) entrance after they enter the campus net .On the main page of the system, it introduces briefly the teaching situation and the direction of the scientific research in schools, and it makes arrangements for the system function choice item listed in the form of menu checking. That is, “course inquiry item”, “scientific research work inquiry item”, “system revising item” and “help item”, in which the “course inquiry item” and “scientific research work inquiry item” may interactive check the related contents that users are interested in and then get into the next sub-item.

In this sub-item, the users can inquire the introduction of the lecturers and their research directions, the introduction of the courses and the related ones. In addition, they are able to do the jobs, such as “course-learning on Internet”, “teaching seminar on-line”, “learning forum on Internet”, “assignments/works exhibition on Internet”, “transferring E-mails and supervising the system”, etc.

In the various functions above, because the contents of the “course-learning” are the network and the content is the multimedia type, it greatly attracts the students, “teaching seminar”, which is equal to a small real-time video conferences, makes the teachers and the students exchange their opinions with each other face-to-face. It may not only solve the difficult problems of the courses after discussion, but also intensify the affinity of their relations. As for the “assignments/works exhibition on Internet”, it offers an opportunity for students to reveal their brilliances and learn something from each other, “learning forum on Internet” is a volume which is used by the students to exchange their learning problems in the form of words. Discussion learning is very popular in the teaching of European and American schools.

The “system revising item” installed in the main page of the system is a pattern plate programmed in particular for teachers to revise the contents of the CAI system; the “help item” is a “system using instruction” programmed in particular for the beginners; moreover, the “information issuing item” installed in particular is taken as the content of real-time information to broadcast continually on the main page.

4. THE FUNCTION REALIZED ON SERVER

Considering the computer network hardware resources that the schools have possessed at the moment and the present conditions of the campus network administration. The CAI is only an independent system “on the software”. Basically, the hardware system of the original campus network is not needed to change too much.

According to the requirements of the client, the server actually should transfer two kinds of information. The first kind is text or hypertext files, which can be realized by a family of TCP/IP(Transfer Control Protocol/Internet Protocol) and its sustained protocols, such as HTTP（Hypertext Transfer Protocol）, NNTP(Network News Transfer Protocol), SMTP(Single Mail Transfer Protocol), POP3(Post Office Protocol Version 3), IMAP(Internet Message Access Protocol), FTP(File Transfer Protocol), and remote control and management. The second one is the “multimedia data” processed into the streaming media data type. The “multimedia data” must be achieved by the family of RTP / RTCP (Real-time Transfer Protocol / Real-time Transfer Control Protocol) and their related protocols. On the other hand, the CAI system itself is a distributed application system. The various kinds of information are stored on the different hosts according to the different type of the relevant teaching contents, respectively. Consequently, the CAI system actually is an applied software system under the sustenance of the network operating system.

According to the above requirements and our real situations, Microsoft Windows 2000 Server is selected as the network operating system. However, on the basis of the product technology index of Microsoft Windows 2003 Server the Microsoft Company issued in the latest, if we choose the latter, it will do well to the expansion of our CAI system function. The system has obvious improvements as follows.

Internet Information Service (Version6.0) has developed into a reliable web server. It not only offers the global web service, but also process FIP server, SMTP server, NNTP server. Compared with Windows 2000 operating system, it has two important features. First, it may adopt the applied programs of ASP.NET exploitative users, stations. ASP.NET, established on the Microsoft.NET Framework, makes use of the Common Runtime to provide users for the translatable and editorial framework, which is able to set up the applied server with powerful function at the end of the server. Second, it employs the brand-new file storage pattern XML and sustenance of Internet standard, such as SOAP. What is more, Windows Server 2003, which has had the Microsoft.NET integrated, enables the clients to use any programming language, such as VB.NET, C++.NET, C#.NET, so that they can finish their program designing tasks more efficiently and quickly. According, the end of the server selects the operating system and makes it greatly improved in the efficient running, extensibility, applied sustenance and administration.

The cluster technology is a universal solving scheme that the operating system provides for the organizing institution needing high elasticity, reliability and availability. It is applicable extensively to TCP/IP network. Then the users will acquire the availability of high level by redundant CPU, storage and data route at the very low cost.

The NLB (Network Load Balancing) technology of the cluster system allows all resources of servers dynamic to improve the capacity.

To support the advanced network technologies: IPV6 (Internet Protocol Version 6) sustains PPPoE of the prospective Ethernet and IPSec by NAT, making the use of Internet more convenient.

As a matter of fact, we adopt the following system service in the server, as well.

Windows Media Service is the necessity for realizing the above teaching on-line. Only if we install Windows Media Server, and supervise these teaching contents, we may call them “programs”, we can carry out the video-on-demand.

We must also install Microsoft Exchange Server, which can offer the real-time data transfer, the Video Conference system, the communicating timely and intensifying dialogue functions. These functions are essential for the CAI system to offer the relevant services. There is no doubt that we have employed and programmed a quite suitable windows interface beneficial to the operation of users in the server, system.

Windows Media Player supplies the convenience with the users to adopt the streaming media technology. The application of the system makes our programming of CAI system benefit a lot.

5. TECHNOLOGY AND RESEARCH DIRECTION

At present, the typical subjects for the streaming media technology research in the computer field include: streaming scheduling, multimedia proxy and caching, streaming application level multicast technologies. The application and research of streaming media technology are becoming the focus which the industry and scientific research institution pay close attention to at the moment.

In addition, the theoretical research on the streaming database, such as, the system architecture in the multimedia database, the administration of multimedia data, the retrieval based on contents and the relevant algorithms, are still in the initial stage. Only if we consider the intelligent optimizing method, soft computing method, etc, we are likely to do the research of the image recognition, video signal process and the related algorithms concerning the above subjects.

6. REFERENCES

[1] http://www.streamingmedia.com
[2] http://www.aciri.org/floyd/papers.htmls
[3] http://www.w3c.org
[4] http://www6.nttlabs.com/HyperNews/get
[5] http://www.iet.unipi.it/lugi/research.htm

Wang Jing is a master student, studying in Department of Electronics and Information Engineering, Kyushu Institute of Design, Kyushu Kyoritsu University, Japan. He graduated from Kyushu University, Japan, in 2004.

Study on Focus in a SIP-based E-learning System

Zeng Qingheng, Hu Ruimin
The Key Laboratory of Multimedia and Network Communications Engineering, Wuhan University,
Wuhan, Hubei Province 430079, China
Email: mmzengqh@msn.com **Tel:** 027-87653005

ABSTRACT

This paper presents a new architecture for e-learning service, which is based on the Session Initiation Protocol (SIP). The key server in the architecture is named Focus. The Focus, maintaining a SIP signaling relationship with each participant in the conference, ensures that each participant receives the proper media streams, and implements conference policies. In this paper, we propose the framework of the Focus and further study its main components related to conference control. To demonstrate the validity of the framework, we test the service set-up delays of the system in LAN, Internet, and wireless LAN access environments, respectively. Experiment results reveals that the service set-up delays are acceptable for an e-learning system. Moreover, the delay under wireless access conditions is longer than those in other access types.

Keywords: Focus; SIP, E-learning, Multimedia conference; Conference control, Mobility.

1. INTRODUCTION

Compared to H.323, SIP [1] is a relatively new standard proposed by IETF in 1999 for multimedia conferencing over IP. It is an ASCII-based, application-layer signaling control protocol that can be used to establish, maintain, and terminate calls between two or more end points. Owing to its nonproprietary advantages in protocol extensibility, system scalability and personal mobility services, SIP has gained more and more supports from industry.

In past several years, we have been involved in the development of an H323-based conference system, which can provide remote conferencing and e-learning service. However, several critical problems such as mobility issue, system flexibility and extensibility, still need to be resolved. By carefully inspecting these problems and the properties of SIP, we find that SIP is able to provide relative ideal solutions. So we latterly design a new SIP-based architecture for an e-learning system. This architecture comprises four components, i.e. web server, manager server, Focus [2] and stream server. The key component in this system is Focus, which maintains a SIP signaling relationship with each participant and is responsible for ensuring that each participant receives the proper media streams. The Focus also implements conference policies. In this paper, we have presented the framework of the Focus and studied its comprising parts concerning conference control. In order to validate the proposed architecture, we have measured its performance in terms of service set-up delay. Results obtained from experiments show that the SIP-based e-learning system has acceptable service establishing latency in LAN, Internet and wireless LAN environments, respectively. Moreover, the delay under wireless access conditions is longer than those in other access types.

The paper is organized as follows: Section 2 gives the architecture of the SIP-based e-learning system and discusses mobility issue. Section 3 presents the framework of the Focus and studies conference control components. Section 4 describes experiment environment and analyses experiment results. Section 5 is our conclusion.

2. SYSTEM ARCHITECUTURE

General Description

The system works in client/server mode and provides lecture contents broadcasting to students in synchronous and asynchronous modes. Synchronous mode provides access to real-time courses while asynchronous mode provides access to stored contents. In synchronous mode, teaching and learning can be seen as a chairperson -controlled conference, because the lecturer controls the process of lecture and decides whether or not requesting students can hold floors.

The architecture of the system is shown in figure 1.

1) Web server contains information and URLs of asynchronous/synchronous courses. It is a front end of manage server. Before using e-learning service, students and lecturers, need enroll themselves in the e-learning system through web server. During the course of lectures, lecturers, for example, visit the web server to fetch the participant list or to change conference policies related to his lecture.
2) Manage server implements the web-based conference management functions and communicates with Focuses and stream servers to provide e-learning service for users. It acts as a conference policy server [2] as well.
3) Focus is a conference server in which a number of conferences can be hold. Focus maintains a SIP signaling relationship with each participant in the conference and is responsible for ensuring that each participant receives the proper media streams. It also sends resource state notification to service users.
4) Stream server keeps the contents of asynchronous lectures. In synchronous lectures, it receives media streams from participants and mixes or transmits streams according to the Focus' instructions.

Before using e-learning service, users should register their potential locations and Focus should register lecture's locations to registrar [1] in order to make use of discovery service offered by SIP. Students get access to live or stored lectures by sending requests to lecture URIs. After receiving an access request, proxies [1] or redirect servers [1] will consult location service for routing that request.

Registrar, proxy or redirect server can be public facilities. The e-learning system makes use of the service provided by these SIP servers but we need not implement them by ourselves.

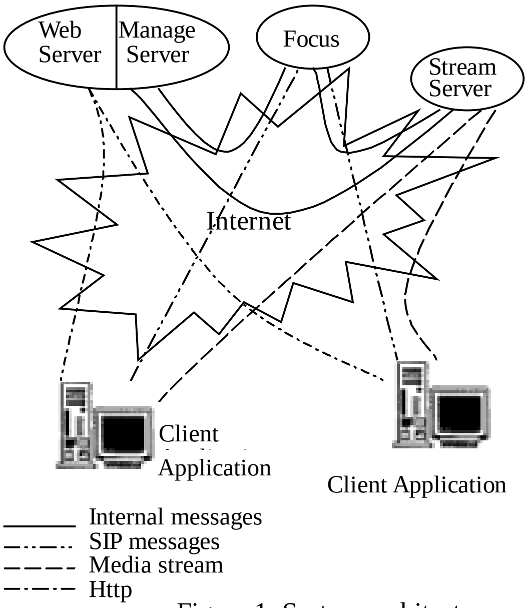


Figure 1: System architecture

Mobility Issues
In an application system, mobility issues are often related to four aspects, i.e. terminal mobility, user mobility, session mobility and service mobility. All these mobility aspects can be directly addressed with location registration and location translation supported by SIP server. For terminal mobility and service mobility, terminal registers each time after receiving the dynamic IP address, while Focus registers each time after server moved or lecture content relocation. The registration functionality allows a callee (user or service) to keep the same logical name even if his physical location changes, and the location service provided by SIP server allows a caller to establish a session with a callee by only knowing its logical name, say SIP URI.

User mobility allows proxies to address a user located at potential terminals by the same logical address. Using SIP forking proxies, users can be reached at any places where they have registered, via the same name, making their device choice transparent to third parties.

Session mobility allows a user to maintain a media session even while changing terminals. There are at least two approaches available, namely third-party call control [3] and the REFER mechanism [4], for a user to transfer media sessions from one terminal to another.

3. FOCUS STUDY

Software Framework Overview
In our e-learning system, Focus is a SIP signal converging unit and a conference control server. Before service, Focus is responsible for registering URIs of synchronous and asynchronous lectures to registrar so that service demanders can address to e-learning servers. During synchronous lectures, Focus maintains SIP signaling relationships with service users and ensures that each participant receives proper audio/video streams. When conference state changes, for example, one student joins or leaves a lecture, Focus delivers notifications to particular event subscribers [5] and informs them about the changing of the conference state.

According to the above description, it is known that Focus should fulfil the following control functions, i.e., conference management, such as conference creation or deletion, user admission and resource management, session control and conference state notification. So we devise a framework for Focus that is illustrated in Figure 2.

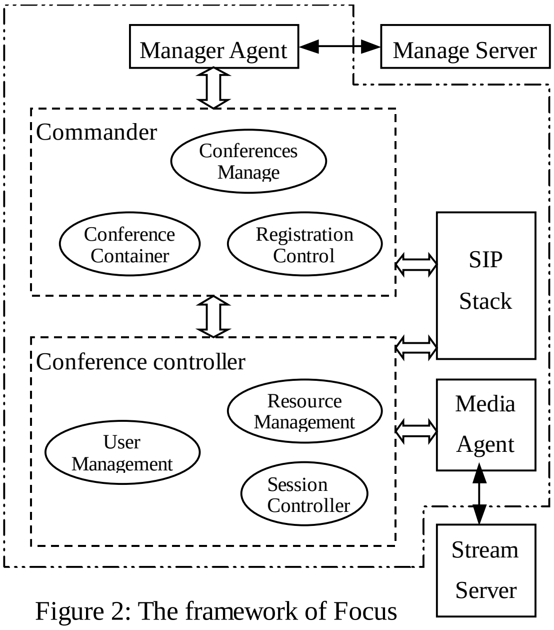


Figure 2: The framework of Focus

1) Manager agent, an internal command translation unit, is responsible for maintaining the communication between manage server and Focus. It receives XML like commands from manage server and translates them into the form that can be understood by Focus. It also sends conference information back on the request of manage server.
2) Media agent, through which Focus remotely configures or instructs stream server to insure that each student receives his requested lectures, is a communication channel between Focus and stream server.
3) SIP stack is compliant with RFC 3261. It contains implementation of SDP [6], for session description. SIP stack sends and receives SIP messages for applications and is responsible for construction, parsing, re-transmitting, filtering SIP messages and maintaining transactions.
4) Commander takes the overall charge of the local conferences management, such as conference configuration, conference creation or deletion, synchronous and asynchronous lectures registration, and instructing conference controller to implement conference policies. Commander also has a default policy for all new conferences and may restrict the operations that new conferences can support.
5) Conference controller is the key component of Focus. It carries out conference control functions and implements user management, resource management, session control and conference state notification.

User Management
The user management sub-component governs user-related properties such as user access right groups and policies. For example, the access control list may include allowing and denying definitions for incoming join attempts. When a user send a access request to Focus, the user management component will look up the access control list to see if the user

is allowed to join the lecture and which role the user will play in the lecture. Unresolved join attempts will be kept in a table until the moderator, a lecturer or a defined auto-mechanism, notified by the system, sends a request making a decision to reject or to accept the request.

On conference initialization, use management gets information from the manage server about conference participants and their conferencing terminals. The information, used mainly in call processing, is crucial and being stored in a user container. During conferencing, user management also need to communicate with other components, such as session controller and media agent, for more detailed information like terminal's video and audio parameters. This more detailed information will be sent back to the manage server later, on its requests.

User management also supports a mass-invitation feature, in which the conference server calls up participants in a dial-out conference, or a lecturer may ask the conference server to invite several registered students at once into his lecture. For mass-invitation situation, user management gets a list of invitees from manage server and passes the list to session controller rather than having to complete signaling exchanges with each participant individually.

Session Controller

Session controller mainly focuses on those activities related to connection setup, modification and disconnection. For session establishment, session controller employs a four-phase procedure. When a user should be invited into the conference, session controller is responsible for requesting participation, negotiating a common set of capabilities, initiating media connections and propagating information among peers. In contrast to the caller process, when a request is addressed to Focus, session controller should initiate a process to alert the moderator to accept or to reject the invitation. If moderator accepts the invitation, session controller checks that the request can be supported and initiates the media connections and sends out state information about the member participating in the conference. In negotiation phase, we use SDP for multimedia session description and offer/answer model [7] for SDP negotiation.

A participant may leave the conference from any states, e.g. by replying negatively to the Invite request, unsuccessful capability negotiating or pressing the disconnect button.

During the conference, session state may get out of synchronous due to the distributed nature of users. For example, the lecturer believes student A is in conference, but student A is in fact not in the conference at all. This may be the result of an unsuccessful disconnect, due to network failure or client program unexpectedly restarting. In order to deal with user unexpectedly disconnected situation and to synchronize session state, session controller should send INFO [8] messages to conference participants at specific intervals. If session controller has not received corresponding responses for a given time from one user, say A, then A will be considered has broken away from the conference and session controller will instruct stream server to stop sending conference media to A and informs other participants that A has left the conference.

For those terminals that do not support INFO method, we estimate session states relative to them by statistic RTCP packet. In a given time, if stream server has not received any RTCP packets from a user, stream server will notify the session controller to stop the singling relationship with that user and session tears down.

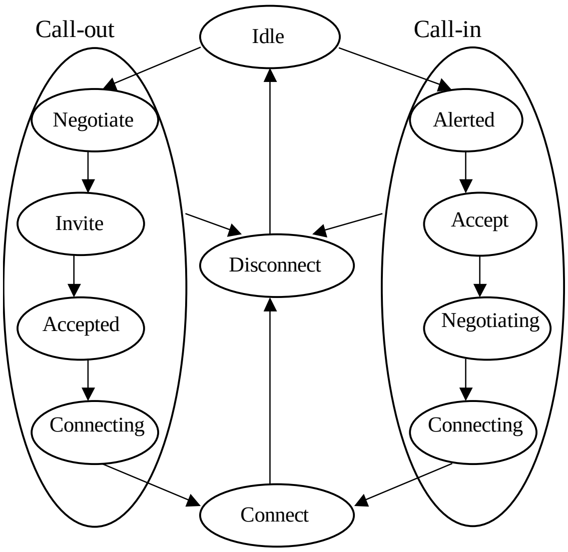


Figure 3: Session state transforming

Resource Management (Floor Control)

Conference applications often have shared resources such as the right to talk, access to a limited-bandwidth video channel, a pointer in a shared application, access to shared files etc. Floor control enables applications or users to gain safe and mutually exclusive or non-exclusive access to the shared object or resource [9].

In a given conference, there may numbers of floors, such as video floor, audio floor and whiteboard. Now, we have only defined audio and video floors. For each floor, we set two associated queues. The holder queue keeps the information or pointers for current floor holders and the pending queue keeps the pending floor requests. These queues are managed with a pre-defined queuing policy, or manually intervened by the moderator. The moderator, often the lecturer, can re-order, add and remove requests in a queue. For example, we define that only four participants can hold audio/video floor at the same time. When a participant make a floor claim, the floor claim will be put into pending queue and the floor moderator will be altered to make a decision whether to accept the request or not. If the number of floor holders is already four and the claim will expire before the time any current floor holder giving up the floor, the moderator will have to further decide whether or not to revoke the input right of one floor holder. If the moderator accepts the request, the claimer will hold the floor and his claim will be transferred from pending queue to holder queue, or else, the claim will be removed from pending queue and the floor claimer have to make the claim again later. In order to deal with the situation in which the moderator was unexpected disconnected from the conference, we also define a priority based first come first served automated queuing policy as an alternative.

To facilitate floor control operations, such as claiming a floor or deleting claims from a queue, we introduce floor control primitives into our system. Detailed floor control primitives can be found in [10]. We adopted its suggestion and only use the subset, i.e. ClaimFloor for claiming a floor, ReleaseFloor for giving up a floor, ChangeConfig for changing parameters of a floor, GrantFloor for accepting a floor claim, RevokeFloor for calling back a granted floor from a current floor holder, RemoveClaims for deleting floor claims from a queue, ReorderClaims for changing the order of claims in a queue. These primitives are exchanged among the conference participant, the conference server and the moderator.

4. EXPERIMENTS AND RESULT DISCUSSIONS

In order to validate the proposed architecture, we have measured its performance in terms of service set-up delay. Call setup times are defined in [11]. Now we show call setup phases in Figure 4 and divide it into two parts:

Post Dialing Delay (PDD) is also called Post selection delay. This is the time elapsed between the caller begin to send Invite request to the callee, and the time the caller hears his terminal ringing.

Connection Delay (CD) is the time elapsed from the sending of an Invite request up to the reception of the corresponding OK response.

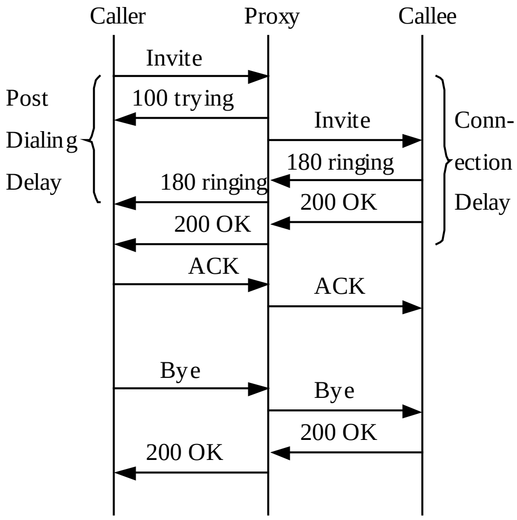


Figure 4: Definition for SIP call delays

In our testing platform, there are four AMD Athlon 40G PCs used respectively as Web server, Focus, Stream server and Proxy which is implemented with an integrated location database and acts in both proxy and registrar modes. The SIP terminals used for testing are soft IP phones, which include the same SIP User Agent and record setup delays. These testing soft phones are audio and video capable. The audio codecs used were G.711 µ -Law,G.711 A-Law and G.723.1. The video codecs used were H.261 and H.263.

Our experiments measured PDD and CD under LAN, WLAN and Internet conditions respectively. In LAN, the transfer rate is 100M and for WLAN (802.11) is 10M. The Internet access rate is 1M. From all scenarios, we use UDP for transport protocol of SIP message.

To get average results of our experiments, we performed 20 calls for each testing case and marked the experiment results in Figure 5.

As shown in Figure 5, PDD for LAN access is about 1.5 seconds and CD is 3.5 seconds, while PDD for WLAN access is 16 seconds and CD is 28.5 seconds. This distinct difference mainly stems from two facts. The first one is that the transfer rate in LAN is much higher than in WLAN. Another is that the wireless access is shared and the radio condition is sensitive, whereas the LAN access is dedicated (switched LAN). Several times, we also notice fluctuations in the results concerning WLAN access. For example, we once obtained CD up to 60 seconds that may be caused by the instability of WLAN.

By comparison with LAN and WLAN access types, PDD and CD for Internet access are respectively 14 seconds and 23 seconds that are moderate. Comparing our results with E.721 ITU-T recommendation, the delays are not superior. However, in the case of e-learning applications, the results are still acceptable. Our past experiences show that users can tolerate such delays

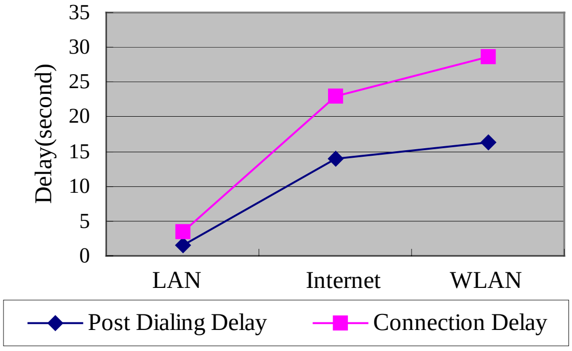


Figure 5: Delays for SIP calls

5. CONCLUSION

Focus is one of the most important components in a SIP-based conferencing system. It maintains SIP signaling relationships with all conference participants and ensures each participant receives the proper media streams. In this paper, we have studied the requirements of the Focus and presented a framework for Focus in the given e-learning system architecture. We also further studied its comprising parts concerning conference control. Finally, in order to validate our implementation, we measured its performance in terms of service establishment delays. The experiment results have shown that in LAN and Internet access conditions, the establishment delays are much less than in WLAN conditions. For an e-learning system, the delays are acceptable. However, if we are about to add some other services, such as authentication and Qos, the service establish delays may increase. We will investigate these topics in the near future.

6. REFERENCES

[1] J.Rosenberg, H.Schulzrinne, *et al.* “SIP: Session Initiation Protocol ”, RFC3261, June 2002.

[2] J. Rosenberg, “A Framework for Conferencing with the Session Initiation Protocol”, Internet draft, IETF, February 12, 2003.

[3] J.Roserberg, J.Peterson, et al. “Best Current Practices for Third Party Call Control in the Session Initiation Protocol”, Internet draft, IETF, Dec. 24, 2003.

[4] R.Sparks, “The Session Initiation Protocol (SIP) Refer Method”, RFC 3515, April 2003.

[5] A. B. Roach, “Session Initiation Protocol (SIP) – Specific Event Notification”, RFC3265, IETF, June

2002.

[6] M. Handley, V. Jacobson, et al. “SDP: Session Description Protocol”, RFC2327, April 1998.

[7] J. Rosenberg, H.Schulzrinne, “An Offer/Answer Model with the Session Description Protocol (SDP)”, RFC 3264, June 2002.

[8] S. Donovan, “The SIP INFO Method”, RFC2976, IETF, October 2000.

[9] Petri Koskeiainen, Henning Schulzrinne and Xiaotao Wu. “A SIP-based Conference Control Framework”, NOSSDAV’02, May 12-14, 2002, Miami Beach, Florida, USA.

[10] WU, X., ET AL. “Use SIP and SOAP for conference floor control”, Internet draft, IETF, Feb. 2002.

[11] T.Eyers, H.Schulzrinne, “Prediction Internet Telephony Call Setup Delay”, IPTel 2000(First IP Telephony Workshop), Berlin, April 2000.

Zeng Qingheng is a Master candidate at Wuhan university. She received a BS in computer science from Huazhong University of Science and Technology in 1999. Her research interests include VOIP，multimedia network communication, wireless network etc.

Hu Ruimin 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, etc.

Java Based Distributed Learning Platform

Zhang Xiaoming, Zhu Jinjun, Wang Jingyang, Qin Min, Zheng Guang
College of Information Science and Engineering, Hebei University of Science and Technology,
Shijiazhuang, Hebei, 050054, China
Email: zhangxiaom@hebust.edu.cn **Tel:** 0311-8613184

ABSTRACT

This paper introduces the principle and method of a java based distributed learning platform. Java played two roles in this project, on one hand, the learning platform is developed based on java, on the other hand the platform can provide a lot of example programs and demos in java language for some courses of computer science specialty. A distributed learning environment is offered for the students in computer science specialty，which can improve the teaching effect.

Keywords: Java，J2EE，Web，UML，Distributed learning platform

1. INTRODUCTION

During years of teaching practices, we find out that some of the concepts in professional curricula of computer science specialty are so abstract that can’t be understood easily. For example, the concepts of process, thread and synchronization in operating system course; the concepts of TCP/IP protocol stack in computer network course; the concepts of inheritance, polymorphism in object-oriented principle course. How can students understand these concepts conveniently and directly is a problem in teaching. An approach to resolve the problem is urgently demanded.

Java is one of the mainstream network programming languages which is object-oriented, Architecture-neutral, Distributed and Multithreaded etc. Much of the features of java are related with the concepts of professional courses in computer science specialty, so we can make use of these features to design some example programs to help the students, which make the abstract concepts instantiation so that to improve the teaching effect.

Based on this idea, we designed JBDLP system(Java Based Distributed Learning Platform). This article illustrates the principle, software architecture of the system and the realization technologies.

2. DESIGN PRINCIPLE

2.1 Goal of the System

The goal of the system is to realize a platform for students to learning conveniently and directly. In order to help the students understand abstract concepts directly, a Java example is designed for each concept; in order to make the students use the platform conveniently, the system is designed as Browser/Server architecture and is realized based on J2EE, which is architecture-neutral.

Java played two roles in this project. On one hand, the project is developed based on Java technology, on the other the example programs and demos in the system are written in Java language. We use Java as a developing tool to make the learning platform convenient to use and we also use Java as an expressing tool to describe the abstract concepts through example programs.

2.2 Organization of the Contents

The learning platform is a well-organized repository which is organized around knowledge point. Here, knowledge point means an abstract concept in a course. The content of the repository includes knowledge points, knowledge description, example description, example code in Java, related Java semantic, related references and related network resource, etc.

For a knowledge point of a course, for example the UDP of computer network course, we can find the theory description of UDP in the repository. And there are more than one Java examples which can show the protocol details are available for the students to browse and download, and both the comments and source code are given to the students. In order to make deeper impression, students can see the visual demos of the examples in their browser. Other assistance such as related Java semantics, related references and related network resource are also available for the students.

In order to develop the Java example programs for teachers, we should establish the associations of the knowledge in professional course of computer science specialty and corresponding Java technologies. So we create and maintain a table shown as Table 1. Through this mapping table, we can write java examples easily for specific knowledge point.

Table 1 Mapping table of knowledge point and Java

Course	knowledge	Java technology	memo
Operating System	MultiThreads	Thread Class，Runnable Interface	
	Synchronization	synchronized	
	……	……	
Computer Network	TCP	ServerSocket Class，Socket Class	
	UDP	DatagramSocket Class, DatagramPacket Class	
	FTP	URLConnection Class	
	……	……	
Object-Oriented	inheritance	extends	
	polymorphism	……	
……	……	……	……

2.3 Function Design

For the convenience of the learner, the platform should have functions as follow:

1) The students can browse knowledge points according to the category
2) The students can use the assistance provided by the platform. The assistance mainly includes the java example programs and visual demos.
3) The students can search for specific knowledge according to course name or knowledge key words.
4) The teacher can maintain the repository.
5) The students can put forward some questions for the specific knowledge point.
6) The teacher can answer the students' question

We can express the JBDLP system through a logical model shown as figure 1.

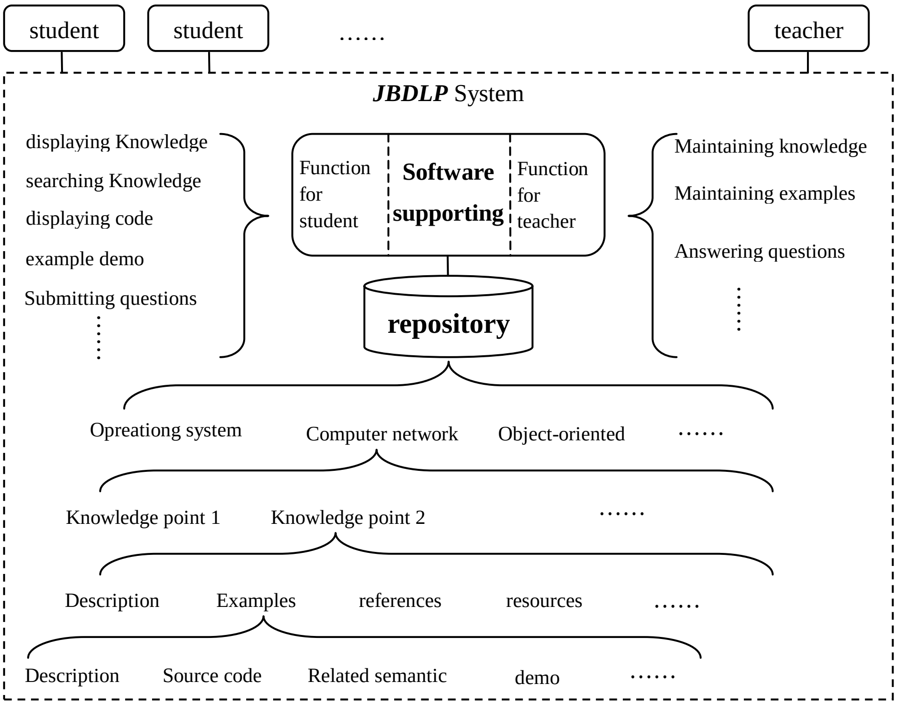


Fig.1 logical model of JBDLP

3. DESIGN OF SOFT ARCHITECTURE

The idea of RUP(Rational Unified Process) is used for developing the system and UML(Unified Modeling Language) is used as the Modeling language. From use case model view, analysis model view, design model view and Implementation model view, we described the software architecture of the system[1].

3.1 Use Case Model

According to the goal of the system, we made the use case analysis to get the actors and use cases. The use case model can be expressed by use case diagram shown as figure 2. For each use case, the description in detail is given through which we can see the interaction of actor and use case. The use case model captures the system requirements.

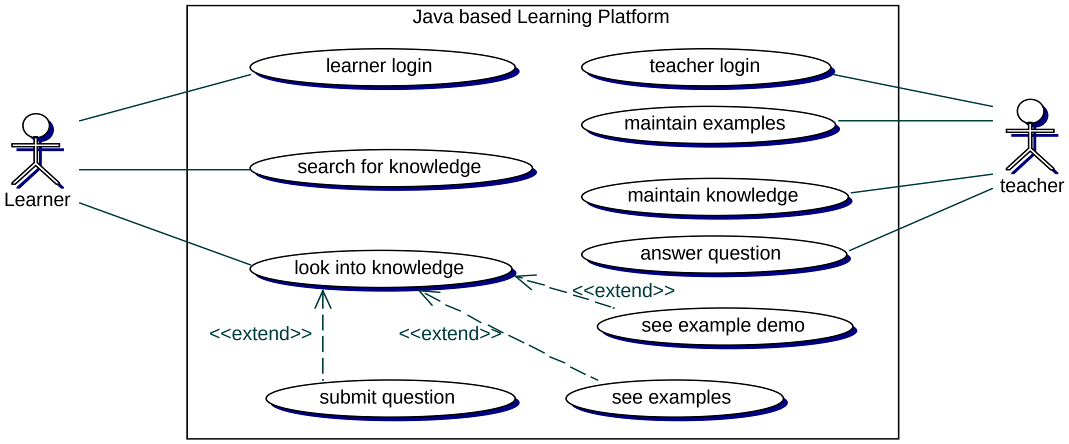


Fig. 2 Use case model

3.2 Analysis and Design

After we have got the use cases and use case descriptions, we can make further analysis to get the entity classes such as knowledge class, course class, example class, reference class, resource class, question class, learner class and teacher class etc. The entity class diagram is shown as figure 3.

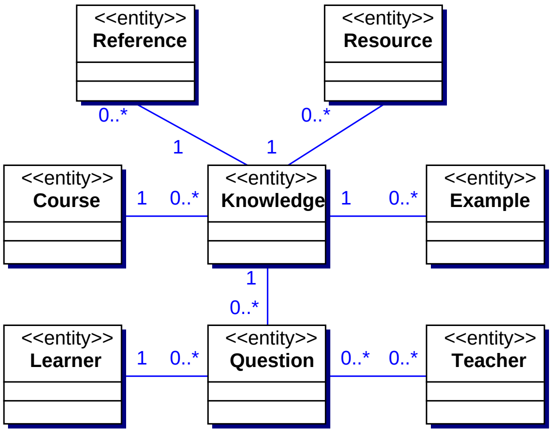


Fig. 3 Entity class diagram

To get the analysis model, we extract boundary class which handles communication between the system and its surroundings and control class which coordinates other objects. The relationships between these boundary classes, control classes and entity classes show the analysis model shown as figure 4.

In the transition from the analysis classes to the design classes, more details related to the target language and execution environment will be incorporated [2]. The design model serves as a high-level view of the source code. For the reason of the article's length, the design model is omitted here and we may discuss the model in other paper.

3.3 Implementation

The system architecture style is designed as layered architecture and MVC design pattern [3] is concerned. The MVC pattern is to separate the user interface from the control part and model part, which can reduce the coupling between the classes. The layered architecture can make the logical structure more clear to understand.

The system is divided into three tiers which are presentation tier, business tier and data tier. At each tier, we can select corresponding java technology from J2EE. At presentation tie, boundary classes will be implemented to show the user interface. Servlet and Applet can be used in this tie. Servlet can create dynamic web pages to the client side and send information to business tier. Applet is used to show the demos on the user's browser such as IE(Internet Explorer). At presentation tie, control classes and entity classes will be implemented as EJB(Enterprise Java Beans). The control part is realized as Session Beans and the entity part is realized as Entity Beans. At data tier, the data store and access services are provided. We use Oracle8i Database server to implement the data tier. [4] [5]

After we have implemented all the ties' programs, we must deploy them properly. The HTML web pages and Applet are deployed to the Web Server such as Apache. The Servlet application are archived as WAR type file and the EJB component are archived as EAR type file and these archives are deployed to the Application Server such as WebLogic. The designed data schema is imported to Data Server such as Oracle. The deploy model is shown as fig 5.

The whole system is developed by Borland JBuilder and we use Together for JBuilder as the modeling tool which is a simultaneous round-trip engineering tool and can keep sync between model and the code.

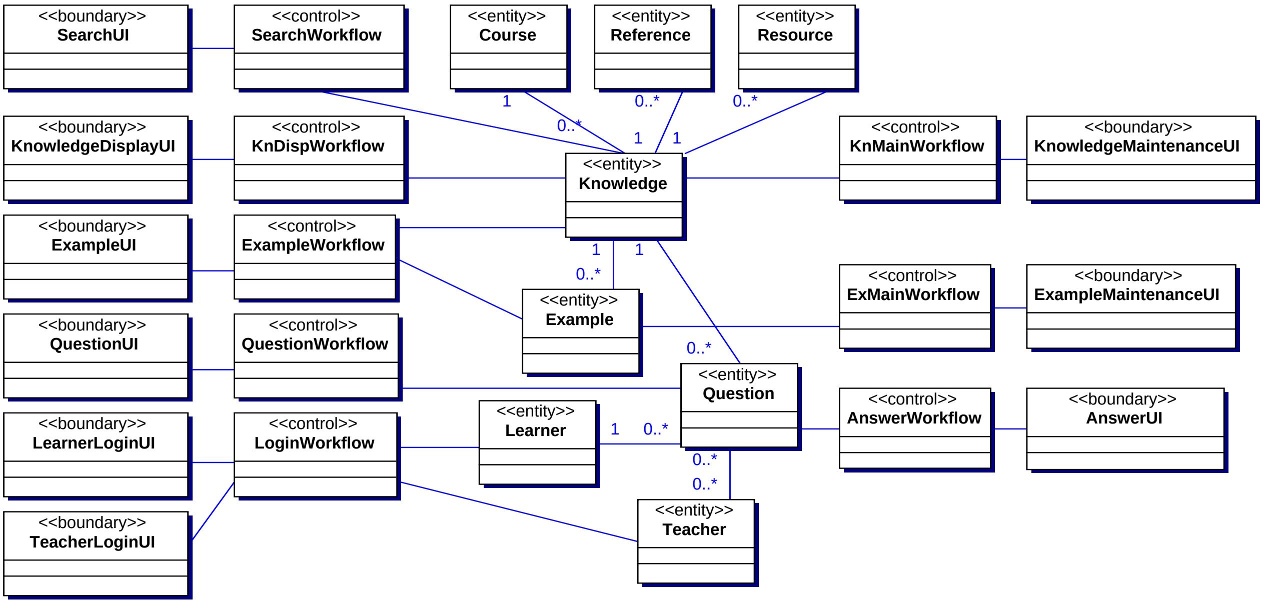

<<boundary>> SearchUI
<<control>> SearchWorkflow
<<entity>> Course
<<entity>> Reference
<<entity>> Resource
<<boundary>> KnowledgeDisplayUI
<<control>> KnDispWorkflow
<<entity>> Knowledge
<<control>> KnMainWorkflow
<<boundary>> KnowledgeMaintenanceUI
<<boundary>> ExampleUI
<<control>> ExampleWorkflow
<<entity>> Example
<<control>> ExMainWorkflow
<<boundary>> ExampleMaintenanceUI
<<boundary>> QuestionUI
<<control>> QuestionWorkflow
<<entity>> Question
<<control>> AnswerWorkflow
<<boundary>> AnswerUI
<<boundary>> LearnerLoginUI
<<control>> LoginWorkflow
<<entity>> Learner
<<entity>> Teacher
<<boundary>> TeacherLoginUI
1
0..*
0..*
0..*
1
1
1
1
0..*
0..*
1
0..*
0..*
0..*

Fig. 4 Analysis class model

4. EXTENDING TO DISTRIBUTED PLATFORM

Because Java technology is used to develop the system, we can install the system on different servers which have different operating system. Teachers in different laboratories are responsible for different professional courses, so the system can be installed on their own server is a good idea. Thus the system can run at a lot of servers which are at different labs and even at different universities so that a large distributed learning platform will be formed.

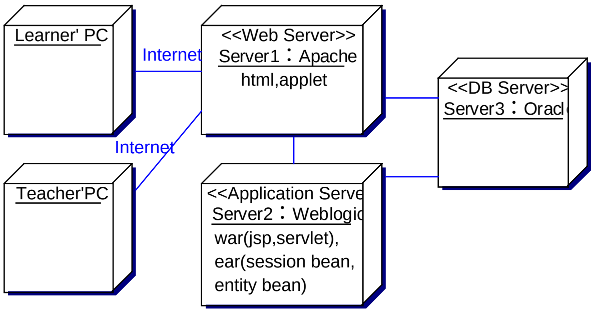


Fig. 5 Deployment model

5. SUMMARY

The Java based distributed learning platform faces the practical problems of students in computer science specialty. It's idea is that example programs written in Java language is used to help students understand the abstract concepts in books. The platform is a useful supplement to the teaching materials and courseware. Thanks to the Java technology, the system can run at different operating systems and becomes a distributed interactive platform for teachers and students. This platform can reduce the teacher's burden and arouse the students' interests to learn by themselves. Above all the teaching effect will be improved evidently.

6. REFERENCES

[1]. Jacobson I., Booch G., Rumbaugh J., The Unified Software Development Process, Addison-Wesley, 1999.

[2]. Jacobson I., Griss M., Jonsson P., Software Reuse: Architecture, process and organization for business success, Addison Wesley, 1997.

[3]. Gamma E., Helm R., Johnson R., Vlissides J., Design Pattens: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1995.

[4]. CT Arrington. Enterprise Java with UML. John Wiley & Sons, 2002.

[5]. Khawar Zaman Ahmed, Cary E.Umrysh. Developing Enterprise Java Applications with J2EE and UML. Addison-Wesley, 2001.

Zhang Xiaoming is a instructor of College of Information Science and Engineering, Hebei University of Science and Technology. He graduated from Hebei University of Science and Technology in 1997 with bachelor's degree; from HeBei University in 2002 with master's degree. He has attended DCABES2002. He has published two books, over 20 Journal papers. His research interests are in distributed information system, software reuse and software component technologies.

Zhu Jinjun is a Full Professor and a head of Network and Database Lab, dean of College of Information Science and Engineering, Hebei University of Science and Technology. He graduated from HeBei University in 1967 with specialty of wireless technology. He has published 4 books, over 30 Journal papers. His research interests are in network and database.

A Distributed Remote Education System Based on CSCW

Rui Hao, Chunyu Ran, Qi Shen
Computer Science, Wuhan University of Technology
Wuhan, Hubei (Province) 430070, China
Email: hao_rui1127@sina.com **Tel**.: (027)87297484

ABSTRACT

Nowadays, the remote education and on-line learning has become the major direction of current education development. For a distributed remote education system, how to separate the functions of its different participators, and what's more important, to realize the cooperative work of each member who takes part in it through network synchronously are the main problems it faces. Fortunately, the appearance of CSCW offers an opportunity to solve the problem. Considering the demands of current distributed remote education, the author makes classifications of different roles in the remote education and clarifies their individual tasks. In addition, based on the classification, a mode designed to reach the standard of CSCW is proposed and some correlative technologies about the CSCW are discussed.

Keywords: Distributed Remote Education, CSCW, Group Behavior, Electronic Whiteboard

1. INTRODUCTION

In this time with the development of computers and the popularization of the network, remote education and on-line learning has become an inevitable direction of the current education development [3]. Lots of countries in the world take it as one of the import means to keep their ability of competition and to improve the quality of their people, therefore pay much attention to the correlative studies and research. From 1995 in Birmingham to 2001 in Duesseldorf, the International Consortium of Discreet Education (ICDE) has hold four conferences in which some discussions about the view of quality, the view about the globalism, the strategy of development, the technical service and maintenance were made. Many effective measures were provided, which have greatly improved the development of the worldwide education career.

The main character of the current remote education is the real-time or unreal-time interchanges between the teachers and the students in different places. The distributing in space and the separate in time are the most significant differences between the current remote education and the traditional educational means, which are also the superiorities of the current remote education. But for a distributed current remote education system, the contents it refers are always involved in many aspects and the data it deals with is far more complicate than the text files. So, it requires a good support of the system for the cooperative work in the network. In this point, the on-line learning systems that designed in traditional ways have displayed their shortcomings in the support of the cooperative work in network, which in fact also has become an import problem that legs off the development of the remote education career. The appearance of the CSCW technology has brought us a hope for solving this problem. Taking the requirement of the distributed remote education into consideration, the article makes some research on the functions of CSCW in constructing the distributed remote education system.

2. CSCW TECHNOLOGY

The definition of CSCW (Computer Supported Cooperative Work) is an environment of distributed computers using the computer, the network and communication technology [5]. It also includes the media technology and the interface between the people and computers to actively organize lots of cooperative members who were separate in the time, distributed in the location but reliant to each other and their behaviors so as to accomplish a task together. Its aim is to design applications and systems supporting many kinds of cooperative work. It is made up of the understanding for group working and the co-operation to support that. In order to achieve this, there should be an environment and tools to support the cooperative work among people. The software and the system that is composed of them are called Group Ware.

2.1 The Main Contents of the CSCW Research

a) Cooperative management and the group sense

This is the most important part of CSCW research. In a CSCW system, the cooperative management is the center that takes control of the cooperation's start, initialization and execution. It covers from the shaping of the cooperation conclusions, the vote or integration about the conclusions to the end of the cooperation, etc. Though the CSCW system offers a virtual working space, which makes it possible to carry out the cooperative work of users separated in their locations, the interchange and the sense in that space are still much less convenient and effective than in the face-to-face working space. Therefore the research on the virtual technology in CSCW system plays a very important role in improving the efficiency.

b) Resolving the conflicts, parallel execution control and consistency.

In the CSCW system, there certainly exist the decision conflicts between the different roles and the problem of visit control caused by the access to the shared resources, which would inevitably lead to the problem of consistency.

c) Support for the dynamic variety of group roles and group constitute

Since in the CSCW system, the users participate in the cooperative work as some role. There should be the support for the group roles so as to effectively realize the right management in the course of cooperation among many users.

d) The capture and analysis of the interchange process between the role-system and the role-role

It is mainly concerned about how to capture and analyze the integrative information in a rapid, safe and effective way, and how to design a high-quality interface of users as well as a model to support the process of interchange between the role-system and the role-role.

2.2 The Classification of the CSCW
People in the information society and cooperative job have various characters. According to the definition of concept of the time and location, the group cooperative work could be classified into four models:

- Asynchronous model: the cooperative working in the same time and place to carry out a task, say, the collective decision and compiling, meetings and so on.
- Distributed asynchronous model: the cooperative working in same time but in different place to carry out the same task, say, the co-design, the co-compiling, group decision, video conference and so forth.
- Asynchronous model: the cooperative working in same place but in different time to carry out the same task, say, the operation in turn.
- Distributed asynchronous model: the cooperative working in different time and different place to carry out a task, like the e-mail, the development of large-scale program, etc.

In these four models, the distributed asynchronous model is the main object of the research.

Nowadays, the main problem lying in the researches and development of CSCW is the shortage of a perfect general architecture model that can support the cooperative system and the application of cooperative sense effectively. In one hand, this is because the number of fields, which the CSCW refers, is considerable, such as sociology, histology, psychology, computer science and so on. Even in the same field, the strength of the support needed by people is not the same. On the other hand, though CSCW is a distributed environment of computers, it still differentiates from the general distributed system. For instance, the distributed system seeks for the transparency in order to let the users feel they are occupying the system alone. But the CSCW systems is aimed at the opacity so as to notice each user the operation of other participators in the shared objects; the distributed system mainly concerns about how to exert the function of every parts of the system effectively, so that the efficiency of the whole system would be the highest while in the CSCW system, it pays more attention to the effective cooperation of every cooperation based on the seek for highest efficiency. So, just the general definition of distributed model is not enough. Many international or domestic institutes and colleges have done much work on this subject. Based on their results, the author would try to construct an architecture model of the current distributed remote education with the help of CSCW technology.

3. REMOTE DISTRIBUTED EDUCATION SYSTEM BASED ON THE CSCW

3.1 The Architecture of the System
Since there exist many shortcomings in the traditional two-layer structure based on the C/S model, such as occupying the system resource, low reliability, boring deployment in the client side, bad ability of migration and so on [3]. The system adopts the distributed application technology based on the J2EE architecture. It is a 3-layer application system built on the platform of Web. There are three layers in the system, the expression layer, the business logic layer and the database layer. The users in the expression layer connect to the business logic layer by the browser through which they visit the database with the help of Service Interface offered by the business logic layer. All the data imported is dealt with in the business layer and connected to the database layer by the DBAgent (the database visit agent). The data from the database would be sent to the expression layer through the Service Interface after be dealt with in the business layer. In the database layer, the database would not maintain a connection for each client any longer but would be connected through the public logic groupware shared by some clients. By doing this, the number of connections is reduced and therefore the performance and security of the data server would be improved a lot. The architecture of system is shown in figure 3-1:

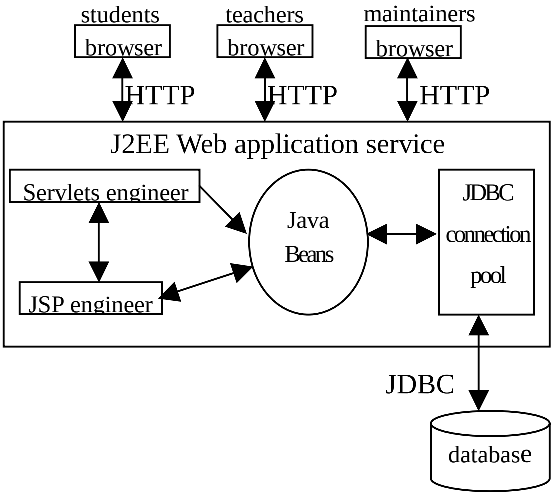


Figure 3-1 The Architecture of Remote Education System Based on J2EE

Comparing to the 2-layer structure, the most advantage of n-tier system structure based on the B/S model is the department of the logic of business from the expression of business. That could make the programmers more focus on the function design and the logic application, which obviously improves the efficiency of the programmers and greatly strengthens the ability of migration.

3.2 The Realization of System Logic Architecture
The remote education has the character of group behaviors. That determines it cannot gain the expected effects without the attendance and the cooperation of every part [2]. So, before we set off to develop the distributed remote education system with the CSCW technology, we have to make an analysis of the members who attend the remote education as a group, we hope to find out the character of tasks and its members and discover the relationship between each other. Then, by applying some relative computer support technology, we could build the logic model on the basis of building a proper teaching group model.

3.2.1 The classification of the group members
The first step of building teaching group model is deciding the group members, that is, the different roles taking part in the group behavior. As a remote education system, there need some person to study, to question and to discuss at first.

Then the person who compile the different courseware, make the teaching plan, answer the questions and offer the guide are needed. Besides, for the system, some correlative jobs, such as design, management and maintenance also need the attendance of members. Therefore, according to the differences in the task classification, we can group the members into three kinds:

1) The teachers: their duties are making the teaching plan, building the architecture of knowledge, compiling correlative courseware, answering the students' possible questions and deciding whether it needs to be discussed publicly, etc.
2) The students: their duties are studying the contents of courseware, putting forward some questions, participating in the discussion and so on.
3) The maintainers: their duties are designing and developing the whole system, maintaining and managing it in order to ensure the normal operation.

3.2.2 The tasks and levels of group members

The task for the group cooperation is accomplishing some specific work with the effort of every member in the group. Meanwhile, the task for the remote education group is through the cooperation between the teachers and the students, the teachers and the teachers, the students and the students, the teachers and the maintainers, to realize the compiling and execution of the courseware, the instruction of teachers to students and the students' understanding about knowledge. Here is the specific course: the teachers decide which courses should be available for the students on the basis of teaching needs and therefore work out the aim of courseware. Based on this, maintainers are organized to design and accomplish the CSCW system. The teachers classify the knowledge architecture according to the practical experience or students' demands, compile all kinds of courseware and exercises with the assistance of the maintainers, and add them to the question database and the courseware database. After registration, the students can choose certain courseware to carry out the on-line learning. If there is any question occurred in the course of study, he can offer it to the teachers in text formation. The teacher could answer the question in time, or he can do it later in other formations, say, the e-mail, the discussion and so forth, depending on their needs. The students could examine their learning effects through the exercises stored in the exercise database. So, the teachers, the students, the maintainers cooperate effectively and make the procession of remote education successfully.

There is certain layer character of the group tasks that is formed by the integration between every members and sub resource and there is a relationship of using and being used between them. Meanwhile, based on the different requirement of the granularity, each task could be classified into more detailed tasks.

3.2.3 The characters of group behavior

The characters of group behavior mean the rules for accomplishing the tasks. Any operation of a task has to obey certain rules that are different between each sub task. The collection of all sub tasks and relationships makes up the character of the group behavior. In this collection, any normal operation of a sub task must be under the premise of the former one's accomplishment, which means that the former has a restriction on the later one. At the same time, the group behaviors that are carried out appear different characters because of the difference in their aims, their members and their contents. For example, in the course of learning, the students are the protagonists and therefore have the right to choose different courseware. However, in the course of teaching, the protagonists have become the teachers who thereby have the right to choose the time of answering questions, to decide the contents of teaching and so forth. Therefore, it is very important to decide the sequence between each task in the course of building the group behavior model.

3.2.4 The classification of the system behavior layer and the construction of the cooperation model.

Taking the character of group behavior into consideration, we can classify the behavior layer and construct the corresponding cooperation model based on the results of former steps. It shows in figure 3-2:

3.3 The Construction of the System Logic Model

After constructing the corresponding cooperation model, we could set off to build the logic model of system. The behavior of remote education group is on the basis of CSCW and WWW. The contents of courseware, the contents of answers (including the questions asked by the students and the standard answers) and the contents of exercises are stored in different databases that can be connected through the Web Server. Besides, there also should be certain media control and CSCW control provided by the system to ensure the cooperative transmission of the video and audio in the network so that we can gain the effect of remote education.

The system is made up of following three sub systems:

- ◆ The classroom teaching system that focuses on the interchange between the teachers and the students. It is the extension and development of the traditional education system. The system provides the support for the media teaching, the show of electronic speech, the view of cartoons about the contents and so forth. Many functions are included in it, say, the learning of the courseware, the on-line order programming and answering the questions, etc.
- ◆ The cooperative learning system whose key point, different from the CAI that focuses on the individual guide, is mainly about the interchanges between the students and students. Its main functions contain the on-line discussion, the group experiments and so on.
- ◆ The administration system which protrudes the relationship between the teachers and teachers or the teachers and the maintainers. It supports the cooperation among many teachers and maintainers and contains many functions such as the maintenance and administration of the system, the compiling and modification of the courseware, and so forth.

All of the three sub systems are integrated with each other in logic. For example, the instruction of teachers to students in network makes use of the integration of cooperative study system to realize supplant and the extension of the classroom teaching system. In the cooperative learning system, the students send the questions occurred in their process of learning to the answer function in the formation of electronic documents. After receiving the questions, the function would search them in the answer database for the correlative answers. If the searches were successful, the answers would be sent back to the students directly, otherwise, the questions would be sent to the teachers. The teachers would answer the

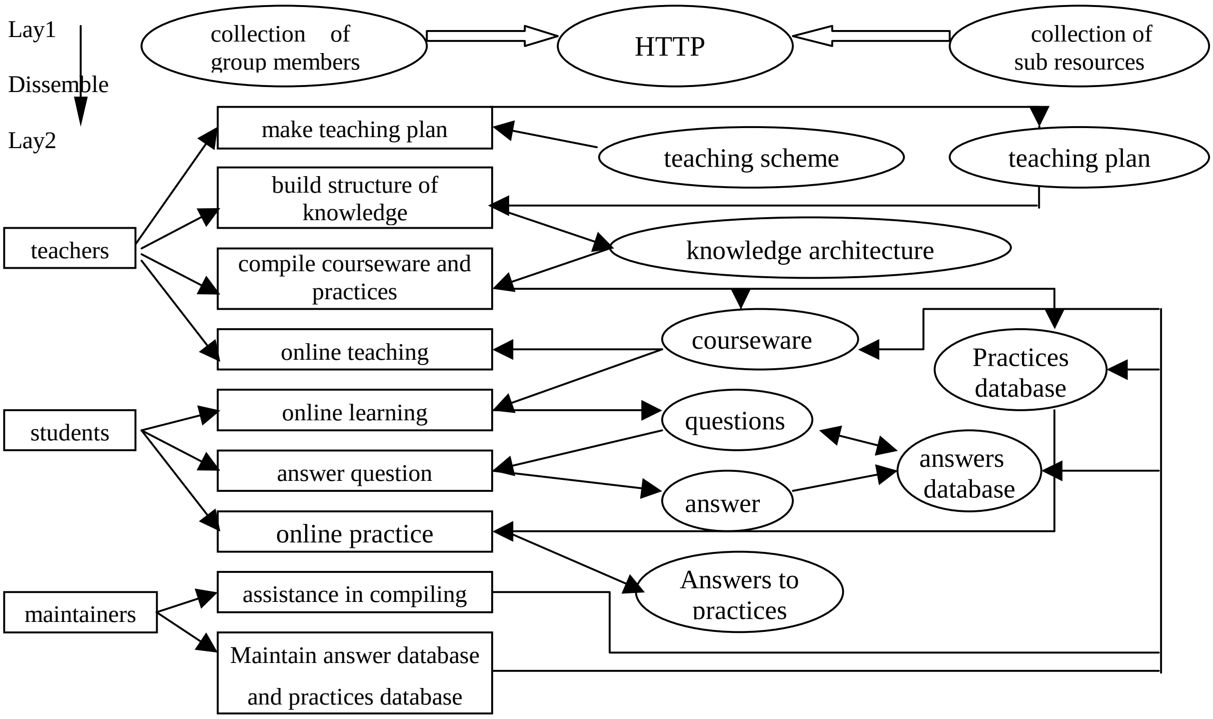


Figure 3-2 The Classification of System Behavior Layer and The Cooperation Mode

questions and deliver them to the discussion area for argument according to different conditions, and add the questions and their answers to the database. The technology of electronic whiteboard is involved in the whole course, which we will describe it later. Besides, since there are many aspects involved in the current media teaching in network, the cooperation of the teachers and the maintainers is in need. Meanwhile, for the purpose of improving the quality of the courseware, it is not enough to make teaching plan, to generalize the knowledge and to compile the contents of courseware depending on the effort of only one teacher. So it also needs the cooperation among many teachers and maintainers to accomplish the work of compiling. The logic model of system is shown in figure 3-3.

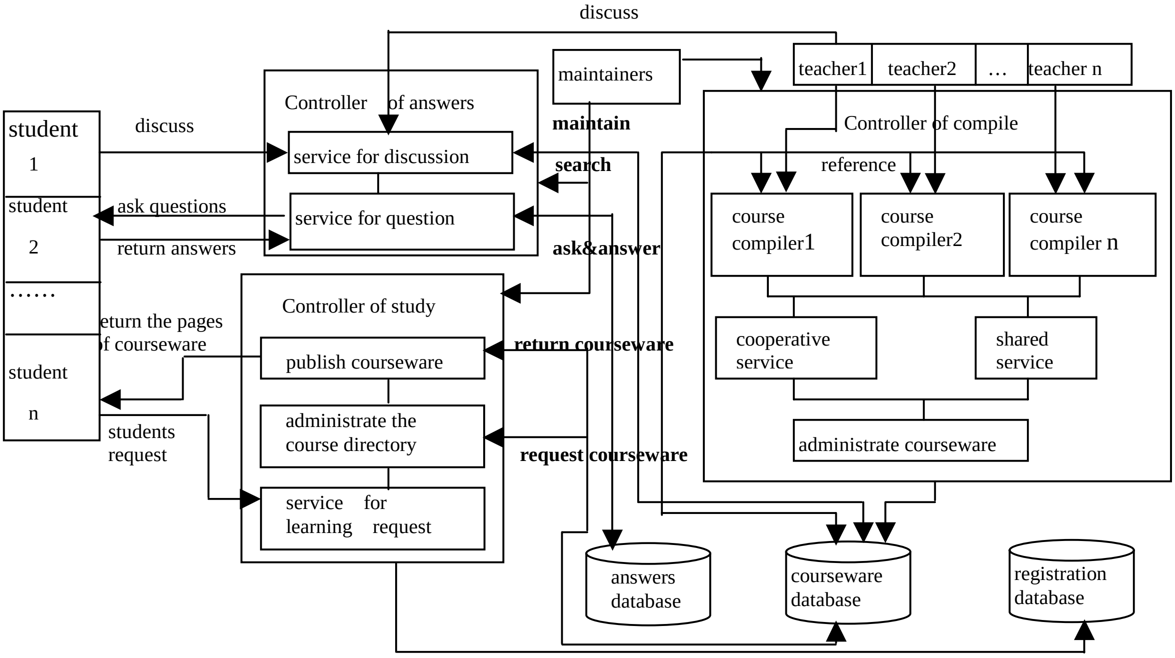


Figure 3-3 The Main Operation Principle of Remote Education System

3.4 The Technology of Electronic WhiteBoard.

In the area of discussion, the technology of electronic whiteboard is in need for the demand of real-time and publicity. The electronic whiteboard, which is similar to the BBS, is the most common area in the system that can be built and controlled by the teachers who take charge of the teaching system. In the condition of being admitted, lots of students could "paste" their individual views in a "public" area in the system, and the system would ensure the dynamic refresh of all students' computers. The students could ask the teachers' questions through the electronic whiteboard and the teachers could decide weather to give them a single guide or to do instruction in groups. By doing this, the transmission and communication between the teachers and students, the students and students could be executed.

The share whiteboard is an application protocol in interchanges between graphics and data based on the communication protocol of many dots in network [4]. An integrated model of master-servant mode and group broadcast mode is adopted in the commutative cooperation system. The master-servant mode is the main structure of the system. The server would send the news gained from the users of the whiteboard to other members of group by means of group broadcast. Therefore, the problems of synchronization, similarity and reliability in the transmission are the key factors for this distributed architecture. All the shared information in the system could be abstracted into the shared objects, which are the basis of cooperation of computers. It offers two kinds of service, one is the synchronization of the shared objects and the other is the sense of them. The synchronous service means how to get a consistency in many copies of objects in the condition of many users, and the sense service is the expression of one user's operation on the shared objects in the interfaces of other cooperative members. For the purpose of describing the contents of whiteboard news, the BBSObj class that describes the information of server whiteboard objects and the BBCObj class that means the information about the client whiteboard objects are designed separately. The explanation of BBSObj class lies in following:

```
   /* BBSObj. Java */
  public class BBSObj{
      public BBSObj ( int number) {
           this.bbNumber=number;
           }
……
private int number;          // the number of whiteboards;
private int count;           // the number of users who
                                    now is using the whiteboard;
private Date createtime;    // the time the whiteboard
                                    is built up;
private String message; // the records of whiteboard sessions;
private int lock;            // the state of current lock
……
           }
```

One time of operation in the shared objects is defined as an event. In one event, except for the dealing about the event inside the clients, the information of this event would be sent to the server that would broadcast it to other users according to the rules of cooperation so as to gain the synchronous renewal of the information. Therefore, the problem of cooperative control, which may otherwise lead to the noisy work environment in cooperative working, must be took into consideration in the course of user's operation in the objects. The cooperative control in the shared objects is consisted of the control of visit to the objects and the control of the parallel. The control of visit is examining the roles and rights of the users after they send out the requirements and then deciding whether the permits to visit are admitted. The control in parallel means the need for adopting the mechanic of parallel control to solve the possible conflicts caused by the operation of many users when they operate the shared space cooperatively in same time. In this system, the cooperative control is accomplished mainly through the method of lock, which is actually a strategy of controlling the medal. There is a lock like the medal in a circle net and the user could talk only when he gets the medal. Otherwise, he has to apply for it. Every shared object in the space would be arranged a unique ID number when it is established, which could be considered as the lock of operation in the shared objects. For any shared object, there is at most only one user who owns the corresponding lock in one moment. There are three states of users' operations on the shared objects: owning the lock, owning the privilege to the lock, not owning the lock. When applying for talk, the user has to offer the name and password for registration, which would be checked by the server. Then the server would add it to the waiting sequence and change the state of the user according to the algorithm of parallel control, such as changing the state from "owning the privilege to the lock" to "owning the lock". Meanwhile, there is an area, which would be examined by the server after receiving it, in the data report for storing the state of lock. If the user has the right, the data report would be transmitted to the arranged group address, but if not, the data report would be abandoned and a reject signal would be sent out. When the server accepting the requirement of giving up the request or the waiting time for receiving data has exceeded over T, it would transmits the right to the next user and a notice to him would be delivered. The users would be deleted from the sequence by the server while exiting the system.

4. CONCLUSION

The remote education is the inevitable direction of the future development of education career. It also plays an important role for all the countries in the continuous development. For a distributed remote education system, how to guarantee the cooperative work and real-time interchanges among the members is a key factor to influence the effect of the instruction. The article makes some research on the functions of CSCW in the remote education. Through the analysis and classification of the roles and tasks of members who take part in the procession of remote learning, an example model of distributed remote education system based on the CSCW is offered. It is on the basis of the mutual relationship between different members. And the technology of electronic whiteboard is adopted. The next step for the research is the further improvement of the system. A commutative function of on-line test would be added and the function of media teaching in this system would be improved gradually.

5. REFERENCES

[1] Ruth Geer, Wing Au. The Online Learning Community: Strategies, Problems and Issues[M]. In: Proceedings of

International Conference on Computers in Education (ICCE’02), 2002.

[2] Bentley. R, Rodden. T, Sawyer. P. Architectural Support for Cooperative Multi-user Interfaces[J]. Computer, 1994, 27(5): 37-46.

[3] Cheng Jianjun. Research and Design of Modern Distance Education System Based on J2EE[D]. WuHan: Computer Science, The Wuhan University of Technology, 2003, 5.

[4] [4]Wang Jun, Zhou Jingli, Yu Shengsheng. Design of Cooperative Work of Shared Whiteboard[J]. J.Huazhong Univ. of Sci. & Tech, Vol.29, No.4, 2001.

[5] [5]Yuan Zhongxiong, Wei Guoqiang. A Distance Instruction Model Based on CSCW[J]. J.Shanghai Institute of Electric Power, Vol.16, No.1, 2000.

Ran Chunyu is a Full Professor and a head of computer system institute in computer science and technology department, Wuhan University of Technology. He graduated from Beijing architectural material college in 1969 with specialty of automatic control and then studied in Najing University and Huazhong University for computer courses. He once went to the Rumanian International Training Center. He has undertaken some research items of the “ninth five-year national key science and technology project”. Now he is carrying on some researches of the “key scientific vertical research projects at provincial bureau level”. More than 30 articles of him have been published and 5 teaching materials are compiled with his attendance. His research areas are computer applications like network database, communication security, and so on.

Hao Rui is a graduate student of computer science and technology department, Wuhan university of technology. He studied in Wuhan University of Technology from 1998 to 2002 and then entered the graduate academe for further study. He is now researching on a project named “automatic system of intelligence buildings”, which is also a research item of the “key scientific vertical research projects at provincial bureau level”. His research areas are computer applications like network database and communication security.

Research and Design of Collaborative Learning System

Kaiyan, Wang **Guzi, Huang**
Computer Network Center, Medical College,Shantou University
Shantou, GuangDong 515031, China
Email: kywang@stu.edu.cn **Tel**: 0086-754-8900470-802

ABSTRACT

The use of Information and Communication Technologies in the education domain has been characterized by the need of providing flexible systems that are adaptable to particular learning situations. In this sense, Software Engineering (SE) has emerged as a software development paradigm suitable for obtaining reusable, flexible, and customizable distributed applications, which would provide great benefits to the e-Learning domain. Nevertheless, this SE-education relationship has not coped with the collaborative aspects and the pedagogic theories underlying the social constructivism that constitutes the basis for collaborative learning. This article describes the process undertaken by the authors when applying SE principles to the development of Computer-Supported Collaborative Learning (CSCL).it also shows a collection of design patterns for developing highly reusable learning objects(LOs).

Keywords: e learning, learning object, design pattern, software engineering, component framework

1 INTRODUCTION

The recent advances in multimedia technology such as the high-speed communication networks, large-capacity storage devices, digitized media, and data compression technologies have greatly changed the way learners communicate with their instructors and with each other, especially in distance education. With the innovation of new network and Internet infrastructures and the development of multimedia technology, the distances perceived by the learners have been virtually diminished and distance learning has become one of the most interesting new directions for education.

Computer-Supported Collaborative Learning (CSCL), partially derived from an evolution of Computer-Supported Cooperative Work (CSCW), is based on a new and strongly interdisciplinary paradigm of research and educational practice. Its main features include highlighting the importance of social interactions (collaboration) as an essential element of learning, the preference for an interpretative approach to the evaluation of the learning process, as well as the role of participative analysis and design of the whole community when creating new technological environments. On the other hand, CSCL has been based on distributed systems technologies in order to support some of its main characteristics, i.e.: communication, collaboration, and coordination.

Several of the above mentioned elements have been embedded in the main commercial products or innovative proposals of e-Learning, although in a marginal way. For example, generic tools that promote collaboration have been introduced without a precise objective and environment, the importance of designing activities and associated workflows have been modeled through standards such as IEEE LTSC (Institute of Electrical & Electronics Engineers – Learning Technology Standards Committee, <http://ltsc.ieee.org>). However, the main stream within e-Learning is still centered on the concepts of knowledge transmission as the basic educational paradigm, and the new proposals are dominated by the immediate application of the new technological 'affordances' and the expected market benefits. Therefore, it is still necessary to advance in order to analyze and embed all pedagogical and technological elements that define CSCL.

2 DESIGN METHODOLOGIES

2.1 SE and Education: The Necessity of Reuse and Adaptation

Educational software in general has been traditionally exposed to the necessity of adaptation and personalization. Such requirements have been expressed by educators who need to use the software in different educational and social contexts, or even, with different pedagogical styles. Thus, too many specific applications have been developed in order to meet the above requirements.

Due to the fact that these applications are usually monolithic, dependant on particular technologies and incompatible among them, teachers usually face great difficulties in order to integrate them in the classroom. These projects present a high failure rate since they are not able to get adapted to new educational situations and to incorporate technological innovations that are continuously emerging.

Software component technology offers the promise of composing tools from elements that may come from different providers. Therefore, it is a reasonable candidate as a potential solution of the educational domain, since it provides the capacity of application reuse and adaptation. When dealing with the problem of reuse in Software Engineering and particularly in component framework, it is essential to take into account the concept of component framework: an extensible set of reusable software components in a particular application domain together with a number of software design patterns that document their use. Components included in a framework can be reused, instantiated and assembled with additional components provided by developers in order to obtain concrete applications faster and with a lower cost.

SE has been employed in several projects in which the idea of component framework has been successfully applied in developing educational applications. However, the issue of supporting collaboration, inherent to the particular CSCL domain, has not been taken into account.

2.2 Objectives

Actually, the objective of obtaining a software component framework for CSCL guided the work of the authors during the last years, within a multidisciplinary group formed by educators, as well as computer engineers.

Nevertheless, building a component framework is not an easy task. A framework developer must face different problems related to both the particularities of the framework domain and the technologies used to support the derived components.

One of the most important problems to take into account in this context is the identification and dimensioning of components. The fulfillment of this task largely depends on how the key concepts and principles of the domain of interest are understood by software developers. In the CSCL domain, this problem is particularly important due to the big separation among abstractions used by experts in collaborative learning (teachers, psychologists, other education practitioners,) and those used by software developers.

2.3 Approaches and Modeling

In this section we will introduce several patterns we used in CSCL design. A pattern is a reusable solution template that organizes proven solutions into categories, since they will probably recur in a different problem domain. Roughly speaking, a pattern is a solution for a given problem or a problem class. The problems emerging from the last section lead to some questions on how these patterns could be applied in order to achieve more dialectic, adaptive, knowledge building-driven pedagogical strategies that would be able to go beyond expositive, static, purely instructional distance learning approaches. Next are a set of four distance learning specific patterns that are being proposed as efficient solutions.

First, *Fine Granularity* pattern is meant to solve the problem of developing high-reusable learning objects. This pattern states that learning objects must be fine granular enough to represent an atomic piece of information. Otherwise, they can only be containers for other learning objects. Figure 1 shows an UML class diagram for this pattern

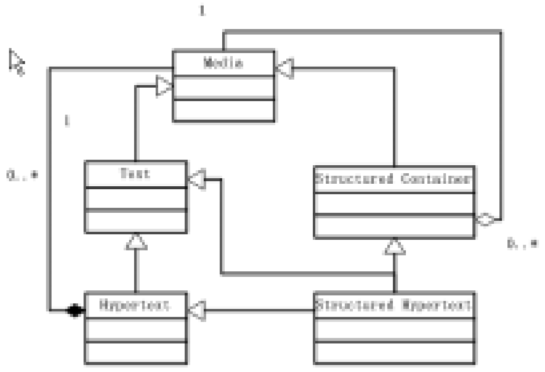


Figure 1 Fine Granularity pattern

In spite of the fact that it is not a completely skeptical approach, since it begins with small, discrete pieces of information, according to criticisms to it, new learning objects are proposed to students, which leads to adaptiveness problems. The next problem is: how to select and present learning objects to students that are adequate to their specific needs? This problem can be solved by adopting a second pattern, named *Learner Adaptiveness*. This pattern can be described by the following general rule: all learning objects must be created and structured so as to be adapted to students' actual learning requirements. Figure 2 shows an UML class diagram for this pattern.

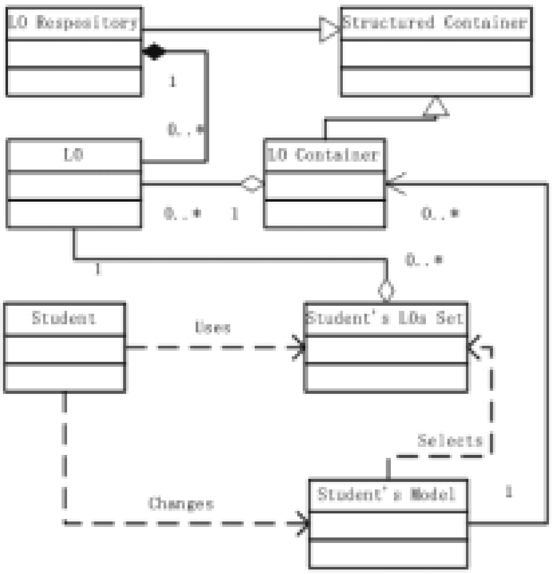


Figure 2 Learner Adaptiveness pattern

The next pattern is about how conversational strategies could replace expositive ones in distance learning courses. A possible solution for this is the *Agent-Agent Interaction* pattern, which states that interaction among students and teachers – named learning agents- can allow learning objects to be shared among agents, thus decentralizing learning objects' managing responsibilities. Figure 3 shows an UML Class Diagram which depicts such pattern.

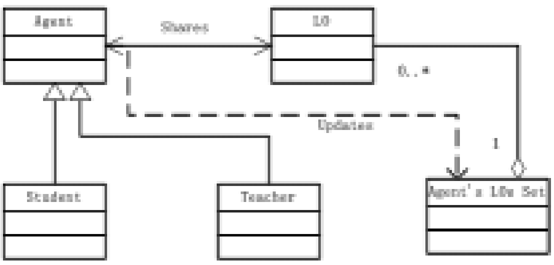


Figure 3 Agent-agent interaction pattern

The last problem to be discussed is as follows: how to ensure that students are able to deduce new concepts from previous ones? A possible solution to this is the *Hybrid Authoring* pattern, which allows them to build new learning objects through interaction among themselves, or among them and other learning objects. Figure 4 shows an UML class diagram that represents such pattern.

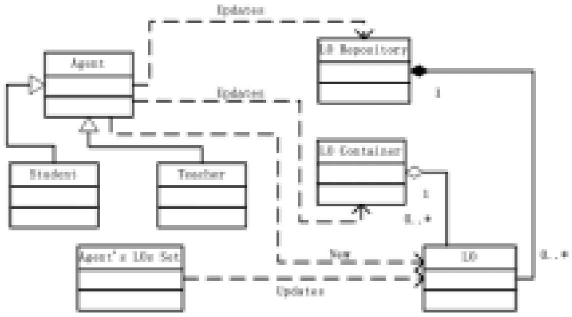


Figure 4 Hybrid authoring pattern

Distance learning environments implement this pattern by including the ability of share learning objects, which gets closer to Socratic dialectics if agents are also allowed to create their own learning objects.

Our distance learning environments implement these patterns by including the ability of present and share learning objects, also with the ability to allow students to create their own learning objects.

3 CONCLUSIONS

In this paper object-oriented software engineering applied to CSCL system is discussed, and a collection of design patterns are presented. Among the concepts and proposed schemes above, we adapted some principles and methods to the need of our projects. It's effective in designing component frameworks in CSCL.

4 REFERENCES

[1] A. Martnez, "Method and Model for the computational support to evaluation in CSCL", Thesis Doctoral. Universidad de Valladolid, 2003.

[2] Saini-Eidukat, B., Schwert, D. P., Slator, B. M., "Designing, Building, and Assessing a Virtual World for Science Education", In Proceedings of International Conference on Computers and Their Applications, 59-65, 1998.

[3] J. Arlow and I. Neustadt, UML and the Unified Process: Practical Object-Oriented Analysis and Design, Addison Wesley Professional, 2001.

[4] C. DiGiano, L., Yarnall, C. Patton, J. Roschelle D. Tatar, and M.Manley, "Collaboration design patterns: conceptual tools for planning for the wireless classroom", Proceedings of the IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'02), 2002.

[5] C. Larman, Applying UML and Patterns, Sebastopol, Prentice Hall PTR, 1997.

[6] J. Arlow and I. Neustadt. UML and the Unified Process: Practical Object-Oriented Analysis and Design, Addison Wesley Professional, 2001.

[7] J. Carey and B. Carlson. Lessons learned becoming a framework developer Software Practice and Experience, vol. 43, pp.789–800, 2002.

[8] Wiley. D., The Instructional Use of Learning Objects, Association for Educational Communications and Technology, pp. 571–577, 2001.

[9] D. W. Johnson and R. T. Johnson. Learning together and alone: cooperative, competitive and individualistic learning, Allyn and Bacon, 1999.

[10] C. Osuna Y. and Dimitriadis. "A framework for the development of educational collaborative applications based on social constructivism", Proceedings of the CYTED RITOS International Workshop on Groupware (CRIWG'99), 1999.

[11] J. Roschelle, J. Kaput, W. Stroup, and T. M. Kahn, "Scalable integration of educational software: exploring the promise of component architectures", Journal of Interactive Media in Education, vol.98, Oct. 1998.

[12] C. Szypersky. Component software beyond object-oriented programming, NY, USA: Addison Wesley, 1998.

Kaiyan Wang is a senior engineer in Shantou University Medical College. He graduated from Shanghai Jiaotong University in 1989, and received his master degree from TongJi Medical College, HuaZhong University of Science and Technology in 1994 with specialty of biomedical engineering. He has designed and developed several multimedia medical coursewares and learning sites, published 5 Journal papers. His research interests are in e-learning, medical courseware development and bioinformatics.

An Individual E-education System Based on Data Mining*

Zhang Xuemei, Zhang Xiaoming, Zhu Jinjun, Yang Kuihe
College of Information Science & Engineering, Hebei University of Science and Technology
Shijiazhuang Hebei 050054, China
Email: zxm@hebust.edu.cn **Tel.:** 0311-3910390

ABSTRACT

E-education supplies individual service for every student. This is the most superiority compare with the traditional education. This paper introduces an individual E-education system based on Data Warehouse and Data Mining. In this system, the data such as registering information, log files for accessing or communion content by network are loaded in Data Warehouse by the module of data collection and process. Then the data in Data Warehouse are analyzed or processed by the module of Data Mining and Analysis Online. So substantive information used directly to individual E-education education are obtained. The final aim is to control the output of education content, the choice of teaching mode and the choice of teaching flow.

Keywords: Data Warehouse, Data Mining, Individual E-education

1. INTRODUCTION

Along with opening of the education market and extending of age in education, E-education will have more and more superiority in competition. E-education supplies individual service for every student. This is the most superiority compare with the traditional education. At the same time, the students in E-education have more difference in mentality, age or knowledge background. This is a difficult problem that how to teach students in accordance of their aptitude, arrange the courses structure and make up a schedule in reason.

Data Warehouse and Data Mining are subjects across statistics, artificial intelligence, pattern recognition, parallel computing, machine learning and database technology. When the students access an education web site large numbers of data will be tracked such as registering information, log files for accessing, communion content by network and so on. Those data can be mined by using Data Mining technology to pick up the information concerning students. And subsequently the access behavior, access frequency and access content can be analyzed to get the cognition to object's colony behavior and fashion. So it can be used to improve the programming on web service side and reach the aim of individual service.

2. BUILD THE DATA WAREHOUSE FOR INDIVIDUAL E-EDUCATION

Data Warehouse is a data set that is facing topic, integration, correlative with time, and steady. It is different with the traditional database that Data Warehouse serves decision by high level[1]. It collects, organizes and stores the data from different geographical position and different constructed information source. At the same time, it gets data used to decision-make and analyze by processing those historical data. The basic frame of Data Warehouse is commonly that load data from different source to Data Warehouse by data collection and disposal tools, as in Figure 1.

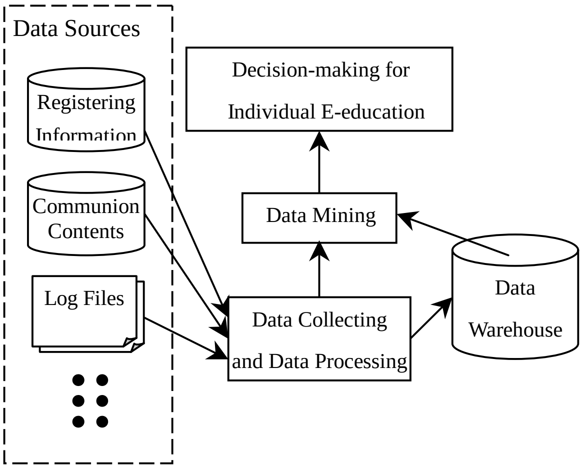


Figure 1 The frame of Data Warehouse

The modules in this frame include:
1) Data source: The data source include the log files for the students' accessing, registering information when they register as users to an education web site, communion contents which discussed by network, the question or answer between students and teachers, and so on.

2) Data collecting and data processing: This module takes out data to Data Warehouse after conversion and integration.

3) Data Warehouse: Two genus data are stored in Data Warehouse. One genus is metadata that is the basic structure cell in Data Warehouse. They are used to track record the data structures and the changes of Data Warehouse. The other genus is fact view that is used to analyze and process it by decision-maker.

4) Data Mining: It includes mostly the online analytical processing and the Data Mining for individual E-education.

3. USE THE DATA MINING BASED ON DATA WAREHOUSE IN INDIVIDUAL E-EDUCATION

The final aim of constructing Data Warehouse is picking up some rules that have important guidance to individual E-education. At the same time, because that all kinds of data are distributed from different data source the tools used to pick up valuable information from a mass of data are necessary. Data Mining is a tool to pick up valuable information. Its basic frame is showed in Figure 2.

* This paper was supported by the fund of Hebei Education Department (Grant NO. 2002120)

The modules in this frame include:

1) Data collection and processing: According to the aim of Data Mining, the correlative data set will be picked up from Data Warehouse. And their consistency and integrality will be examined.

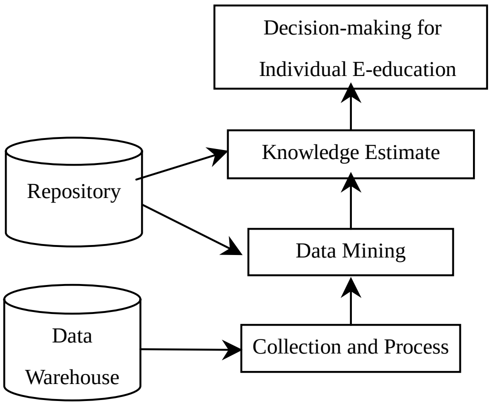


Figure 2 The frame of Data Mining

2) Repository: It is used mainly to Data Mining and knowledge estimate. The search operation in Data Mining can be guided by using the knowledge in repository. And the data as a result of Data Mining can be estimated.

3) Data Mining: The correlative data will be picked up from Data Warehouse to analysis or process such as clustering, estimation, assorting, prediction, relevancy and description.
Classific mining can be used to classify the students. So the individuation requirement of students is mastered and students are grouped rationally. The grouping students can discuss efficiently with each other.

Relevancy analysis and time sequence analysis can be used to group the web pages. So the course program, schedule and advice of learning methods for different student can be supplied to a perfect learning process for different.

4) Knowledge estimate. The knowledge helpful to individual E-education will be looked for or selected. And the data as a result will be expressed by a format like conception, rule, disciplinarian, restriction and so on.

The Data Mining based on Data Warehouse is a kind of deep-seated process of data in Data Warehouse. It also is a method or tool for realization of decision-making. Through abstracting from substantive Historical data in Data Warehouse the internal relation between each other can be obtained. At the same time, substantive information used directly to individual E-education can be obtained.

4. CONCLUSIONS

In the system of individual E-education base on Data Mining, firstly the data in different source are loaded in Data Warehouse by the module of data collection and processing. Secondly the data in Data Warehouse are analyzed or processed by the module of online analysis and Data Mining. Sequentially a series of information or knowledge used directly to individual E-education are obtained. The final aim is to control the output of education content, the choice of teaching mode and the choice of teaching flow. Data Mining is the kernel part in this system.

5. REFERENCES

[1] W.H.Inmon, Building the Data Warehouse Third Edition, John Wiley & Sons, New York, China Machine Press, Beijing, 2003.

[2] David J. Hand, Principles of Data Mining, MIT Press, China Machine Press, Beijing, 2003.

[3] Olivia Parr Rud, Data Mining Cookbook, John Wiley & Sons, New York, China Machine Press, Beijing, 2003.

[4] Si-chun Yang, “Analysis and Research on Data Mining Based on Data Warehouse”, Vol 13, No.9, Sep. 2003, pp. 86-88.

[5] Li-xin Zhang, “The Learning Environment In Network And The Learner”, China E-education, Vol 188, No. 9, Sep. 2002, pp. 28-30.

Zhang Xuemei is a lecturer in College of Information Science & Engineering, Hebei University of Science and Technology. She graduated from Heilongjiang Institute of mining in 1991 with a bachelor's degree. And she graduated from Jiaozuo Institute of technology in 1998 with a master's degree. She has published one book and over 10 Journal papers or academic conference papers. Her research interests are in database, data mining in network and e- education.

The WSE and its Application on the Encryption in the Remote Education System

Ran Chunyu, Bai Lin，Hao Rui
Computer Science, the Wuhan University of Technology
Wuhan, Hubei (Province) 430070, China
Email: Tiger_cloud@163.com **Tel.:** (027) 50105644

ABSTRACT

This article introduced the fountain and development of the WSE1.0 technology and its new implements about the security, routing, and Attachments. And some Web Service protocols which based on the GXA frame. At the end of this article, we introduce the new pulse of the WSE technology--WSE2.0.

Keywords: WSE, SOAP, WS-Security, WS-Routing, DIME, WS-Attachments

1. INTRODUCTION

To guarantee the currency on Web Service between different companies, the main XML Web providers (include Microsoft, IBM and Verisigns) put forward some new protocols. To support these new protocols, Microsoft released the Web Services Enhancements 1.0 for Microsoft.NET(WSE). The WSE is an all-new class library that is used for realizing the advanced Web service protocols, which based on the GXA frame(include the WS-Security, WS-Routing, DIME and WS-Attachments). The WSE is composed by some classes that implement these new protocols, and a group of filtering term that loaded by Microsoft ASP.NET. These filtering terms will prevent the input or output SOAP messages, and explain or create a SOAP flag for supporting the demands. The WSE provide a design model to the developers that are very close to the bottom. The filter pipe, which the WSE provided can handle every kind of complicated details of a message, then the developers can deal with the more complicated problem without suffered by the limits of those particular design model.

2. THE APPLICATIONS OF THE WSE

These are the characteristics of the WSE structure:

- The WS-Security proceeds to sign on the SOAP messages. The digital signatures are allowed to be signed before the change of SOAP messages.
- The WS-Security proceeds to encrypt to the SOAP message. The encryption creates an encrypted message, and only the owner of the private key can read the content of the messages. The WSE supports both the encryption and the decryption.
- The WS-Security proceeds to certificate the SOAP messages. The certificate of SOAP message is particularly efficient when the routings are more than one time.
- The WS-Routing uses WSE to route the SOAP messages. The sending of the SOAP messages uses the logic name, and it provides the transparent of the network topology.
- The WS-Attachments uses the DIME protocol to add attachments on the SOAP messages. Add attachments in the SOAP messages can make a file be sent directly without be serialize to the XML format.

The WSE is a kind of engine that implements the advanced Web service protocols to the SOAP messages. It requires writing the flags in the outputting SOAP messages and reading the flags from the inputting SOAP messages; it also requires converting the SOAP message's text. For example, encrypt the outputting message's text and decrypt the inputting message's text according to the WS-Security models. In the WSE, this function is realized by the filter. The outputting filter writes the flag to the message; the inputting filter reads the flag from the message and checks the availability of the flag. In addition, both the outputting filter and the inputting filter can convert the contents of the messages. Figure 1 explained the WSE filter model.

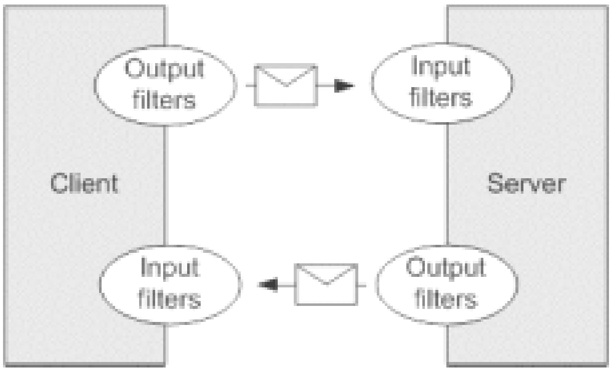


Figure 1 WSE filter model[9]

2.1 WSE application on XML safety

The Web service has been able to realize the safety on some degree. For instance, use the safety transmission such as SSL to guarantee the safety of XML Web service, but this method is only worked on point-to-point transmission. It means that the receiver must contact with the sender to verify the SOAP message again if the SOAP message must route more than one times and it used the SSL during the whole routing. One of the methods that the WSE help to build an expandable distribute applications is to send the safety certificates in SOAP messages by using the definitions in the manual of the WS-Security. It is realized by a safety certificate, which awarded by a client who is reliant to both the sender and the receiver. These safety certificates will be added to the SOAP message as the client sends the first SOAP request. It is no longer needed to put forward another requests to the client or another revising user to verify the integrality of the security tokens when the Web server receive the SOAP request. And this method is capable only if the origin of the safety certificate is credible. By way of the substitute, the code verify to the reliability of the certificate is finished before carrying out the XML Web service. [1] Without returning the source of the certificate to verify, it saved one or more networks request at least. And it improves the retractility of the applications markedly. The figure 2 explains the steps:

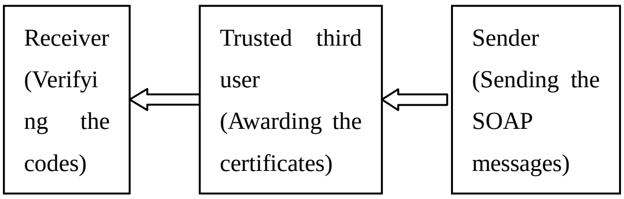


Figure 2

The digital signature and the encryption of the SOAP messages can make the SOAP XML Web service safer. The digital signature of the SOAP messages is to verify a SOAP message and find out whether it is changed after the digital signature by a SOAP message receiver. When the quantity of the SOAP message is one, it only permits the certain XML Web service to read the context of the massage to guarantee the safety of the XML Web service. [3]

2.2 The routing of the SOAP message

For designing a WSE distribute application to a application that realizes the network topological applying that transparent to the client, establishing a midst computer that configured to run the WSE router is necessary. The client sends the SOAP messages to the WSE router, the WSE router relegate the SOAP messages to a host computer with XML Web service.

One of the advantage of use the WSE router which installs the distribute application is be able to change the host of the XML Web service off-line without change the codes and configures on the client. The administrator on the host of the WSE router can make any needed changes to reorientate a SOAP message to another server. For implement this, the administrator need to prepare a backup server to provide the host service to the XML Web service, and the router still orientate the SOAP messages to the main server. Then the administrator should prepare a Web.config file to save the content in the cache and save the URL of the backup server to the new cache. A cache is a XML file that includes the ultimate address, which the router received. When the main server was displaced, the Web.config file and the cache are both displaced by the host of the new router. Then other SOAP messages will be dynamic routed to the backup server, all of this is unknown to the clients, they just still send messages to the router.

2.3 Use the SOAP messages to send the attachments.

The WSE supports the Direct Internet Message Encapsulation (DIME), this protocol defines a mechanism that can send the attachments in the SOAP messages. A XML Web service sends out a large text document or binary data is familiar, such as sending out a picture document in a SOAP message. The SOAP message could not transmit these large files according to the original intention of the designers, because it is used to transfer the simple XML texts. The files must be serialized to the XML format for adding these files to a SOAP message inside to these documents, but this method makes the new messages two times larger than the original files. The DIME protocol use the mechanism that encapsulate the entire content of the original files to the SOAP messages, and this mechanism can remove the needs of serializing the files to the XML format.

The DIME and the WS-Attachments constituted a simple and valid solution, which can pack the attachments and the SOAP messages together. They make it possible to include any kinds of data in the SOAP messages, so the JPEG image, digital signature file and many other DIME data pack can be sent by the SOAP message. They also defined a mechanism that can quote the additional data from the SOAP messages, which packed in the data packs.

3. THE ENCRYPTION OF SOAP MESSAGES IN THE REMOTE EDUCATION SYSTEM

This is an application that used on the registration of the users in the remote education system. The "Register" class represents the registered categories such as the teacher, student or administrator; the "Card" class represents the categories of the certificates, which are needed during the registration; the "CardNumber" represents the serial number of the certificates. There is the sensitive information in the "Register" class, so we must encrypt the SOAP message to guarantee the safety during the registration.

```
<soap:Envelope soap:xmlsn="http://....../soap-envelope">
  <soap:Header>
  ...
  </soap:Header>
  <soap:Body>
  ...
    <x:Order Type="Register" x:xmlns="http://....../order">
      <x:Register Type="Student">
        <x:Card Type="ID-Card">

<x:CardNumber>123456789123456</CardNumber>
          <x:ExperationDate>1108</ExperationDate>
        </x:Card>
      </x:Register>
      ...
    </x:Order>
    ...
  </soap:Body>
</soap:Envelope>
```

There is the sensitive information in the "Register" class, so we encrypted the SOAP message to guarantee the safety during the registration. The follow message includes the same information, but the "Register" class is substituted by an "EncryptedData" class that includes the encrypted information. The "EncryptedData" class is quoted by the "DataReference" class in the Security flag.

```
<soap:Envelope soap:xmlsn="http://....../soap-envelope"
  xmlns:xenc="http://....../xmlenc#"
  xmlns:xsig="http://....../xmldsig#"
  xmlns:wsse="http://....../secext">
  <soap:Header>
    <wsse:Security>
      <xenc:ReferenceList>
        <xenc:DataReference URI="#OrderID"/>
      </xenc:ReferenceList>
    </wsse:Security>   ...
  </soap:Header>
  <soap:Body>
```

```
   ...
      <x:Order Type="Register" x:xmlns="http://....../order">
         <xenc:EncryptedData Id="OrderId">
            <xenc:EncryptionMethod
               Algorithm= "http://....../xmlenc#tripledes-cbc"
            <xsig:KeyInfo>
               <xsig:KeyName>My Symmetric Key</xsig:KeyName>
            </xsig:KeyInfo>
            <xenc:CipherData>
               <xenc:CipherValue>...</CipherValue>
            </xenc:CipherData>
         </xenc:EncryptedData>
      ...
      </x:Order>
      ...
   </soap:Body>
</soap:Envelope>
```

4. THE EXPECTANCY OF THE WSE 2.0 TECHNIQUE

1) The WSE 2.0 technique is not supported or permitted to use on the production.
2) The WSE 2.0 technique is a new technique, which based on the new strategy frame, enhanced safe mode, message oriented design model, and it has the function that supports the safety, routing and attachment in the multi-host environments. [7]
3) The WSE 2.0 technique simplifies the coding by applying the safety strategies that can be run on the Microsoft .NET frame by the developers and the administrators. [4]
4) The Web services communication can use the Kerberos tickets, X.509 certificate, user name/password credence and other usual security tokens that based on the binary and the XML. [2]
5) The safe mode of the WSE 2.0 provides the strategy droved basis to the Web services for traverse the trusted region. A trusted service can establish the security tokens of search and affirming, or establish some grouping safe dialog and make the authentications of the same dialog faster.
6) Using the WSE Web services can be run in many circumstances such as ASP.NET, independent and executable applications and NT Services etc. And it can use many different protocols to communicate such as the HTTP and TCP.

5. CONCLUSIONS

As a new technique, the WSE technique has broad future. It's an important chessman to the Microsoft. Using the WSE technique can improve the software's interaction, efficiency and safeties markedly. And it will be more popular after the advance of the WSE 2.0.

6. REFERENCES

[1] Sending Files, Attachments, and SOAP Messages Via Direct Internet Message Encapsulation Jeannine Hall Gailey

[2] What to expect from Web Services Enhancements Tim Landgrave, Builder.com | 27 February 2003

[3] WS-SECURITY INTEROPERABILITY TEST RESULTS Author: Jonathan Stephenson

[4] Web Services Security in The .NET Framework By Mansoor Ahmed Siddiqui

[5] Encrypting SOAP Messages Using Web Services Enhancements Jeannine Hall Gailey Web Services Enhancements 1.0

[6] Microsoft Delivers Latest Developer Tools for Building Advanced WestGlobal and WRQ Redmond, WA, USA. December 16, 2002.

[7] Microsoft Previews Web Services Enhancements 2.0 By Clint Boulton July 15, 2003

[8] The SOAP Header expanding: WS-Routing and WS-Referral Xiaolu Chai (in Chinese)

[9] MSDN documents Microsoft Company

Ran Chunyu is a Full Professor and a head of computer system institute in computer science and technology department, Wuhan University of Technology. He graduated from Beijing architectural material college in 1969 with specialty of automatic control in engineering department. He once studied in Najing University and Huazhong University for computer courses, and then went to the Rumanian International Training Center for further study. He has undertaken some research items of "ninth five-year national key science and technology project". Now, he is carrying on some major researches of "key scientific research projects at provincial bureau level". More than 30 articles of him have been published in native or foreign journals and 5 teaching materials are compiled with his attendance. His research areas are computer applications like network database, communication security, graphics and images, and so on.

Solution of the Wigner-Poisson Equations for RTDs

M. S. Lasater[1], C. T. Kelley[1], A. G. Salinger[2], D. L. Woolard[3], and P. Zhao[4]
[1]Center for Research in Scientific Computation and Department of Mathematics
[4]Electrical and Computer Engineering Department
North Carolina State University, Raleigh, North Carolina, 27695-8205, USA
[2]Sandia National Laboratories P.O. Box 5800, MS-1111
Albuquerque, New Mexico, 87185, USA *
[3]U. S. Army Research Office
U. S. Army Research Laboratory, RTP, North Carolina, 27709-2211, USA
Email: tim_kelley@ncsu.edu **Tel.:** 1-919-515-7163

ABSTRACT

We will discuss a parametric study of the solution of the Wigner-Poisson equations for resonant tunneling diodes. These structures exhibit self-sustaining oscillations in certain operating regimes. We show numerically that the phenomenon corresponds to a Hopf bifurcation, using the bias across the device as a continuation parameter. We will describe the engineering consequences of our study and how it is a significant advance from some previous work, which used much coarser grids. We use the LOCA package from Sandia National Laboratory. This package, and the underlying NOX and Trilinos software, enables effective parallelization. We report on the scalability of our implementation.

Keywords: Wigner-Poisson Equations, Resonant Tunneling Diode, Hopf Bifurcation, Continuation.

1. INTRODUCTION

Semiconductor technology has developed to the point where the next generation of electronic devices will operate at the atomic level. Since the device scale is so small, design problems arise immediately. Currently, we do not have the technology to observe and collect all relevant data from such small devices. Furthermore, even if we had this capability, the device physics are determined by quantum mechanics and not by classical electromagnetism. A fundamental result of quantum mechanics is that the act of observing a quantum system will have an impact on the results we obtain. Thus, physically measuring how a normal quantum system is functioning would require an account of the effects of the observer on the reported data. To avoid this issue, engineers and physicists researching these quantum devices are working to develop an accurate model of these quantum systems from first-principle physics. One particular nanostructure we are interested in is the resonant tunneling diode (RTD).

A RTD is created by taking a slab of semiconductor and placing a second kind of semiconductor (one that has a larger band-gap) into this semiconductor. Since the second type has a larger band-gap, this effective creates potential barriers within the structure. Figure 1 is a diagram of an RTD.

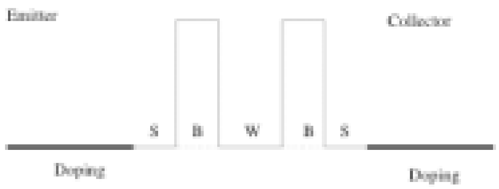


Figure 1 Diagram of RTD

The second type of semiconductor is represented by the dotted lines in the diagram. The potential barriers are also shown in the diagram and are represented by (B). Between the two barriers is a section of the original semiconductor. This is the quantum well (W) that is contained between the two barriers. Far from the barriers, the original semiconductor is doped (represented by the darker lines). Doping is where atoms that contain more (or less) electrons that the semiconductor itself are embedded into the semiconductor to create (or take away) extra electrons in the structure. Between the barriers and the doped regions are areas where the original semiconductor exists. These areas are called spacers (S).

Classically, if a particle runs into a potential barrier and it does not have enough speed, it will be reflected back. Since quantum mechanics treats electrons as waves instead of particles, an electron at any speed that encounters a barrier still has some probability of passing through the barrier. This effect is known as "quantum tunneling" and is the basis of this device. If a voltage difference is applied across the device, electrons will start to move along the device, tunnel through the barriers, and reach the other side, thus creating a current

Numerical simulations [1], [2] have shown that current oscillation can be expected for certain voltage differences, and that these current oscillations have a high frequency in the terahertz (THz) regime. With these numerical simulations, engineers and physicists are hoping to understand what physical mechanism creates these intrinsic oscillations and determine what physical parameters (i.e. doping profile, barrier height and width, well width, etc.) are conducive to sustaining and controlling these oscillations in hopes of producing a viable high frequency power source. This work is attempt to create a faster and more accurate RTD simulator to aid the engineers in these goals.

*Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000

2. MODEL DESCRIPTION

The model used to describe the electron transport in these devices is the Wigner-Poisson equations [3]. These equations consist of a nonlinear PDE that describes the time-evolution of the distribution of the electrons in the device coupled with Poisson's equation which incorporates the potential effects of the electrons into the model. The first of these equations can be given by

$$\frac{\partial f}{\partial t} = W(f) = K(f) + P(f) + S(f) \qquad (1)$$

Here, *f=f(x,k,t),* is the distribution of the electrons. It is a function of the position of the electron, *x*, the momentum of the electron, *k*, and time, *t*. The position variable *x* ranges from 0 to *L*, the length of the device, and the momentum variable *k* ranges from -∞ to ∞. The time-derivative of *f* is comprised of three terms. The first term, *K(f)*, represents the kinetic energy effects on the distribution and is given by

$$K(f) = \frac{-hk}{2\pi m^*}\frac{\partial f}{\partial x} \qquad (2)$$

Here, *h* is Planck's constant and *m** is the effective mass of the electron. The second term, *P(f)*, is the nonlinear term in the equation and is for the potential energy effects on the distribution

$$P(f) = \int_{-\infty}^{\infty} f(x,k')T(x,k-k')dk' \qquad (3)$$

where *T(x,z)* is given by

$$T(x,z) = 4\int_{0}^{L_c/2}[U(x+y) - U(x-y)]\sin(2yz)dy \qquad (4)$$

In this equation, *U(x)* is the electric potential as a function of position, and L_c is the coherence length. *P(f)* is nonlinear in *f* since *U(x)* depends on *f* through Poisson's equation. The final part of the time-derivative describes the scattering processes that occur between the electrons

$$S(f) = \frac{1}{\tau}\left[\frac{\int_{-\infty}^{\infty} f(x,k')dk'}{\int_{-\infty}^{\infty} f_0(x,k')dk'} f_0(x,k) - f(x,k)\right] \qquad (5)$$

Here, τ is the relaxation time, and $f_0(x,k)$ is the equilibrium Wigner distribution. This is the steady state solution to Eq. (1) when there is no voltage difference across the device. The boundary conditions for *f* impose the incoming electron distributions. That is, at *x*=0 and for *k* > 0 (electrons with positive momentum that are moving right) we have

$$f(0,k) = \frac{4\pi m^* k_B T}{h^2}\ln\{1 + \exp[\frac{-1}{k_B T}(\frac{h^2k^2}{8\pi^2 m^*} - \mu_0)]\}$$

and at *x*=*L* and for *k* < 0 (electrons with negative momentum that are moving left) we have

$$f(L,k) = \frac{4\pi m^* k_B T}{h^2}\ln\{1 + \exp[\frac{-1}{k_B T}(\frac{h^2k^2}{8\pi^2 m^*} - \mu_L)]\}$$

k_B is Boltzmann's constant, *T* is the temperature, µμ_0 is the Fermi energy at *x*=0, and µμ_L is the Fermi energy at *x*=*L*.

The electric potential *U(x)* is made up of two parts. The first part is from the electrostatic potential created by the electrons in the device. We will denote this part by *u(x)*. The second part is from the potential barriers in the device created from the heterojunction of the two different semiconductor materials. We will denote this part by Δ*(x)*. To get *u(x)*, we must solve Poisson's equation

$$\frac{d^2u}{dx^2} = \frac{q^2}{\varepsilon}[N_d(x) - \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x,k')dk'] \qquad (6)$$

q is the charge of the electron, ε is the dielectric constant, and $N_d(x)$ is the doping profile. The boundary conditions for Poisson's equation are where the voltage difference across the device is incorporated. We have that

$$u(0) = 0, u(L) = -v \qquad (7)$$

where $v \geq 0$ is the applied voltage. Once we have solved for *u(x)*, we have *U(x)* = *u(x)* + Δ*(x)*.

3. DISCRETIZATION

To numerically solve for the distribution, we discretize both the domain and equations using a finite difference method. For the *x*-domain, we use N_x grid points where $x_i = (i - 1)*\Delta\delta x$, $i = 1,2,\ldots, N_x$ and $\Delta\delta x = L/(N_x - 1)$. These grid points are evenly spaced across [0, *L*]. For the *k*-domain, we first truncate from -∞ to ∞ to $-K_M$ to K_M, where K_M is a maximum momentum we consider. We use N_k grid points where $k_j = (2*j - N_k - 1)*\Delta\delta k/2$, $j = 1,2,\ldots, N_k$ and $\Delta\delta k = 2*K_M/N_k$. These grid points are evenly spaced across ($-K_M$, K_M). So numerically we want to compute an approximation to the distribution at each grid point. That is for each $i = 1,2,\ldots, N_x$ and $j = 1,2,\ldots, N_k$, calculate a f_{ij} such that $f_{ij} \approx f(x_i, k_j)$.

To approximate the spatial derivative term in Eq. (2), we use a second-order upwind differencing scheme. For the integral terms in Eqs. (3) and (4), we use the midpoint rule in *k* and the trapezoid rule in x for their approximations. Finally, for solving Poisson's equation, we use a standard three-point central differencing scheme. This discretization converts the continuous nonlinear PDE problem to a nonlinear ODE for the solution for *f* at the grid points.

4. CONTINUATION METHODS

We are interested in studying the steady-state Wigner distribution, *f*, of the Wigner-Poisson equations, $\partial f/\partial t = W(f)$, as a function of a system parameter, *v*, which is the applied voltage difference across the RTD. So, in the end, we are trying to find the steady-state Wigner distribution, *f(v)*, which satisfies the nonlinear equation

$$W(f(v)) = 0 \tag{8}$$

as we vary the parameter *v*. To do this, we use continuation methods

.

Continuation methods map out solutions to nonlinear equations that depend on a parameter as a function of this parameter. Continuation algorithms generate a sequence of parameters $\{v^m\}$ along with a corresponding sequence of steady-state solutions $\{f^n\}$ that satisfy $W(f^m(v^m)) = 0$. Since we know that when $v = 0$, the steady-state solution is given by the equilibrium Wigner distribution f_0, then the first terms in these sequences are $v^1 = 0$ and $f^1 = f_0$.

We will now present three common continuation methods. A standard technique for solving nonlinear methods are the zero-order continuation, first-order equations is through Newton's Method, and each continuation method uses it to solve the nonlinear equation. The three continuation methods are zero-order continuation, first-order continuation, and the pseudo-arclength continuation.

Assume we have just computed f^m for some v^m, and now we want to compute the next steady-state solution f^{m+1} for a v^{m+1} that is close to v^m. The zero-order continuation uses Newton's Method to solve $W(f^{m+1}(v^{m+1})) = 0$ using f^m as an initial iterate. This method is called zero-order since it does not attempt to incorporate the effects of changing the parameter v^m to v^{m+1} in our initial iterate for f^{n+1}.

The first-order continuation method considers such a change by trying to approximate the sensitivity of the steady-state solution *f* to the parameter *v*, given by $\partial f/\partial v$, at the previous steady-state solution f^n. To compute this value, we differentiate $W(f(v))$ with respect to *v* to get

$$\frac{\partial W}{\partial v} = \frac{\partial W}{\partial f}\frac{\partial f}{\partial v} = W'(f^m)\frac{\partial f}{\partial v} \tag{9}$$

So solving for $\partial f/\partial v$ involves solving a linear equation where the coefficient matrix is the Jacobian of *W* with respect to *f* evaluated at f^n, denoted by $W'(f^n)$, and the right hand side is $\partial W/\partial v$. To evaluate $\partial W/\partial v$ at the previous steady-state solution f^n, we use a forward difference approximation

$$\frac{\partial W}{\partial v} \approx \frac{[W(f^m(v^m + \delta)) - W(f^m(v^m))]}{\delta} \tag{10}$$

where δ is some small perturbation. Once we have the approximation to $\partial f/\partial v$, the initial iterate for Newton's Method to solve for f^{n+1} at v^{m+1} will be $f^n + \partial f/\partial v * (v^{m+1} - v^m)$.

The final continuation method, pseudo-arclength continuation [4], is useful when continuing around turning points. Turning points are parts of the steady-state solution branches where the branch turns around. When a turning point occurs, the Jacobian matrix becomes singular. So applying Newton's Method is difficult as we approach the turning point since the Jacobian matrix is becoming singular, making the linear solves for the Newton steps harder. Pseudo-arclength continuation handles this problem by augmenting the nonlinear equation $W(f(v))$ with an artificial parameter *s* (the arclength parameter) and an additional arclength equation. So the system we are solving now is

$$\begin{aligned} W(f(v,s)) &= 0 \\ n(f(s), v(s), s) &= 0 \end{aligned} \tag{11}$$

where the first equation specifies that the solution is on the steady-state solution branch, and the second equation specifies the step to take in the parameter *s*. Suppose we have the point (v^m, f^n) on the solution curve and the next solution point to be computed is (v^{m+1}, f^{n+1}). For the next continuation step, the arclength equation is given by

$$n(f(s), v(s), s) = \frac{\partial f^T}{\partial s}(f - f^i) + \frac{\partial v}{\partial s}(v - v^i) - \Delta s$$

where $\Delta \varDelta s$ is the step taken in the parameter *s*. A geometric interpretation of the (v,f) points that satisfy $n(f(s), v(s), s) = 0$ can be given. Suppose *a*, *b*, *c*, and *d* are real numbers and (x_0, y_0, z_0) is a three-dimensional vector. It is a result from analytic geometry that the three-dimensional vectors (x, y, z) that satisfy $a(x - x_0) + b(y - y_0) + c(z - z_0) - d = 0$ lie in the plane perpendicular to the three-dimensional vector (a, b, c) at a distance away from (x_0, y_0, z_0) which is determined by the size of *d*. Similarly, if (v^{m+1}, f^{n+1}) satisfy the arclength equation, then the point will lie in the (v,f) plane perpendicular to the gradient of $(v(s), f(s))$ at some distance away from (v^m, f^n) which is determined by the size of Δs. An example of tracing a one-dimensional system with a one-dimensional parameter using pseudo-arclength continuation is given in Figure 2 to further illustrate this point.

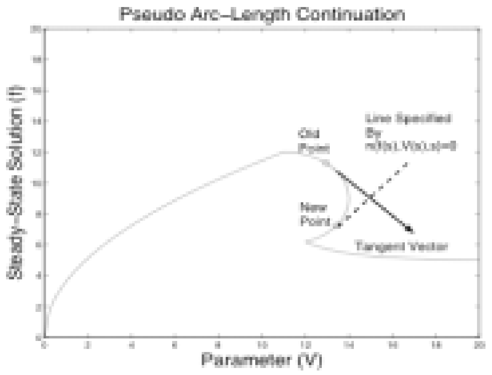


Figure 2 Arclength Continuation

5. LOCA – LIBRARY OF CONTINUATION ALGORITHMS

To implement these continuation methods into our RTD simulator, we used LOCA (Library of Continuation Algorithms), a software library developed at Sandia National Laboratories [5]. This software library was created for large scale bifurcation and stability analysis. It is a part of Sandia's Trilinos project. Trilinos is a collection of Sandia's parallel solver algorithms, and LOCA uses several other parts of Trilinos in its continuation methods. To solve the nonlinear equations, LOCA relies on NOX, Trilinos nonlinear solver. To solve the linear equations created in the Newton iterations, AztecOO, Trilinos's preconditioned Krylov solver, is used. To determine the stability of the computed steady-state solutions, LOCA determines the eigenvalues of the Jacobian of the nonlinear equation *W(f)*. For a given steady-state solution *f** of the nonlinear ODE *df/dt = W(f)*, it is a well-known result [] that the eigenvalues of the Jacobian of *W* at *f**, *W'(f*)*, determines the stability of *f**. If all of the eigenvalues of *W'(f*)* have negative real part, then *f** is asymptotically stable. If any of the eigenvalues of *W'(f*)* have positive real part, then *f** is unstable. To check the eigenvalues of *W'(f*)*, LOCA utilizes Trilinos's eigensolver Anasazi.

6. PRECONDITIONER DEVELOPMENT

The nonlinear solver in the continuation method used for our application was Newton-GMRES. This is an inexact Newton Method, where the linear solution for the Newton steps are solved the Krylov iterative method GMRES [6]. To reduce the number of iterations GMRES takes and therefore reduce the computational burden of the simulation, a preconditioner was developed. When solving the linear equation $Ax = b$, where A is a n by n matrix and x,b are n-dimensional vectors, a preconditioner is another matrix M multiplied into the equation (so now we solve $MAx = Mb$) where the new coefficient matrix MA is an easier matrix for an iterative method to handle. Usually, M is an approximate inverse to A. When solving the linear equations in Newton's Method, the coefficient matrix is always the Jacobian matrix. If we look at Eq. (1), and ignore the last two terms, we get the approximation that $W(f) \approx K(f)$. Since K defined in Eq. (2) is a linear operator, we know $\partial K/\partial f = K$. So an approximation to the Jacobian is $W'(f) \approx K$. Therefore, the preconditioner we use is $M = K^{-1}$.

7. PARALLEL SIMULATOR

To parallelize our evaluation of *W(f)*, we take our domain in *(x, k)* space and distribute among different processors. Here, we decided that each processor would get a contiguous block of *x*-space and all of the corresponding *k*-space that went with each. By splitting the data between the processors this way, we ensure that the integrals in *k*-space can be performed by each processor independently. This splitting, though, will require communication between the processors that calculate the spatial derivative term in Eq. (2). The Poisson solve was not parallelized and is performed by the main processor before everything else is calculated. Once *U(x)* is known, the main processor sends out a copy to rest of the processors. The processors then compute their part of *W(f)*, and return this to the main processor.

The parallel runs reported in this paper were performed on a IBM Blade Center with Xeon 2.8 GHz processors at the North Carolina State University's High Performance and Grid Processing.

8. NUMERICAL RESULTS

The first thing we did with our simulator was to verify our numerical simulation with others that were previously published [1]. While these previously published used a very coarse grid (N_x = 86, N_k = 72), their computational time to analyze the current output for v = 0 to v = 0.480 volts took a few days. Our improved simulator was able to match these results while reducing the computational time to a few hours, while not using any parallel processing. Figure 3 is a plot of the current output versus applied voltage for the coarse grid. The results for the finer grids do not match those in Figure 3, and we are currently exploring the reasons for the difference, which we believe are new physics. We will report on this in future work.

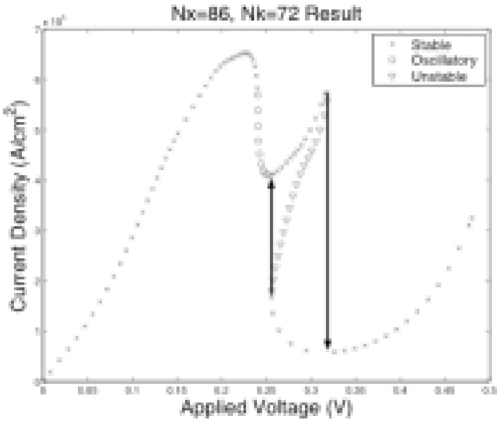


Figure 3: Coarse mesh simulation

Since our simulator was directly computing the steady-state solutions and the previous simulations were using time-accurate methods to reach steady-state, we were able to identify unstable steady-state branches while the time-accurate simulation missed these. These unstable steady-state branches were able to explain the hysteretic effects found on this grid. If the applied voltage is started at zero and is increased, the current stays on the higher stable branch until the voltage is 0.318. The current then drops to the lower stable branch and continues on. If the applied voltage is started at 0.48 volts and is decreased, the current will stay on the lower stable branch until the voltage is 0.25, and then jumps up to the higher stable branch.

Table 1 shows that the preconditioner we use is scalable. The number of GMRES iterations for each Newton step and the number of Newton iterations for each continuation step are essentially independent of the mesh.

Table 1 Krylovs/Newton as mesh is refined

N_x	N_k	Avg. Newton Its. Per Continuation Step	Avg. Krylov Its. Per Newton Step
86	72	2.24	156
172	144	2.51	167
344	288	2.41	180
688	576	2.42	196

As we refine the grids, the number of Newton iterations per continuation step and the number of Krylov iterations per Newton iteration are remaining relatively constant which we expect of a scalable preconditioner.

Table 2 reports on the parallel efficiency of the entire application. The results show that roughly 40% of the code is running in scalar mode.

Table 2: Parallel efficiency

# of Procs.	Linear Solve Time (sec.)	Speedup Factor	Efficiency (%)
1	431.21	---------	--------
2	263.69	1.64	82.0
4	115.71	3.73	93.3
8	75.23	5.73	71.6
16	45.83	9.50	59.4

The grid used in this table is N_x = 688, N_k = 576. As we increase the number of processors used for this job, the efficiency stays above 70% up to 8 processors.

Table 3 presents scalability results of the parallel simulator.

Table 3: Scalability

N_x	N_k	# of Procs.	Avg. *W(f)* Evaluation Time (sec.)
172	144	1	0.1209
344	288	4	0.2814
688	576	16	0.5505

As we quadruple both the number of unknowns and the processors, the function evaluation time should stay flat if the simulator is scaling perfectly. From the table, we see the function evaluation time is doubling. The scaling is consistent with the speedup, telling us the code is 40% serial.

9. CONCLUSIONS

The results from coupling the RTD simulator with LOCA look very promising. We are able to duplicate previously published results at lower computational cost and able to tackle finer grids that before were computationally infeasible. The results from these finer grids seem to indicate that important physics was not resolved in the grid which has been most widely used. We are currently working on getting Fast Fourier Transforms into the simulator to handle the two *x* convolutions in Eq. (4) and the *k* convolution in Eq. (3) that are apart of evaluating the potential energy term *P(f)*. This is the most computationally intensive term, and we anticipate further speedup once the FFTs are incorporated, and we hope they will also improve the efficiency and scalability of our parallel simulator.

10. ACKNOWLEDGEMENTS

The research of MSL, CTK, and PZ was partially supported by a United States Army Research Office DURINT grant. The research of CTK was also partially supported by United States National Science Foundation grants DMS-0209695 and DMS-0070641.

11. REFERENCES

[1]. P. Zhao, H. L. Cui, and D. L. Woolard. Dynamical Instabilities and I-V Characteristics in Resonant Tunneling Systems. Phys. Rev. B, 63:75302, 2001.

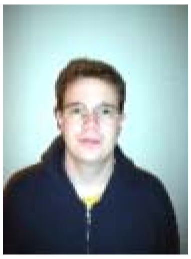
[2]. P. Zhao, H. L. Cui, D. L. Woolard, K. L. Jensen, and F. A.Bout. Simulation of Resonant Tunneling Structures:Origin of I-V Hysteresis and Plateau-Like Structure Journal of Applied Physics, 87:1337-1349, 2000.

[3]. F. A. Bout. and K. L. Jensen. Lattice Weyl-Wigner Formulation of Exact Many-Body Quantum-Transport Theory and Applications to Novel Solid-State Quantum-Based Devices. Phys. Rev. B, 42:9429-9438, 1990.

[4]. H. B. Keller. Lecture on Numerical Methods in Bifurcation Problems. Springer-Velag, New York,1987.

[5]. Andrew G. Salinger, Nawaf M. Bou-Rabee, Roger P Pawlowski, Edward D. Wilkes, Elizabeth A. Burroughs, Richard B. Lehoucq, and Louis A. Romero. LOCA 1.0 Library of Continuation Algorithms: Theory and Implementation Manual. Technical Report SAND2002-0396, Sandia National Laboratory, March 2002.

[6]. C. T. Kelley. Iterative Methods for Linear and Nonlinear Equations, Volume 16 of Frontiers in Applied Mathematics. SIAM, Philadelphia, PA. 1995.



Matthew Lasater is a graduate student in the Mathematics Department at North Carolina State University. His research interests are in numerical methods for bifurcation analysis and the application of those methods to simulation of RTDs. He is the author of three papers, has visited Sandia National Laboratory for two summer internships, and has given several presentations at conferences.

C. T. Kelley is a Drexel Professor of Mathematics at North Carolina State University. His research interests are in numerical methods for the solution of nonlinear equations and optimization problems, parallel computing, and on applications to semiconductor modeling and design, and groundwater flow simulation. He graduated from Purdue University in 1976 with a degree in applied mathematics, and spent a postdoctoral year at the US Army Mathematics Research Center in Madison Wisconsin in the 1977-8 academic year. He has been on the faculty of North Carolina State University since 1978. Kelley serves on the council of the Society for Industrial and Applied Mathematics (SIAM), as editor-in-chief of the SIAM Journal on Optimization, as co-chair of the organizing committee for the 2005 SIAM Annual Meeting, and is on several editorial boards and conference organizing committees. He is the author of three books, all published by SIAM, and over 100 papers.

One New Method and Its Parallelization of Perturbation Expansion for Coupled System of Acoustic and Structure

Deng Li[1], Suzuki Masabumi[2], Hagiwara Ichiro[2]
[1]Japan Research Institute, Limited, Engineering Department
Kudan Bldg. 1-5-3 Kudan-Minami,Chiyoda-ku,Tokyo, 102-0074 Japan and
Tokyo Institute of Technology, Graduate School of Science and Engineering
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
Email: deng@mech.titech.ac.jp, **Tel:** +82-3-5734-3630
[2]Suzuki Masabumi and Hagiwara Ichiro
Tokyo Institute of Technology, Graduate School of Science and Engineering
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
Email: hagiwara@mech.titech.ac.jp, **Tel:** +82-3-5734-3555

ABSTRACT

We introduce a new method for coupled eigenvalue problem, its theoretical approach, finite element approximation and its error estimation, perturbation analysis and its parallelization.

Keywords: coupled eigenvalues, decoupled eigenvalues, and perturbation.

1. INTRODUCTION

In this paper, we study a numerical method to calculate the eigenvalues of the coupled vibration between an acoustic field and a structure.

Figure 1. is a coupled sample inside of a car. A typical example of the structure in our present study is a plate which forms a part of the boundary of the acoustic region. Fig2 (a) is a 3D coupled model. By using Fourier mode decomposition in the z-direction, we reduce it to a 2D coupled eigenvalue problem (see Fig2 (b)) from which we can get the exact solution.

Now the 3D coupled vibration system is given by the following system of partial differential equations:

$$\begin{cases} -c^2\nabla^2_{x,y,z}P-\omega^2P=0, \\ P|_{\Gamma_0}, \\ \dfrac{\partial P}{\partial n}|_{S_0}, \\ D\nabla^4_{y,z}U-\omega^2\rho_1U=P|_{S_0}, \\ U|_{\partial S_0}=\dfrac{\partial^2U}{\partial\sigma^2}=0, \end{cases} \qquad (1)$$

Fig 1

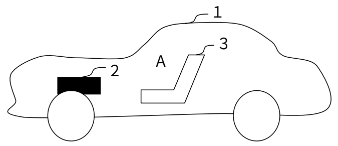
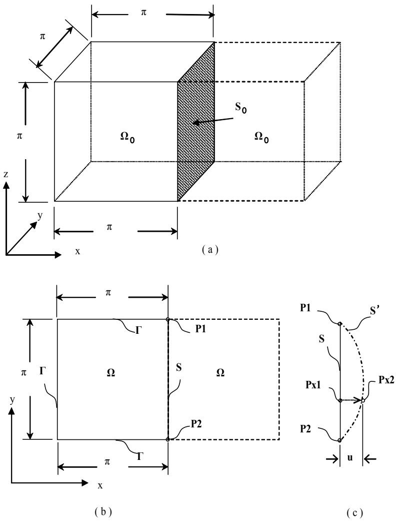


Fig 2



where

Ω_0 : a three-dimensional acoustic region,
S_0 : a plate region,
$\Gamma_0=\partial\Omega_0 \setminus S_0$: a part of the boundary of the acoustic field,
∂S_0 : the boundary of the plate,
P_0 : the acoustic pressure in Ω_0,
U_0 : the vertical plate displacement,
c : the sound velocity,
ρ_0 : the air mass density,
D : the flexural rigidity of plate,
ρ_1 : the plate mass density,
n : the outward normal vector on $\partial\Omega$ from Ω_0, and
σ : the outward normal vector on ∂S_0 from S_0.

We can reduce the problem to a two-dimensional eigenvalue problem. Due to the symmetry of the domain Ω_0 in the z direction and the boundary conditions, we can apply Fourier mode decomposition to P and U in the z direction:

$$\begin{cases} P = \sum_{m=1}^{\infty} P_m(x,y)\sin(mz), \\ U = \sum_{m=1}^{\infty} U_m(x,y)\sin(mz), \end{cases}$$

where *m* represents the Fourier mode. Now we are ready to transform (1) into the reduced from(2), which we will study intensely for the remainder of the paper.

We define a new perturbation method to obtain a perturbation series which expresses the eigen-pairs for the coupled case by those from the decoupled case, i.e., we introduce a parameter ε as follows:

$$\begin{cases} -\nabla_{x,y}^2 P + (-\omega^2/c^2 + m^2)P = 0 \ \text{ in } \Omega \\ P|_\Gamma = 0 \ \text{ on } \Gamma \\ \dfrac{\partial P}{\partial x}|_S = \varepsilon \rho_0 \omega^2 u \ \text{ on S} \\ \mathrm{D}(\partial^4 u - 2m^2 \partial^2 u) - \omega^2 \rho_1 u = \varepsilon P|_S \ \text{ on S} \\ \mathrm{u(0)}{=}\partial^2 u(0) = u(\pi) = \partial^2 u(\pi) = 0 \ \text{ on S} \end{cases} \tag{2}$$

If parameter *ε=0*, then the coupled problem will be separated into two uncoupled problems. The first represents Acoustic Vibration:

$$\begin{cases} -\nabla_{x,y}^2 P + (-\omega^2/c^2 + m^2)P = 0 \ \text{ in } \Omega \\ P|_\Gamma = 0 \ \text{ on } \Gamma \\ \dfrac{\partial P}{\partial x}|_S = 0 \ \text{ on S} \end{cases} \tag{3}$$

The second represents Plate Vibration:

$$\begin{cases} \mathrm{D}(\partial^4 u - 2m^2 \partial^2 u) - \omega^2 \rho_1 u = 0 \ \text{ on S}, \\ \mathrm{u(0)}{=}\partial^2 u(0) = u(\pi) = \partial^2 u(\pi) = 0 \ \text{ on } \partial\mathrm{S}. \end{cases} \tag{4}$$

2. AN OPERATOR THEORETICAL APPROACH FOR THE COUPLED EIGENVALUE PROBLEM

For a rather general 2D bounded domain with its boundary, the eigenvalue problem (1) is written in the following weak form:

$$\begin{cases} \rho_0^{-1}\left[<< p,q >> + m^2((p,q))\right] = \omega^2 \left[\dfrac{1}{\rho_0 c^2}((p,q)) + (u,\gamma sq) \right], \\ D\left[(u'',v'') + 2m^2(u',v') + m^4(u,v)\right] - (\gamma sp, v) = \omega^2 \rho_1 (u,v), \end{cases}$$

where

$$<< p,q >>= \int_\Omega \nabla_{x,y} p g\nabla_{x,y} \overline{q} \ dxdy,$$

$$((p,q)) = \int_\Omega p\overline{q} \ dxdy, \text{ and}$$

$$(p,q) = \int_0^\pi u\overline{v} \ dy \ \left(= \int_0^\pi u(\pi,y)\overline{v}(\pi,y) \ dy \right).$$

We use standard notation for Sobolev spaces and introduce the following function spaces:

$$V_S = \left\{ p : \ p \in H^1(\Omega), \ p|_\Gamma = 0 \right\} \text{ and}$$
$$V_P = \left\{ u : \ u \in H^2(S), \ u|_{\partial S} = 0 \right\}.$$

We define a collection of bilinear forms on the function spaces V_S and V_P:

$$a_S(p,q) = \frac{1}{\rho_0}\left[<< p,q >> + m^2((p,q)) \right], \ p,q \in V_S,$$
$$b_S(p,q) = \frac{1}{\rho_0 c^2}((p,q)), \ p,q \in V,$$
$$a_P(u,v) = D\left[(u'',v'') + 2m^2(u',v') + m^4(u,v) \right], \ u,v \in V_P,$$
$$c_\theta(u,q) = (u,\gamma sq), \ u \in V, q \in V_S, \text{ and}$$
$$\overline{c}_\theta(p,v) = (\gamma sp, v), \ v \in V, p \in V_S.$$

Then the weak formulation is rewritten into

$$\begin{cases} a_S(p,q) = \omega^2\left[b_S(p,q) + c_\theta(u,q) \right], \\ a_P(u,v) - c_\theta(p,v) = \omega^2 b_P(u,v). \end{cases}$$

We are now ready to state the operator theoretical formulation. Define operators A_S, A_P, T, and T^* using the Riesz representation theorem as follows:

$$\begin{cases} b_S(p,q) = a_S(A_S p, q), \\ b_P(u,v) = a_P(A_P p, q), \end{cases}$$

$$\begin{cases} c_\theta(u,q) = a_S(T^* u, q), \\ \overline{c}_\theta(p,v) = a_P(Tp, v). \end{cases}$$

The operators A_S and A_P are self-adjoint and compact. The operators T and T^* are compact. By using these operators, the eigenvalue problem is transformed into form

$$\begin{cases} a_S(p - \omega^2(A_S p + T^* u), q) = 0, \\ a_P(-Tp + u - \omega^2 A_P u, v) = 0, \end{cases}$$

$$\begin{cases} p = \omega^2(a_S p + T^* u), \\ u - Tp = \omega^2 A_P u, \end{cases}$$

which leads to the matrix formulation

$$\begin{bmatrix} I & 0 \\ -T & I \end{bmatrix} \begin{bmatrix} p \\ u \end{bmatrix} = \omega^2 \begin{bmatrix} A_S & T^* \\ 0 & A_P \end{bmatrix} \begin{bmatrix} p \\ u \end{bmatrix}. \tag{5}$$

A finite element formulation follows from this immediately.

In practical computations, the matrix equations are written into

$$\begin{cases}(K_a-\omega_h^2M_a)\tilde{p}-\omega_h^2L\tilde{u}=0,\\ -L^T\tilde{p}+(K_a-\omega_h^2M_a)\tilde{u}=0,\end{cases}$$

where K$_a$ and M$_a$ are the stiffness and mass matrices for the acoustic field, K$_p$ and M$_p$ are the stiffness and mass matrices for the plate, and L and L^T are the coupling matrices.

The precise definitions of K_a, M_a, K_p, M_p, L and L^T are as follows: for $u\epsilon V_P$ and $p\epsilon V_S$,,

$$\begin{aligned}&(K_a)_{ij}=\frac{1}{\rho_0}\left[<<p,p>>+m^2(p,p)\right],\\ &(M_a)_{ij}=\frac{1}{\rho_0c^2}((p,p)),\\ &(K_p)_{ij}=D\left[(u'',u'')+2m^2(u',u')+m^4(u,u)\right],\\ &(M_p)_{ij}=\rho_1(u,u),\\ &(L)_{ij}=(u,\gamma sp),\ \text{and}\\ &(L^T)_{ij}=(\gamma sp,u).\end{aligned}$$

3. PERTURBATION ANALYSIS FOR THE COUPLED VIBRATION PROBLEM

We introduced a coupling parameter ε in (2). Consider the perturbed version of (5):

$$\lambda\begin{bmatrix}I & 0\\ -\varepsilon T & I\end{bmatrix}\begin{bmatrix}p\\ u\end{bmatrix}=\begin{bmatrix}A_S & \varepsilon T^*\\ 0 & A_P\end{bmatrix}\begin{bmatrix}p\\ u\end{bmatrix}.$$

We introduce a Hilbert space V as follows: $V=V_s\,x\,V_p$ with a inner product A_S+A_P and define a vector x in V and operators A and B acting on V as follows:

$$\varphi=\begin{pmatrix}p\\ u\end{pmatrix},$$

$$A=\begin{pmatrix}A_S & 0\\ 0 & A_P\end{pmatrix},\ \text{and}\ B=\begin{pmatrix}0 & 0\\ T & 0\end{pmatrix}.$$

Then the original eigen-pair x and λ is changed into the eigen-pair $x(\varepsilon)$ and $\lambda(\varepsilon)$ of the eigenvalue problem:

$$(A+\varepsilon B^*)\varphi(\varepsilon)=\lambda(\varepsilon)(I-\varepsilon B)\varphi(\varepsilon).$$

Then when $\varepsilon=0$ we obtain:

$$A\varphi(0)=\lambda(0)\varphi(0).$$

We impose the following normalization condition for φ(0) and φ(ε):

$$(\varphi_i(0),\ \varphi_j(0))=\delta_{ij}\ \text{and}\ (\varphi_i(\varepsilon),\ \varphi_i(0))=1,$$

where δ_{ij} is the Kronecker delta function.

This leads to the un-symmetric formulation with reduced conditions

$$\begin{cases}\lambda_i(\varepsilon)=\lambda_i\ (0)+\lambda_i^{(2)}\varepsilon^2+\lambda_i^{(4)}\varepsilon^4+\lambda_i^{(6)}\varepsilon^6+...\\ \varphi_i(\varepsilon)=\varphi_i\ (0)+\varphi_i^{(1)}\varepsilon+\varphi_i^{(2)}\varepsilon^2+\varphi_i^{(3)}\varepsilon^3+...\end{cases}$$

where the coefficients of the perturbation series $\lambda(\varepsilon)$ and $\varphi(\varepsilon)$ will now be derived.

Using the finite element formulation, we derive the solution to the coupled eigenvalue problem. Note that

$$\lambda_h^2\begin{bmatrix}K_a & 0\\ -\varepsilon L^T & K_P\end{bmatrix}\begin{bmatrix}p_h\\ u\end{bmatrix}=\begin{bmatrix}M_a & \varepsilon L\\ 0 & M_P\end{bmatrix}\begin{bmatrix}p_h\\ u\end{bmatrix},$$

where

$$\lambda_h^2(\varepsilon)=\omega_h^{-1}(\varepsilon),\ \varphi_{\mathrm{h}}(\varepsilon)=\begin{bmatrix}p_h\\ u\end{bmatrix}.$$

There are two orthonormality conditions for the eigenvector:

$$\varphi_{ih}(0)\begin{bmatrix}K_a & 0\\ 0 & K_P\end{bmatrix}\varphi_{jh}(0)=\delta_{ij}\ \text{and}\ \varphi_{ih}(\varepsilon)\begin{bmatrix}K_a & 0\\ 0 & K_P\end{bmatrix}\varphi_{ih}(0)=1.$$

Hence,

$$\begin{aligned}&\lambda_i^{(2n+1)}=0,\\ &\lambda_i^{(2n)}=(T^T+\lambda_iT)\varphi_i^{(2n-1)}+\lambda_iT\varphi_i^{(2n-3)}+...+\varphi_i,\\ &\varphi_i^{(n)}=\left\{\sum_{i\neq j}\frac{1}{\lambda_i-\lambda_j}\left\{\begin{matrix}\sum_{k=1}^{n-1}[\lambda_i^{(n-k)}(\varphi_i^{(k)},\varphi_j)-\\ (T^T+\lambda_i,T)+\\ \lambda_iT\varphi_i^{(2n-3)}+...+\varphi_j]\end{matrix}\right\}\right\}.\end{aligned}$$

When the higher order remainder terms are small we have

$$\lambda_1(\varepsilon)=\lambda_1(0)+\varepsilon 2\lambda_1(2)+\mathrm{O}(\varepsilon^4). \qquad (6)$$

While the results were derived from 2D equation, they are useful in the 3D case as well.

4. NUMERICAL RESULTS

We will show several sets of numerical results to investigate the validity of (6) in our presentation.

5. FUTURE WORK

We expect to obtain a mathematically rigorous estimation of the magnitude of the convergence radius of the perturbation series in the near future.

We will also calculate 3D problems with more complicated shapes of the acoustic region such as the cases of motor vehicles by using the finite element method and to get the error estimation for such calculations.

6. REFERENCES

[1]. Deng, L., Kako, T. and Hagiwara, I., Development of coupled structural-acoustic eigen-pair expression from decoupled eigen-pair (1st report, Induction by finite perturbation series), Transactions of the Japan Society of Mechanical Engineers (Part C), Vol.63 (1997), pp. 3446-3453 (in Japanese).

[2]. Deng, L. and Kako, T., Finite element approximation of eigenvalue problem for a coupled vibration between acoustic field and plate, Journal of Computational Mathematics, Vol.15, No.3 (1997), pp. 265-278.

[3]. Craggs, A. and Stead, G., Sound transmission between enclosures-A study using plate and acoustic finite elements, ACOUSTICA, Vol.35 (1976), pp. 89-98.

[4]. Babuska, I. and Osborn, J.E., Eigenvalue problem, in Handbook of Numerical Analysis, Vol.2, Finite Element Methods (Part 1), 1991, Elsevier, Amsterdam and New York, pp. 683-692.

Deng Li has been a Chief Engineer in the Engineering Division of the Japan Research Institute, Ltd. (Tokyo) since 1997. She has also been a researcher in the Graduate School of Science and Engineering at the Tokyo Institute of Technology since 2002. She is on leave and is currently a Visiting Scientist in Computational Sciences at the University of Kentucky through 2007. Deng graduated from Sun Yat-Sen University (Guangzhou) in 1984 with a B.S. in Mathematics and Mechanics. She received her M.S. (1994) and Ph.D. (1998) degrees in Computer Science from the University of Electro-Communications (Tokyo).

Parallel Reservoir Integrated Simulation Platform For One Million Grid Blocks Case*

Pan Feng, Cao Jianwen, Sun Jiachang
Research & Development Center for Parallel Software,
Institute of Software, Chinese Academy of Sciences
P.O. Box 8718, Beijing, P.R. China
Email: { pan,cao,sun }@mail.rdcps.ac.cn **Tel:** 86 10 6255 3467

ABSTRACT

This article first provides a brief introduction to the numerical reservoir simulation and a parallel numerical reservoir integrated simulation platform from RDCPS (Research & Development Center for Parallel Software, Institute of Software, Chinese Academy of Sciences), including Pre-Processing, Simulator (for a Three-Dimensional & Three-Phase Black Oil models), Post Processing, seamlessly integrated with parallel computers. We then present key technologies of the simulator, such as the nonlinear and linear solvers, communications among processors, parallel I/O, etc., and corresponding resolvents. Finally, some results with the platform to solve one million grid blocks cases from Chinese oil fields will be given in the article, which can show that the simulator has a very robust portability, high-speed for deadline and good scalability for the tested cases. As application software, our object is always focusing on meeting deadlines from oil industry. Now, for one million grid blocks' case with 20 ~ 30 years production, its elapsed time with 16 processors is less than 12 hours on parallel computers based on Myrinet or QsNet, namely "to submit a case just before off-duty and get its result just before on-duty". A decreasing line of elapsed time appears for a one million grid blocks case. The developing trace of the simulator along with parallel computers can be also inferred.

Keywords: parallel numerical reservoir simulation, fine residual-oil distribution, one million grid blocks problem, integrated simulation platform, parallel computer, deadline, grid computing.

1. INTRODUCTION

Numerical reservoir simulation [1,2,3,4] describes quantitatively the flow of multiple phases in permeable media with computers, predicts the behavior of an oil field, and makes production schedules (determined by properties of oil field, market demand, investment strategy and government regulations).

Fine Numerical Reservoir Simulation

After several decades' production, oil layers of the main oil fields in China, such as Daqing oil field, Shengli oil field, are all in high-water, high-scattered. How to successfully carry out the second and the third production with the considerable, effective plan? A key problem is to know their fine distributions of residual oil with large-scale numerical reservoir simulations. In addition, experts from oil fields say that the simulation is truly valuable for them, if the elapsed time can be within twelve hours (to submit a case just before off-duty and get its result just before on-duty) or even within eight hours (to submit a case just before on-duty and get its result just before off-duty).

Because of sluggishness of application software on parallel computers and strict apply deadline in oil industry, the size of most simulation models in China still ranges from 100,000 to 300,000 grid blocks, far from the number of ones for the fine distribution of residual oil. So much geognostic message is ignored and the fine distribution cannot be given. In recent years, on the one hand, the number of geognostic models' grid blocks is up to more than several million even ten million with new methods, which meets the fine numerical reservoir simulation. On the other hand, the ever-increasing computing ability of current parallel computers provides petroleum industry computing ability up to Tera peak performance. So fine reservoir simulation is brought forward, which studies fine residual-oil distributions with a single sand layer, not traditional coarse-layer of several single-sand layers and matches the relevant geological description information. But the number of grid blocks increases to one million, even a few hundred million. Large-scale computations come into being.

In 1997, the Center for Petroleum and Geosystems Engineering (CPGE), University of Texas at Austin on IBM SP also did some valuable work. Their largest test involved four million grid blocks and 32 million unknowns and took approximately 23 minutes to run on a 128-processor IBM SP. In 1997, J.W.Watts [5] also gave a prediction of what the reservoir simulation state of the art will be in 2007 and speculation regarding certain aspects of simulation in 2017. He predicted that the largest black-oil simulations would use about 100 million grid blocks.

Parallel Numerical Reservoir Simulator of RDCPS

The reservoir simulator group of RDCPS in Institute of Software of Chinese Academy of Sciences has been developing advanced techniques for numerical reservoir simulation since 1992. During these years some progress has been made, such as high effective parallel solver for Three-Dimensional & Three-Phase problems [5,6,7], parallel simulator for Three-Dimensional & Two-Phase problems [8], and parallel simulator for Three-Dimensional & Three-Phase problems [9,10,11,12,13] of black oil model. Our final aim is to generate an integrated process for oil field management: geognostic model – numerical reservoir model -- numerical simulation -- auto analysis and fit -- prediction of residual oil distribution, namely, Pre-Processing, Simulator, Post Processing.

Black Oil Model

It is well known that the mathematical reservoir model is a

*This work was supported in partial by Major State Basic Research Project (No. G19990328, Parallel Computations of the Super-large Scale Reservoir Simulation (2003-2004) (Cooperated with China National Offshore Oil Corporation), and National Natural Science Foundation Project (No.60303020, 2004.1-2006.12).

system of coupled nonlinear equations. It exhibits behavior typical of the solutions of both parabolic and hyperbolic partial differential equations. It can be expressed as,

$$\nabla \cdot [\frac{\rho_o^o K K_{ro}}{\mu_o}(\nabla P_o - \rho_o g \nabla Z)] + q_o^o = \frac{\partial(\varphi \rho_o^o S_o)}{\partial t}$$

$$\nabla \cdot [\frac{\rho_w K K_{rw}}{\mu_w}(\nabla P_w - \rho_w g \nabla Z)] + q_w = \frac{\partial(\varphi \rho_w S_w)}{\partial t}$$

$$\nabla \cdot [\frac{\rho_g K K_{rg}}{\mu_g}(\nabla P_g - \rho_g g \nabla Z) + \frac{\rho_o^g K K_{ro}}{\mu_o}(\nabla P_o - \rho_o g \nabla Z)] + q_g + q_o^g$$
$$= \frac{\partial[\varphi(\rho_o^g S_o + \rho_g S_g)]}{\partial t}$$

$$S_o + S_w + S_g = 1 \qquad P_{cow} = P_o - P_w \qquad P_{cog} = P_g + P_o$$

More details refer to [1,2,3,14].

2. KEY TECHNOLOGIES

First a reasonable strategy is to parallelize old sequential simulators, namely to divide a large-scale memory needed by them into nearly equally small ones with "divide and conquer". In addition, a transparent idea is used for avoiding special knowledge about petroleum engineering.

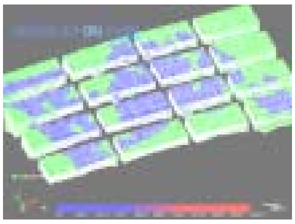
We also follow some rules. Just as Prof. Thomas F. Russell of department of mathematics in University of Colorado at Denver told me in an email that in most situations the engineer's highest priority is to get an answer, speed is the second priority, because of deadlines and accuracy is third. So the simulator chooses full implicit method to solve nonlinear equations of black oil model with Newton-like and Krylov subspace methods [5,6,7,15,22], uses domain decomposition method [16,20], transparent boxes method [8] and category method [7,11,12] to parallelize (the requirement of memory is up to 4 G for 1M grids problems) for increasing speed because of deadline, devises communication encapsulation (the customized communication library for various parallel computers), multi-tree structure [7,17] and minimization with relevant analysis [18] (communication on some distributed memory machines is the same important as computation), utilizes unformatted, direct file I/O, minimization I/O times with memory, and parallel file system to deal with large I/O requirement [19], etc.

3. ONE MILLION CELLS' SIMULATION

In this section, some results of one million grid blocks cases from Chinese oil fields will be given, which can show that the simulator has a very robust portability, high-speed for deadline and good scalability. Foremost, as application software, it can meet deadlines from oil industry, the elapsed time of which with 16 processors can be less than 12 hours on parallel computers with higher bandwidth and lower latency networking hardware, such as Myrinet and QsNet, namely "to submit a case just before off-duty and get its result just before on-duty".

A senior engineer (Baoshu LI) from Daqing oil field of China National Petroleum Corporation (CNPC) suggested us to solve a one million grid blocks' case within 24 hours in 1996. We simulated one million grid blocks case from Daqing oil field with sequential codes under only single user status on SGI Power challenge XL (16 195MHz MIPS R10000 processors, 4GB of Memory) in, did not completed after eighteen days in 1997. It is in 1998 that RDCPS developed the simulator, a Three-Dimensional, Three-Phase Black Oil Simulator on a quad-node Linux PC-Cluster (each with a single 450MHz Intel Pentium III processors, 256MB of Memory).

Cases with One Million Grid Blocks



In the paper, there are two one million grid blocks' cases from Daqing oil field used. Case 1 in China is a big-difficulty one with a 3-dimensional and 3-phase grid (199 x 87 x 67), 291 wells, and 31.5 years production period (from 1966. 12 to 1998.6), which covers one of 60s of total Daqing field. It amounts to 1,159,971 grid blocks and 3,479,913 unknowns. Its grid blocks system and the 16-processor domain decomposition are given correspondingly in the up Figure, which is all exported by post processing of our platform. Case 2 is a middle-difficulty 1,047,200 grid blocks one, which is a 175 x 176 x 34 fine grid refined from a 35 x 44 x 34 coarse grid, with 36 wells, production period from 1960.5 to 1973.12. Its total unknowns are 3,141,600).

Robust Portability

Since 1998, our simulator are ported onto more than thirteen parallel computers, including self-made PC-Cluster, SunWay-I, Dawning 2000 and 3000 series, DeepComp 1800 and 6800 series, etc. So the code of the simulator becomes modulized more and more, accommodating various parallel computers. Its modulization lies in option of compilers (LINUX, AIX, IRIX), communication (MPI, PVM, Multi-tree mode), I/O (NFS, Parallel FS), etc. And PC-Cluster is becoming popular more and more, which also is concluded in TOP500 and Chinese TOP100. PC-Cluster makes it possible from hardware condition for us to develop large-scale parallel, high-complex, high-efficient software. RDCPS set up the first such cluster in 1998 and owns three clusters now.

High-Speed for Deadline

As application software, the simulator's object is always focusing on meeting deadlines from oil industry because experts from oil fields say that the simulation is truly valuable, the elapsed time of which is within twelve hours (to submit a case just before off-duty and get its result just before on-duty) and within eight hours (to submit a case just before on-duty and get its result just before off-duty). A decreasing elapsed time line appears for case 1, just as Figure 1. In October 1999, it cost about 64 hours to complete case 1 a Linux PC-Cluster (16 nodes, each node with one 500MHz Intel Pentium III processor and 256MB of Memory (320 MB for main service node)) self-made by RDCPS. For the case, the performance of the simulator running the cluster was quite the same as that of Parallel VIP (Landmark) on SGI-Origin 2000. In 2000, supported by the Hi-Tech Research and Development Programme of China (863 Programme, "fine reservoir numerical simulation of large-scale whole fields" cooperation with the Daqing oil field) from Ministry of Science and Technology of the People's Republic of China, the simulator was ported onto many parallel computers, such as Dawning 2000-II (released on Feb. 2000 in China), Dawning

3000(released on February 2001 in China) etc. The elapsed time for the case 1 was firstly less than 12 hours, "to submit a case just before off-duty and get its result just before on-duty", which met the deadline of petroleum industry. We completed the same problem within one hour on the two Teracluster, one is the DeepComp 1800 of Legend Corp. at Academy of Mathematics and System Science, CAS., which owns 256 nodes (Red Hat Linux, 1Gigabit Ethernet and Myrinet2000), each with dual 2 GHz Intel Pentium IV Xeon. The other is the DeepComp 6800 of Legend Corp. at Computer Network Information Center (CNIC), CAS., which has 256 nodes (Quadrics QsNet Interconnect), each with quad Itanium2 1.3 GHz.

Now, for one million grid blocks' simulation with 20 ~ 30 years production, its elapsed time with 16 processors is less than 12 hours on parallel computers connected with Myrinet or QsNet, namely "to submit case just before off-duty and get its result just before on-duty". The developing trace of our simulator along with parallel computers can be also inferred from Figure 1.

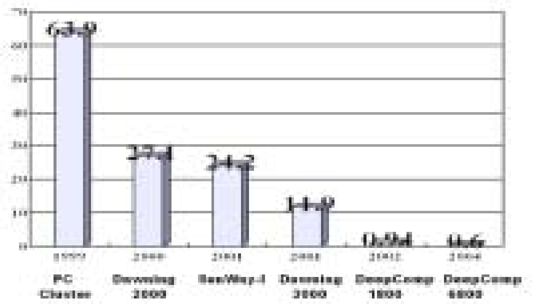


Figure 1

Good Scalability

The result of case 2 on Dawning 2000-II is given in Table 1. The communication time is 2.08 hours (24% of elapsed time) for 16 processors, 1.77 hours (36% of elapsed time) for 32 processors, and 1.05 hours (40% of elapsed time) for 64 processors. Compared with 16 processors case, the relative efficiencies (compared to the 16-processor's result) of 32 processors and 64 processors are 85%, 80% respectively. Scalability of our simulator v1.0 on Dawning 2000-II for this problem is pretty good, a near linear scalability.

Table 1: Elapsed Time and Relative Speedup for Case 2 on Dawning 2000-II (Motorola PowerPC604e 333MHz, Local memory 512M, Myrinet Netcard)

Number of Processors	16	32	64
Elapsed Time (Hr.)	8.37	4.92	2.61
Relative Speedup	1.00	1.70	3.20

For the case 1, we show computation results of our simulator v2.1 on a Teracluster (CNIC, 256 nodes, each with quad Itanium2 1.3 GHz, Quadrics QSNet Interconnect). All the results are one processor on each node for fully observing scalability (because the communication and I/O performance of several processors in each node are quite different from that of one processor in each node in this parallel computer, the results of one processor in each node are more valuable for us). From Table 2 we can find that the relative efficiencies (compared to the 16-processor's result) of 32 processors, 64 processors, 128 processors, and 256 processors are 95%, 95%, 66% and 37% respectively. A good scalability of our simulator v2.1 for this case is presented.

Table 2: Elapsed Time and Relative Speedup for Case 1 on Teracluster (256 nodes, each with quad Itanium2 1.3 GHz, Quadrics QsNet Interconnect)

Number of Processors	16	32	64	128	256
Comm. Time(Hr.)	0.78	0.57	0.42	0.37	0.32
Elapsed Time (Hr.)	4.83	2.56	1.28	0.92	0.81
Relative Speedup	1.00	1.89	3.77	5.25	5.96

From Table 2, we can observe that the communication times are decreasing from 16 processors to 256 processors (we also observe a little increase in the case of 512 processors). It indicates that message size and complexity affect each other during communication of various processors. It is also concluded that the elapsed time of the results more than 128 processors do not decrease largely for case 1. There are two reasons, one is message-passing cost, and another is parallel I/O. They are the key difficulties and focuses for parallel computer manufacturers and parallel software developers forever.

4. CONCLUSION AND FUTURE WORK

From the article we can conclude that a parallel numerical reservoir integrated simulation platform developed by Research & Development Center for Parallel Software (RDCPS), Institute of Software, Chinese Academy of Sciences, can meet strict deadlines from petroleum industry for one million grid blocks' problems with 20 ~ 30 years production. And it has a very robust portability, high-speed for deadline and good scalability. One obvious problem is the large ratio of communication and parallel I/O time in total wall time when the number of processors is 512 or above. How to adjust message size and communication complexity for optimizing communication? Parallel file system becomes indispensable in this case.

With grid computing springing up, RDCPS is aiming at developing a new parallel numerical reservoir integrated simulator platform which can solve 1 ~ 10 G grid blocks' system with efficiently using 512 ~ 1024 processors or about within computing nodes on GRIDs. Why we do it? Our simulator is not only integrated application software, but also a very good application benchmark for parallel computers, especially for Grid Computing. There are large-scale I/O operations, message passings, and computations. It can give us an all-sided checking for GRIDs. Of course, there are much more requirements from oil fields with increasing production years in the future.

5. ACKNOWLEDGEMENT

We thank Dr. Shangmeng WEN and Cunyu WANG for their part work, Prof. Wenlan CHEN for her helpful discussions of several years, and Prof. Xuebin CHI for giving us a good chance to complete our simulations on their new Teracluster. We gratefully acknowledge one million grid blocks cases from Daqing oil field and Shengli oil field. We also thank Computer Network Information Center (CNIC), CAS., Academy of Mathematics and System Science, CAS., Dawning Corp., and Legend Corp. for using their computers.

6. REFERENCES

[1] K.Aziz, A.Settari: Petroleum Reservoir Simulation, Applied Science, 1979.

[2] Calvin, C., Mattax, R. and Dalton, L.: Reservoir Simulation (SPE Monograph series, vol. 13): Society of Petroleum Engineers, 1990.

[3] Fanchi, J. R.: Principles of Applied Reservoir Simulation, 2nd ed.: Gulf Publishing 2001

[4] Peaceman D.W.: Fundamentals of Numerical Reservoir Simulation", Elsevier Scientific Publishing Company, 1977

[5] Watts, J.W.: Reservoir Simulation: Past, Present, and Future, paper SPE 38441,proceedings of the 14th SPE Symposium on Reservoir Simulation, Dallas, TX, June 8-11, 1997.

[6] Jianwen CAO, Feng PAN: Some Parallel Preconditioners and Application to Reservoir Simulations of Black-oil Model, Proceedings of 5th Chinese Parallel Computing, ISBN7-5369-2751-7, pp.71-75; Sept. 1997, Xi-an, China.

[7] Jianwen CAO, Choi-Hong LAI: Numerical Experiments of Some Krylov Subspace Methods for Black Oil Model, An International Journal of Computers and Mathematics with Applications, Elsevier, 2002, Volume:44, Issue:1/2, PP125-141

[8] Jianwen CAO: Efficient and Effective Solvers with Preconditions in the Parallel Software of Large-Scale Reservoir Simulation, PhD Dissertation, Jun. 2002

[9] Shangmeng WEN, Jiachang SUN: Paralleling the Initialization Part for A Large Sequential Numerical Software， International Conference on HPC ASIA, Singapore, Sept.22-25, 1998, 270-275.

[10] Jianwen CAO, Feng PAN, Jiachang SUN, Wei Liu: Large-Scale Parallel Reservoir Simulation on Distributed Memory Systems, Proceedings of 2001 International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Hubei Science and Technology Press, ISBN 7-5352-2722-8. pp.98-103; Oct.2001,Wuhan, China.

[11] Jianwen CAO, Feng PAN, Jifeng YAO, Jiachang SUN, Guozhong ZHAO: The Implementation of a Parallel Software of Petroleum Reservoir Simulation and Its Application, Journal of Computer Research and Development, Vol.39 No.8 pp.971-980; August 2002,

[12] Feng PAN: Parallelizing Whole Arrays in Large-Scale Sequential Numerical Software with Classify Methods, Proceedings of 6th Chinese Parallel Computing, ISBN 7-81024-675-5, pp.251-254; Oct. 2000, Changsha, China.

[13] Feng PAN: The Category of Program Structure and Its Applications in Parallelization, Proceedings of 4th International Conference/Exhibition on HPC in Asia-Pacific Region, Vol.2, pp1007-1009, IEEE Computer Society ISBN 0-7695-0589-2; May 2000, Beijing China.

[14] Wei LIU, Jianwen CAO, Alberto MEZZATESTA and Peng ZHU: Parallel Reservoir Simulation on Shared and Distributed Memory System, SPE64797, 2000

[15] Bell.J.B.: Mathematical Structure of the Black-Oil Model for Petroleum Reservoir Simulation, SIAM Journal on Applied Mathematics 49, 1989, 749-783.

[16] Jianwen CAO, Numerical Experiments and Analyses of Krylov Subspace Methods on One Kind of PDE, Journal on Numerical Methods and Computer Applications, 1999, Vol.20, No.4, PP255-265

[17] Zhiyuan MA, Fengjiang JING, Xiangming XU, Jiachang SUN: Simulation of black oil reservoir on distributed memory parallel computers and workstation cluster， SPE 29937, 505-512, 1995.

[18] Yuqing XIONG, Jianwen CAO, Xiang ZHANG: OilCL: A Communication Library for Oil Reservoir Numerical Simulation Parallel Computing, Chinese Journal of Computers, 2000, V23, No.7, PP744-749

[19] Hao ZHU, Jianwen CAO: A Comparison Based Tool for Checking Parallelised Large Sequential Programs, 2000,The Fourth International Conference/Exhibition on HPC in Asia-Pacific Region, Vol 1, HPC-ASIA 2000, IEEE Computer Society

[20] Feng PAN, Jianwen CAO: Some Parallel I/O Discussions of Large Scale Parallel Reservoir Simulations on Linux Clusters, Proceedings of 7th Chinese Parallel Computing, Aug. 2002, Chengdu, China.

[21] Feng PAN: Convergence of Multiplicative Schwarz Scheme in Strictly Diagonal Dominant Systems, Proceedings of 6th Chinese Computational Mathematics, pp.119; Oct. 1999, Shanghai, China.

[22] Jiachang SUN, Jianwen CAO: Large scale petroleum reservoir simulation and parallel preconditioning algorithm research, Science in China Ser.A, Mathematics 2004, vol.47 Supplement, April 2004.

Pan Feng is a engineer of research and development for parallel software (RDCPS), Institute of Software, Chinese Academy of Sciences. He got the Bachelor degree from Heilongjiang University in 1995 and the Master degree from Jilin University in 1998, majoring computational mathematics. His specialty is research, development, and application of parallel software with mathematics, especially for petroleum engineering. He is a holder of the second-class prize of "the State Scientific and Technological Progress Award (2000)".

The Study of Distributed Hydrologic Data Integration Based on CORBA *

Lou Yuan-Sheng, Wang Zhi-Jian,, Ai Ping, Zhou Xiao-Feng, Xu feng
Computer and Information Engineering institute, Hohai University
Nan Jing, Jiang Su 210098, China
Email: wise.lou@163.com **Tel:** 13951836094

ABSTRACT

Flood control and other water conservancy task requires coherence and centralized hydrologic information service. In order to satisfied this requirements, this paper shows a Distributed Hydrologic data Integration Method. Data acquisition components collect the data, and transmit them to database in different hydrologic Information center. Data service components provide information service to application or other components. The application of active data services based on CORBA event services makes the data transmitted to database or other component in time.

Keywords: Hydrologic data, Distributed, CORBA , Data Integrate, Component.

1. INTRODUCTION TO HYDROLOGIC DATA COLLECTION

Traditional hydrologic data collected by manual station, and send to different level hydrological data center by telegram code, decode system transform the code to actual data and write them to database. This method not only have low efficiency, but also easy to make mistake. The hydrological telemetering technology reduces the middle tache of Hydrologic data collection and transmission, so the reliability of the data gained effectively guarantee. The developments of communication and network make the distributed computing use widely, these status provide new approach to hydrologic data integration.

The hydrological telemetering system based on network consists of metering stations (which are located in the rivers, lakes and ocean), center stations (sub-central stations) and control center, in which one center station connects many metering stations and one control center connects many center stations. According to the analysis, the hydrological telemetering system based on network mainly includes six modules: SCADA, data preprocess, discharge calculation, data transfer, database maintenance and information query [1]. How to integrate these functions is the guarantee of the hydrologic data integration and how to provide reliably service.

The hydrological telemetering system based on network advanced the reliability and veracity of the data, but on account of the distributed of collection and transmission, there are some problem should be solved:

1) **The coherence of the data:** because of the collected data may separated store in data center station and control center, this maybe lead to the difference of the same data.

2) **The betimes of the data:** the collected data must be transmitted into different database in time.

3) **The difference of database system and storage method:** the data service must be solving this difference.

Distributed application aim at how to deploy, manage, and maintenance the sharing source in network computing environment. It composes the dispersing computing source to a logic integer, and shields the different environment, this makes the software components communicated and cooperated through middleware. The data integration system must have expansibility, and implement “plug and play” of different data source, it is difficult to realize this requirement by traditional method [2]. Data access middleware based on distributed object technology can solve these problems, it can provide data service to multilayer application better [3].

The main distributed object technology supported components and middleware include CORBA(Common Object Request Broker Architecture, CORBA) [4] (Object Management Group), COM/DCOM(Microsoft),and EJB(Enterprise JavaBeans) (SUN). CORBA is a standard independence individual programming language. The system developed according to CORBA can run in almost of in common used machine and operating system. By the characteristic of CORBA, it is easy to construct multilayer distributed components application to solve the integration of data and application. The method of distributed hydrologic data integration this paper illustrated is based on CORBA.

2. DISTRIBUTED HYDROLOGIC DATA INTEGRATION

2.1 Distributed Hydrologic Data

Hydrologic data is the important proportion of water conservancy foundation database. It provides information service to water conservancy, government, and public by data service center. The collection and integration of hydrologic data accomplished by the hydrological telemetering station, center station, and control center. The hydrological telemetering system takes the role of collect data, center station, and control center take the role of date storage and management.

In hydrological telemetering system, the metering station is made up of sensors and RTUs (Remote Terminal Unit), in which RTUs connect network via wireless or wire channels and receive instructions and transfer data via network. Then the center stations receive the data from the metering stations and deliver the instructions from control center or send the instructions to the metering stations. Meanwhile, the center stations store the data into the local database and send them to the service requesters. The control center is responsible to

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2001AA113170 (National “863” Plan) and the innovation fund of Hohai University (2002402052/2003410343).

manage the whole system and provide information services to users. Generally, the center stations also need to provide information services to the local place.

2.2 Framework of Distributed Hydrologic Data Integration
As the distributing characteristic of Hydrologic data, distributed hydrologic data integration method composed of a series of software components. Based on CORBA service, these components cooperated accomplished data acquire, storage, and service. These components defined as CORBA component assembled by CORBA object, and provide service with the interface. (Figure 1)

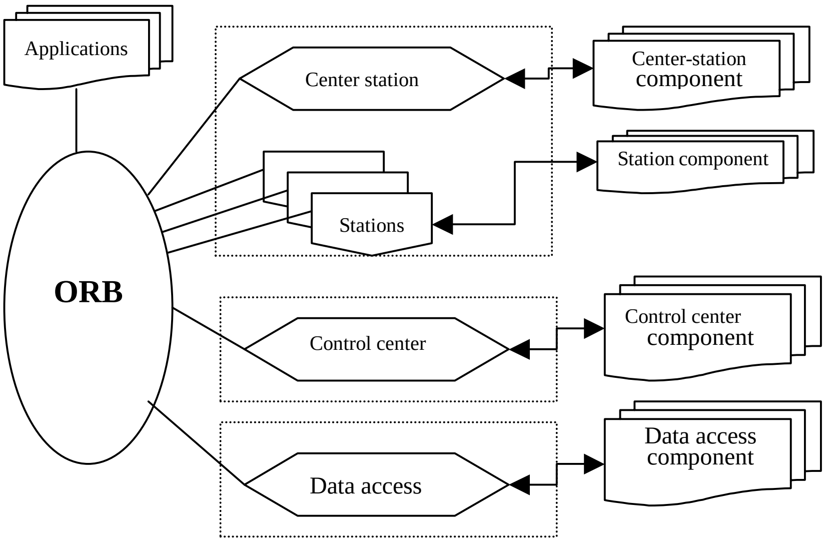


Figure 1. Sketch map of hydrological data integration based on CORBA framework

Hydrologic data collection accomplished by hydrological telemetering system. The basic structure of the six class basic component of hydrological telemetering system is the following:

Component = {functions, interfaces}

Functions are the function which the component provided, interfaces are the interface for the accessing according with the standard of CORBA.

The function and the interface of the six class service of hydrological telemetering system is defined as the following:

SACDA_components={functions={configuration, Calling }，interfaces={command，data}}

data_transfer_components ={ functions={ transmission }，interfaces={location, data, protocols}}

discharge_calculation_ components = { functions={ discharge calculation }, interfaces={ data}}

data_preprocess_components={ functions={ data preprocess }, interfaces={ data, rules choice }}

database_maintenance_components ={ functions={ insert, delete, update, backup, recover, synchronization, active push on/off}, interfaces={ data, database location }}

information_query_ components ={ functions={ retrieve data, presentation data}, interfaces={ data, query }}

In order to facilitate the construction of the application, according to the requirements of the function, the six class basic components can be assembled to the station components and the center station components.

The station components are simple，it is assembled by the SCADA service and the data transmission service, and packaged into components with standard interface:

Station_ component = {SACDA_ component, data_transfer_ component, station_interfaces}

The center station set the database to store the telemetering data and the state of the metering station, the structure of components is only need to compose the station components, and provide unify interface:

Center_station_component={Station_components, center_station_interfaces}

By the way to use the station components and center station components，it is easy to construct application. At the same time, by the composition of six classes' basic component, the user can get agility and advantageous service.

The management and service of the data dominated by data control center, these data access component include:

1) **Data monitor component:**
Used to monitor the equipment of hydrological telemetering system, if new data appears, they access in time.

2) **Data management component**
Get the data from equipment of hydrological telemetering system by the station components and center station components，then write the data into different database.

3) **Data service component**
Provide data service for different consumer by the data service interface.

The service provided by these components distributed in the server of hydrologic information network, the same function implemented by the same component, so the coherence and reliability of the data can be guaranteed.

3. DATA SERVICE

After the hydrologic data stored into the database in control center, the application can query directly. It distributed in the network environment, the developer can construct data access middleware by CORBA components; by shielding the different of data source, the middleware can implemented multi data source integration and service.

3.1 Data ServiceBased on Middleware

To implement heterogeneous data sources integration, it must be realized the entirely transparent accessing to different data sources. Not only ensure data sharing, but also ensure the autonomy of database management system, in order to guarantee the heterogeneous system access the heterogeneous data source by the unified API independent specifically database.

In domain application, the development of software components always goes along with domain data. The steady characteristic of domain knowledge is the precondition to construct a currency standard. Data Access Middleware (DAM) constructed in the middle of software components and data source can shield the difference of data source. The software components need not to care about the form and the location of the data in the database, they only contact with DAM, the actual data access executed by DAM. DAM defined and implemented by CORBA components. (Figure 2)

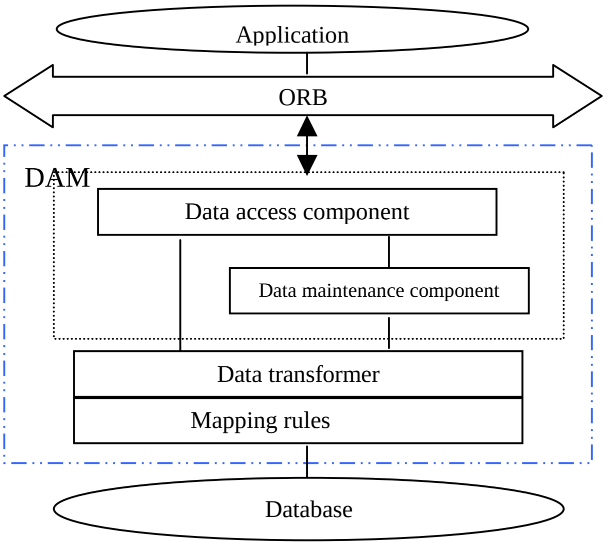


Figure 2 Data Access Middleware

DAM mainly includes four parts: Data access component, Data maintenance component, Data transformer, and Mapping rules.

Data Access Component

It is described with IDL (Interface Define Language), and provide a unify data view to the user. It encapsulates the data in database as the attribute of Interface. When the user query the data, it deliver the request to Data transformer, by the matching of data field, data transformer translate the request to SQL, after database management execute the SQL and get the result, the result would be translated and send to client through the transformer and data access component.

Some data access objects IDL define:

```
Module Db {
   exception DbException
  //define error message
  {string reason;};

  struct condition{
    // define query condition architecture
  }
  typedef sequence<condition> conditionSeq;
  struct selectAttr   {
    ….// define query select architecture
  }
      typedef sequence < selectAttr > selectSeq;
      Data interface   ……
  interface DataSet {
    // get the value of attribute number  currently object queried
string getString(in long index)   raises (DbException);
    // next record
boolean next()   raises (DbException);
       ……  //other operation ；
     };

interface DataObj{
     DataSet getDataSet(in selectSeq selseq,in conditionSeq conseq)
                   raises (DbException);
};
```

DataSet Object：According to DataSet Object, the result of data access queried returned, the object include some operation to data.

DataObj Object：To separate the data to different interface according to its contact and function. DataObj Object finishes the query, and returns a DataSet object to client.

Data Maintenance Component

It used by the data access component to accomplish the data maintenance, and respond to the data changes.

The main IDL define:

```
Module DbM {
  exception DbMException
  {string reason;};

  interface DbMaintain {
    void close()   raises (DbMException);
    void   execute(in   conditionSeq   conseq)   raises (DbMException);
       ……  //other operation；
  };
};
```

Data Transformer

It divided the query, across the rules take the each part corresponded to the fields of the data, and combine them to one or more SQL access to database, get the result returned to data access components with data set. Data access components access the database by call on the method of data transformer.

Mapping Rules

To define some rules in order to shield the semantic and the form difference of the data.

3.2 Active Data Service

Active Service is the technology which the clients get the necessary information without send out any instruction. Active service technology can be used to reflect the change of hydrologic data to application or write to different database. Hydrologic data active service can be implemented by the “Push” technology of CORBA event service.

Event service is one of the CORBA service，it import the communication concept of CORBA events. It define two roles for the object concerned with communications：Suppliers and Consumers [5] Suppliers produce event data; consumers receive and process event data. Event produced by one supplier, and transmitted to random number consumers. Suppliers and consumers are separated entirely: suppliers need not to know the number and ID of the consumers, at the same time; consumers need not to know the received events provided by witch supplier.

In order to support this pattern, CORBA event service import event channels witch is a middle object between suppliers and consumers. Event channels allow multi suppliers and consumers asynchronous communication without known each other. The suppliers and consumers connected with event channels are not interaction directly, they obtain agent object, event channels completed the event exchange over these agent. By this way, the event channels separate the suppliers and consumers.

Active data service in hydrologic data integration system can be implemented by “Push” technology of CORBA event service (figure3), this makes the new data or changing data enter database in time, and send to application and data service immediately.

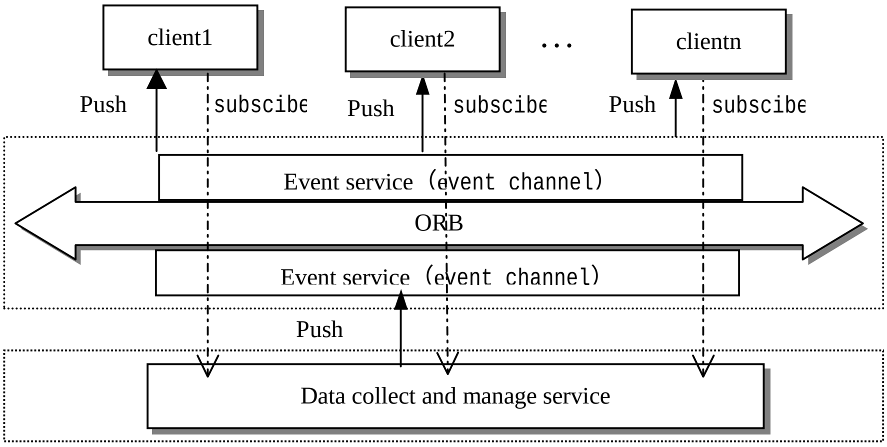

client1
client2
...
clientn
Push
subscibe
Push
subscibe
Push
subscibe
Event service（event channel）
ORB
Event service（event channel）
Push
Data collect and manage service

Figure3 Active hydrologic data service

At first, the client “subscribe” the necessary real time service, then the client can do its own work without care about when the “subscribe” event data arrive. The server side service would monitor the data changes, when pass muster data appeared, it “push” them to the client who “subscribe” over even channels.

The implement of active data service include function interface define, establish of suppliers and consumers, and event transmission. The suppliers and consumers need connected with event channels after they established.

4. CONCLUSION

The most important characteristic of water conservancy application is depended on domain data. The data with coherence, reliability, and betimes is the foundation for decision-making and project-designing.

The distributed hydrologic data integration system based on CORBA this paper presented has these characteristics:

(1) Expansibility
The system developed by components method based on CORBA makes it have better expansibility.

(2) Heterogeneous integration
Because the system based on CORBA standard can span different platform, this method is easy to integrate different system.

(3)Reliability
The adoption of uniform software for data access ensure the coherence of the data, the adoption of active data service makes the data transmitted to application or database in time.
This paper abstract the distributed hydrologic data integration system software by using CORBA, which would be valuable for building other water conservancy systems distributed in the networks.

5. REFERENCES

[1] AI Ping, Wang Zhi-jian, NI Wei-xin. The Application of Middleware Technology in Integration of Hydrological Telemetering System. HYDROLOGY, VOL: 22, 2002(6): 32-35

[2] WANG Ning,CHEN Ying,etc,Design of a Heterogeneous Data Integration System Based on CORBA．JOURNAL OF SOFTWARE 1998，5 vol.9

[3] Lou Yuan-Sheng, Yin Yan-Min, Wang Zhi-Jian，

Design and Implementation of Unify Data Access Middleware Based on CORBA，MINI-MICRO SYSTEMS，2001.10，vol.22

[4] Object Management Group. The Common Object Request Broker: Architecture and Specification, Revision 2.3. June 1999

[5] Object Management Group. CORBAservices: Common Object Services Specification, Updated December 1998. . 4-1~4-12

Lou Yuansheng is an Associate Professor of Computer and Information Engineering institute, Hohai University. He graduated from Hohai University in 1990, and get his bachelor's degree. He worked in Yellow River Conservancy Committee from Jul.1990 to Aug.1995. He studied in Hohai University from Sep.1995 to Dec.2001 as a graduate student, and obtained his doctor's degree with specialty of Computer Application. He is a member of Water Conservancy Academy. He has attended the study of two “863” Plan Project, and has finished the study and development of several important water conservancy domain projects.

Wang Zhijian is a Full Professor and superintendent of Computer Application graduate school, dean of Computer and Information Engineering institute, Hohai University. He graduated from Nanjing University in 1981. He works in Jiangsu Compute Technique Academe as an engineer from Jan.1980 to Agu.1983.He obtained his doctor's degree with specialty of Computer software in 1989.He is a syndic of Jiangsu computer Academy and Jiangsu software institute. He has presented papers more than 40, and published 4 monographs. As a dominant, he has finished many important projects in different domain.

Two Kinds of Novel Evolutionary Fuzzy Controllers* — Control Algorithm Analysis

Wang Pan[1,2*]**, Xu Chengzhi**[1]**, Zhang Jianjian**[1]**, Wan Junkang**[1]
[1]**Wuhan University of Technology, Wuhan, 430070, P.R. China**
[2]**Huazhong University of Science and Technology, Wuhan, 430074, P.R. China**
E-mail: jfpwang@tom.com **Tel:** 86-27-87858435

ABSTRACT

This paper presents two kinds of adaptive fuzzy control strategies based on evolutionary computation (EC). Principles, methods and steps of these two algorithms are analyzed. In these strategies, some key parameters of two self-regulated fuzzy controllers (a multi-regulated-factors fuzzy controller and qualitative-quantitative self-regulated fuzzy controller) are optimized by EC. Both linear and nonlinear quantization functions as quantized formula are employed and ITAE index is applied as fitness function.

Keywords: Evolutionary Computation, Complex Systems, Fuzzy Control

1. INTRODUCTION

Since 1980s, evolutionary computation (EC) which is composed of genetic algorithm (GA), evolutionary strategies (ES), evolutionary programming (EP) have been employed in natural science and social science deeply and widely [1]. Unique advantages have being showed and many achievements have been made with the aid of EC.

Fuzzy control is a widely accepted control scheme, in which objectives are controlled by fuzzy set theory on the basis of human experience and knowledge. For its prominent feature of being independent of mathematical models of controlled objectives, fuzzy control may offer an effective approach to the control problems with uncertainty. Furthermore, fuzzy control is robust to the variant parameters of controlled objective, so it can be applied to a large range of nonlinear or time-variant systems [2].

Some research has been made to optimize controller by combining fuzzy control with EC especially GAs. There are three ordinary strategies in evolutionary-fuzzy control: (1) regulating fuzzy rules with EC [3] (2) regulating membership functions with EC [4] (3) regulating both rules and membership functions with EC [5].

Because of a larger number of parameters to be optimized, some of strategies have great computational complexity. Two novel self-regulating adaptive fuzzy control strategies (a multi-regulated-factors fuzzy controller and qualitative-quantitative self-regulated fuzzy controller) are presented in this paper based on EC. In such strategies, some key parameters are optimized by an adaptive EC presented by authors. Both linear and nonlinear quantization functions as quantized formula are employed and ITAE index is applied as fitness function. Simulated results illustrate these control strategies have satisfactory dynamic, steady and robust performance in MIMO systems, chaotic systems and delayed systems.

2. SYSTEM STRUCTURE AND STRATEGOES

2.1 The control system structure

The structure of the control system is shown in **Fig.1**.The relationship between controller and controlled object is shown in **Fig.2**. Generally, the algorithm also fits m-input-n-output situation)

Where $\mathbf{R}=(r_1, r_2, \cdot\cdot\cdot, r_n)^T$; $\mathbf{Y}=(y_1, y_2, \cdot\cdot\cdot, y_n)^T$; $\mathbf{e}=(e_1, e_2, \cdot\cdot\cdot, e_n)^T$; $\mathbf{ec}=(ec_1, ec_2, \cdot\cdot\cdot, ec_n)^T$; $\mathbf{E}=(E_1, E_2, \cdot\cdot\cdot, E_n)^T$; $\mathbf{EC}=(EC_1, EC_2, \cdot\cdot\cdot, EC_n)^T$; $\mathbf{NQ_1}=(NQ_{11}, NQ_{12}, \cdot\cdot\cdot, NQ_{1n})^T$; $\mathbf{NQ_2}=(NQ_{21}, NQ_{22}, \cdot\cdot\cdot, NQ_{2n})^T$; $\mathbf{Ku}=(Ku_1, Ku_2, \cdot\cdot\cdot, Ku_n)^T$. r_i, y_i are ith input and output, e_i, ec_i, E_i, EC_i are the errors and error rates of input branches before and after quantization respectively, NQ_{1i}, NQ_{2i}, ($i = 1,2, \cdot\cdot\cdot, n$) are correspondent nonlinear quantization factors.

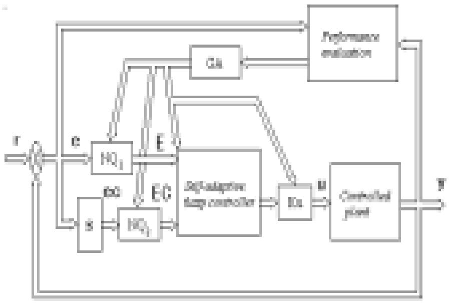


Fig.1 Structure of Control System

* This work was partly supported by NSF of China (Grant No.: 60174039)

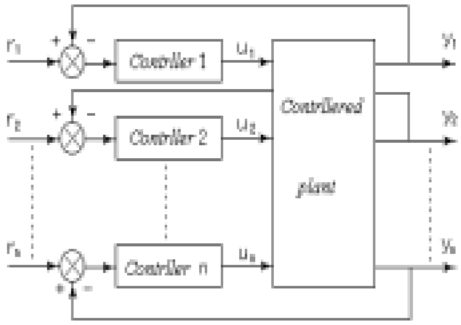


Fig.2 Relation Diagram Of Controllers And Controlled Plant

2.2 Quantization factor and adaptive fuzzy control algorithm

Conventional fuzzy controller quantizes error and error rate linearly. The quantizing formulas are illustrated as follows:

$$Ke = \frac{n}{Xe} \tag{1}$$

$$Kec = \frac{m}{Xec} \tag{2}$$

$$E = INT(e * Ke + 0.5) \tag{3}$$

$$EC = INT(ec * Kec + 0.5) \tag{4}$$

In the past studies, we present a novel nonlinear quantization method. In this method, variations of quantized value of e and ec increase as absolute values of e and ec decrease, which has been proved it can improve control performance effectively [6,7]. The correspondent formulas are as follows:

$$NQ_{1i}：E_i(x_i) = l_1 \frac{e^{\lambda_i \cdot x_i} - e^{-\lambda_i \cdot x_i}}{e^{\lambda_i \cdot x_i} + e^{-\lambda_i \cdot x_i}} \quad \ldots\ldots\ldots\ldots\ldots(5)$$

$$NQ_{2i}：EC_i(x_i^*) = l_2 \frac{e^{\mu_i \cdot x_i^*} - e^{-\mu_i \cdot x_i^*}}{e^{\mu_i \cdot x_i^*} + e^{-\mu_i \cdot x_i^*}} \quad \ldots\ldots\ldots\ldots\ldots(6)$$

Where $E_i \in [-n，n]$, $EC_i \in [-m，m]$, $X_i \in [-Xe,Xe]$, $X_i^* \in [-Xec, Xec]$, l_1、l_2 λ_i、μ_i can be calculated by the following equations:

$$E_i(Xe_i) = n，E_i(\frac{Xe_i}{2}) = a；$$

$$EC_i(Xec_i) = m, EC_i(\frac{Xec_i}{2}) = b；$$

If $n=m=6$ and $a=4,b=4$ (obviously both a and b are more than 3):

$$l_1 = l_2 = 4\sqrt{3},\ \lambda_i = \frac{\ln(2+\sqrt{3})}{Xe} \quad \mu_i = \frac{\ln(2+\sqrt{3})}{Xec}$$

For the adaptive fuzzy controller, two strategies are adopted independently, we respectively name strategy 1 and 2 as multi-regulated-factors fuzzy control and qualitative-quantitative self-regulated fuzzy control:

Strategy 1 multi-regulated factors fuzzy control

$$U_i^{'} = \begin{cases} < c_{0i}E_i + (1-c_{0i})EC_i > + U_{0i} & E_i = 0 \\ < c_{1i}E_i + (1-c_{1i})EC_i > + \ U_{1i} & E_i = \pm 1 \\ < c_{2i}E_i + (1-c_{2i})EC_i > + U_{2i} & E_i = \pm 2 \\ < c_{3i}E + (1-c_{3i})EC_i > + U_{3i} & E_i = \pm 3 \end{cases} \tag{7.1}$$

Where $n=m=3$, c_{ji} is the jth quantized factor of the ith controller and $\boldsymbol{U}_0 = (u_{1i}, u_{2i}, u_{3i})^T$ is the steady control value. Obviously, $c_{ji} \in [0，1]$. All c_{ji} composed the multi-regulated-factors of the fuzzy controllers. Formula (7.1) can be seen in some references, but it's hard to select parameters especially for multivariable systems. In our work, a novel EC algorithm is employed for searching "optimal" or "satisfactory" parameters.

Strategy 2 Qualitative-quantitative self-regulated fuzzy control:

$$U_i = \frac{\beta_{1i}|E_i|}{\beta_{1i}|E_i| + \beta_{2i}|EC_i|} E_i + \frac{\beta_{2i}|EC_i|}{\beta_{1i}|E_i| + \beta_{2i}|EC_i|} EC_i + U_{0i} \tag{7.2}$$

Where β_{1i}, β_{2i} are weights of E_i and EC_i . U_{0i} is the steady state control value. $\beta_{ji}(j=1,2;\ i=1,\cdots\cdots,n)$ should be variant with time. For simplicity, we fetch β_{ji} fixed values to the situation of $E_i * EC_i \geq 0$ （trend to the equilibrium point） and $E_i * EC_i < 0$ (leave the equilibrium point).

From the above algorithm we presented, it is evident that the controlling value is composed of quantized errors, error rates and steady state control value. The weights of E and EC are adaptive as their own relative values.

3. EVOLUTIONARY OPTIMIZATION

3.1 chromosome and mechanism of coding

For **Strategy 1**:Chromosome is composed of the following gene strings:

{ Ku Xe Xec U$_0$ C$_0$ C$_1$ C$_2$ C$_3$ }

For **Strategy 2**:Chromosome is composed of the following gene strings:

{ Ku Xe Xec U$_0$ β$_1$ β$_2$ }

Where $\beta_1 = (\beta_{11}, \beta_{12} \cdot\cdot\cdot \beta_{1n})$, $\beta_2 = (\beta_{21}, \beta_{22} \cdot\cdot\cdot \beta_{2n})$ are the values of β_{1i} while $E_i * EC_i \geq 0$ and $E_i * EC_i < 0$ respectively. Parameter-genes could be added or deleted from the chromosome when necessary. Each chromosome is coded by four-digit-decimal or floating points. Such disposal has the advantages of dwindling search space, improving the search efficiency over conventional method — optimizing the parameters set in control table.

Assume the length of the controller's chromosome is L_i, the total length is：

$$L = L_1 + L_2 + \cdots\cdots + L_n = \sum_{i=1}^{i=n} 4m_i \tag{8}$$

3.2 Fitness function

Performance criterion is significant in evaluating the effect of control systems. In our research, weighted ITAE is adopted as the main performance criterion.

The fitness function is defined as follows:

$$\begin{gathered} ITAE = \int_0^\infty t\left|\alpha \bullet E^T\right| dt \\ \alpha = (\alpha_1, \alpha_2, \cdots \alpha_n) \in R^{1\times n} \quad , E \in R^{l\times n} \end{gathered} \tag{9}$$

Where E is *n*-dimensional a vector，whose elements are the errors of input branches; α are *n*-dimension weighted vector of errors.

After discretization:

$$ITAE = \sum_{i=1}^{n} \sum_{j=1}^{K} \alpha_i e_i t_j \Delta t \tag{10}$$

Where *K* is terminated step.

The fitness function adopted in this paper is as follows:

$$f(x) = c_1 + (c_2 - c_1)\frac{(ITAE)_{\max} - ITAE(x)}{(ITAE)_{\max} - (ITAE)_{\min}} \tag{11}$$

Formula (11) represents the *ITAE*-value of individuals is mapped into $[c_1,c_2]\in[0,1]$ in term of the rule of "the smaller，the better". And (*ITAE*) max 、（*ITAE*）min are the max-value and min-value of each generator while *ITAE*(*x*) means the value of individual of the present generator.

3.3 Evolutionary operators

Pc and Pm are two key parameters of GA. In conventional GA Pc and Pm are determined through experience and the same for all individuals. In our newly adaptive GA, Pc and Pm for each individual may be different and are closely related to fitness value of each individual. The basic principle is: 1) Pc and Pm are set large to evolve the "bad" individuals when fitness value is small; 2) Pc and Pm are set small to prevent the "good" ones when fitness value is large; 3) The more the value of *f (x)* is, the steeper the variance of Pc and Pm is. Consequently, the whole evolution process is accelerated. We adopt the following adaptive possibility function:

$$P_c = \begin{cases} k_1 - k_2 \exp\{k_3(f - \bar{f})\} & f \ge \bar{f} \\ k_4 & f < \bar{f} \end{cases} \tag{12}$$

$$P_m = \begin{cases} k_1' - k_2' \exp\{k_3'(f - \bar{f})\} & f \ge \bar{f} \\ k_4' & f < \bar{f} \end{cases} \tag{13}$$

Where $k_1, k_2, k_3, k_4, k_1', k_2', k_3', k_4'$ are constants to be determined; $P_{c_1}, P_{c_0}, P_{m_1}, P_{m_0}$ are the maximum and minimum of crossover probability and mutation probability respectively. We set $K_4 = P_{c_1}, K_4' = P_{m_1}$. $\bar{f}$ is the average fitness.

$$\Theta \quad \left.\begin{aligned} k_1 - k_2 \exp\{\, k_3(f_{\max} - \bar{f})\} &= Pc_0 \\ k_1 - k_2 &= Pc_1 \end{aligned}\right\}$$

$$\therefore Pc_1 + k_2 - k_2 \exp\{k_3(f_{\max} - \bar{f})\} = Pc_0 \tag{14}$$

and $\quad \Theta \quad k_1 - k_2 \exp\{k_3(\frac{f_{\max} - \bar{f}}{2})\} = Const_1$

it follows that:

$$Pc_1 + k_2 - k_2 \exp\{k_3 \frac{f_{\max} - \bar{f}}{2}\} = Const_1 \tag{15}$$

we obtain

$$k_2 - k_2 x^2 \overset{x=\exp\{k_3(f_{\max}-\bar{f})/2\}}{=} Pc_0 - Pc_1 \tag{16}$$

$$k_2 - k_2 x = Const_1 - Pc_1 \tag{17}$$

Formula (16) divided by (17) , then we get

$$\frac{k_2(1-x^2)}{k_2(1-x)} = \frac{Pc_0 - Pc_1}{Const_1 - Pc_1} \tag{18}$$

simplifies formula (19),we have

$$1 + x = \frac{Pc_0 - Pc_1}{Const_1 - Pc_1} \tag{19}$$

that is

$$x = \frac{Pc_0 - Const_1}{Const_1 - Pc_1} = \frac{Const_1 - Pc_0}{Pc_1 - Const_1} \tag{20}$$

From the definition of *x*, we have

$$k_3 = \frac{2}{f_{\max} - \bar{f}} \ln \frac{Const_1 - Pc_0}{Pc_1 - Const_1} \tag{21}$$

Moreover, Θ k_1-k_2=Pc_1， we have the following results:

$$k_2 = \frac{1}{1-x}(Const_1 - Pc_1) = \frac{-(Const_1 - Pc_1)^2}{Pc_1 - Const_1 - Const_1 + Pc_0} = \frac{(Const_1 - Pc_1)^2}{2Const_1 - Pc_1 - Pc_0} \tag{22}$$

$$k_1 = Pc_1 + k_2 = \frac{Const_1^{\,2} - Pc_1 Pc_0}{2Const_1 - Pc_1 - Pc_0} \tag{23}$$

Base on the same method, k_1', k_2', k_3' can be calculated.

For simplification, we set $k_3 = k'_3$ =1, then：

$$k_1 = \frac{\exp\{f_{\max} - \bar{f}\}\, p_{c_1} - p_{c_0}}{\exp\{f_{\max} - \bar{f}\} - 1} \tag{24}$$

$$k_2 = \frac{P_{c_1} - P_{c_0}}{\exp\{f_{\max} - \bar{f}\} - 1} \tag{25}$$

$$k'_1 = \frac{\exp\{f_{\max} - \bar{f}\}\, p_{m_1} - p_{m_0}}{\exp\{f_{\max} - \bar{f}\} - 1} \tag{26}$$

$$k'_2 = \frac{P_{m_1} - P_{m_0}}{\exp\{f_{\max} - \bar{f}\} - 1} \tag{27}$$

Elitist reservation and immigration: in each generation, we copy the 2%~5% best chromosomes directly to the next generation, without the operations of reproduction, crossover and mutation. Meanwhile, 20%~40% worst individuals are deleted and randomly selected new individuals are inserted into the population. In this way, the system is open, compared with the closed population of conventional genetic algorithm. By such disposal, the possibility of finding the optimum is obvious improved.

4. CONCLUSIONS

This paper presents two kinds of adaptive fuzzy control strategies based on evolutionary computation (EC). Principles, methods and steps of these two algorithms are analyzed. In next paper, empirical studies are fulfilled on these two algorithms' tracking ability, robustness, and anti-disturbance.

5. REFERENCES

[1] Z. Michalewicz, Genetic Algorithm + Data Structure=Evolution Programs, Berlin: Heidelberg: Springer-Verlag, 1996.

[2] C. C. Lee, "Fuzzy Logic In Control Systems: Fuzzy Logic Controller-Part Ⅰ & Ⅱ", IEEE Trans. Syst., Man, Cybern, Vol. 20, No.2, 1990,pp. 404-435.

[3] T. C. Chin, X. M. Qi, "Genetic Algorithms For Learning The Rule Base of Fuzzy Logic Controller", Fuzzy Sets And Systems, No. 97, 1998,pp. 1-7.

[4] S. Patric, G. Francois: "A Genetic Algorithm For Optimizing Takagi-Sugeno Fuzzy Rule Bases", Fuzzy Sets And Systems, No. 99, 1998, pp. 37-47.

[5] C. France, L. Richard, "Constraining The Optimization of A Fuzzy Logic Controller Using A Enhanced Genetic Algorithm", IEEE Trans. Syst.,Man,Cybern-Part B, Vol. 30, No.1, 2000, pp. 31-46.

[6] Wang Pan, *et al*, "A New Fuzzy Neural Network Control Algorithm Based on Nonlinear Quantitation and Simulation Studies", Proceeding of ICAIE'98, Wuhan: HUST Press, 1998.

[7] Wang Pan, *et al*, "Study on a New Qualitative-Quantitative Adaptive Fuzzy Control Algorithm", Chinese J. Systems Engineering And Electronics, Vol. 20, No.11, 1998, pp. 55-57.

[8] Jin Qibing, Gu, Shisheng, "Parameter-Converged Rapidly Neural PID Control for Multivariable Systems", Chinese J. Control and Decision, Vol. 13(Suppl), 1998, pp. 448-452.

[9] Fang Jian'an., Shao Shihuang, "Control of A Kind of Chaotic System Using Genetic Algorithm and Fuzzy Logic", Proceedings of Industrial Technology, 1996, Shanghai.

Wang Pan is a Full Associate Professor and a head of Institute of Control and Decision, Wuhan University of Technology. He received the B.S. degree in industrial automation from Wuhan University of Technology, Wuhan, P. R. China, and the M.S. and Ph. D. degrees in systems engineering from Huazhong University of Science and Technology, Wuhan, P. R. China. He has published over 30 Journal papers, 15 Conference papers. His research interests are intelligent control, decision analysis, and biomedical intelligent information systems.

A Study of the Mixed Fuzzy PID Controller in the Accurate Orientation

Chen yunji Shen keyu
Wuhan University of Technology Wuhan 430063
Email: Chenyj@mail.whut.edu.cn

ABSTRACT

The door-style crane plays an important role in the construction of hydroelectric station. At the same time, it has high degree of difficulty. This paper analyzes the control system of the accurate orientation of the door-style crane, and put forward that the fuzzy PID structure composed of the common integral controller and the double dimension fuzzy controller. In order to test and verify the validity of the fuzzy control strategy and the feasibility of the fuzzy controller, this paper set up the math model about the real system, and use MATLAB and SIMULINK situation simulate tools to group the frame chart of the fuzzy control system, completes the design and simulation of the whole system, gives out the computer simulate result.

Keywords: fuzzy controller position error transfer function Laplace converts

1. INTRODUCTION

“The double-orientation crane” is the key in the construction of hydroelectric. As a important facilities. The operation of all the available crane usually depends on the driver’s experiences in our country today. Because of the parallax and the long distance between the sluice and the driver’s cab, the orientation of the sluice is not enough accurate, so increase the difficulty of operation, especially the lift hook of the crane is operated under water, consequently need to improve supervision about the state of the run. About the requirement of the accurate orientation of the crane and the lift hook, we design the fuzzy PID structure composed of the fuzzy controller and the PLC control system that control the output of the frequency converter, achieve the accurate orientation of the crane and the lift hook. In order to test and verify the validity of the fuzzy control strategy and the feasibility of the fuzzy controller the paper set up the maths model about the real system, completes the design and the simulation of the whole system, evaluates the result of the simulation, provides the dependable information data for study, analysis, design and adjust the real system.

2. CHOICE OF THE SCHEME OF CONTROL [1]

The accurate orientation of the crane and the under-water operation of the lift hook are both influenced by the outside natural condition (for example the direction and speed of wind, the direction, speed and pressure of water), the gradient of orbit and the operational experience of the driver, however all of the factor maybe change anytime. So it is difficult to input the function data to make the system high accuracy. If use the common PID control, it is difficult to reach the ideal accurate orientation.

When we conceptualize the scheme of control, proposals the fuzzy PID structure. The position error should be found out between the crane’s position on the big vehicle’s orbit and required accurate position when the big vehicle is oriented accurately, the position error is directly related with the output of the frequency is higher, same, the output of the frequency converter is related with the position change, at the same condition of error, if the rate of error change is greater, the speed of frequency converter will be greater.

According to the above-mentioned analysis, it is validity to use the fuzzy PID control structure in quality. The paper designs the scheme of the fuzzy control about the accurate orientation of the big vehicle. After the simulation by computer, the result indicates that the design scheme is validity.

In this system, the input variable of the fuzzy controller is the position error and the rate of error, the output variable is the frequency of the frequency converter.

3. FUZZY PID CONTROLLER [2][3]

In the common fuzzy control system, considering the simple and the high-speed of the realization of the fuzzy controller, it is usual to use the double dimension fuzzy controller. The input variable of this controller are system is error E and error change

EC ,consequently it has the effect of proportion and calculus, but it has not the effect of integral, the same as the effect of the common PD controller. In addition, the fuzzy controller has exclusively the process of quantitative. So the system that uses this fuzzy controller will has well situational character, however can’t remove the static error.

According to the theory of linear control, integral control can remove the stable error, but the situational respond is slow, the proportional control has a fast situational respond. However proportion-integral control not only has a high stable precision, but also has a high situational respond, consequently introduce PI control strategy into fuzzy controller and compose the compound control of fuzzy –PI(or PID). Improve the situational character of the fuzzy controller. But now, this compound controller has many structures. All of the structures have good effect of removing surplus error. However only the structure that mixed PID controller can remove the vibration of limit, this mixed PID controller is composed of the common integral controller and the double dimension fuzzy controller. The output of the common PI controller $u_{\mathrm{i}} = \mathrm{K}_{\mathrm{I}}\sum_{\mathrm{i}} e_{\mathrm{i}}$ add the output of the double dimensions u_{f} to make up the output of the output of the mixed controller, it means $u = u_{\mathrm{i}} + u_{\mathrm{f}}$. This system is called the no-error fuzzy control system. In order to improve the stability of the fuzzy, the double dimensions controller uses the structure and the parameter of the common controller, introduces the integral in chart 2.1, $K_I = 0.02$.

4. REALIZATION AND SIMULATION OF THE BIG VEHICLE’S RUN [4]

4.1 Founding of the systemic model

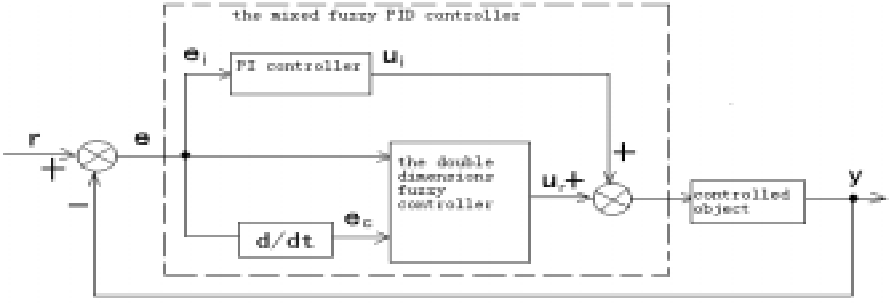
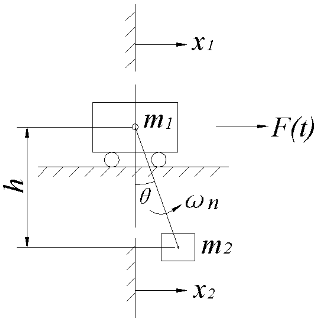

the mixed fuzzy PID controller
e_i
PI controller
u_i
r
+
e
the double dimensions fuzzy controller
u_f +
+
-
d/dt
e_c
controlled object
y

Chart2.1 the mixed fuzzy PID controller When the big vehicle (or the small vehicle) starts or stops, the hanged goods will swing. Chart 3.1 is the model of the run of the big vehicle (or the small vehicle), supposes that F_m is resistance (from friction and wind) when the big vehicle (or the small vehicle) , F_d is the drive force , F_Z is brake force. a is the acceleration, F(t) is the general force when it starts or stops. From the second law of Newton, gives out the below equation:

$$\begin{cases} m_1\ddot{x}_1 - \dfrac{m_2 g}{h}(x_2 - x_1) = -F(t) & (4.1) \\ m_2\ddot{x}_2 + \dfrac{m_2 g}{h}(x_2 - x_1) = 0 & (4.2) \end{cases}$$

Model of mechanics



m_1 ——the general mass of the big vehicle (include transform mass of organization) ;

m_2 ——the mass of the hanged goods ;

x_1 、 x_2 ——the displacement of m_1, m_2

h ——the height of the hanged goods ;

F（t） ——the general force ;

start state： $F(t) = F_d - F_m$;

brake state： $F(t) = -(F_Z + F_m)$ ；

θ ——the swing-angle when the goods swings ;

g ——gravity acceleration, g =9.81 m/s^2。

4.2 Discussing about the movement of the goods when the crane starts or brakes

the relationship of the swing-angle θ and x_1、x_2：

$$\theta = \frac{x_2 - x_1}{h}$$

（when the vibration is little, θ is small, $\theta \approx tg\theta$） (4.3)

the above formula（3.2） multiplies m_1 / h and the formula（3.1）multiplies m_2 / h, then the difference of them is :

$$\ddot{\theta} + \frac{(m_1 + m_2)g}{m_1 h}\theta = \frac{F(t)}{m_1 h}$$

(4.4)

$$\therefore F(t) = m_1 h\ddot{\theta} + (m_1 + m_2)g\theta$$

(4.5)

Supposes that the drive force made by the electric machine (or the brake force made by the brake) and the resistance is constant （F(t) is constant）, So the calculus formula（4.4）is the equation of the system excited by the input, the result of this equation is :

$$\theta = \frac{F(t)}{(m_1 + m_2)g}(1 - \cos\omega_n t) \qquad (4.6)$$

ω_n is the constant frequency of the goods：

$$\omega_n = \sqrt{(1 + \frac{m_2}{m_1})\frac{g}{h}} \qquad (4.7)$$

From the formula（4.6）, if $m_1 << m_2$, $\omega_n \approx \sqrt{\frac{g}{h}}$.

Provides that in the design of the big crane , the best acceleration $a \le 0.1$m/s^2 ,when it starts or stops, from the formula（4.5） ,gives out the most biggest vibration angle：

$$\theta_{\max} = \frac{2F(t)}{(m_1 + m_2)g} = \frac{2a}{g} = \frac{2 \times 0.1}{9.81} = 1.16^\circ$$

Considering the start of the crane，from the Newton Formula，supposes that the crane's speed is V_d , so the drive power P is：

$$P = F(t) \cdot V_d = [m_1 h\ddot{\theta} + (m_1 + m_2)g\theta] \cdot V_d$$

(4.8)

Additional，supposes that the output torque of the electric machine is T_d , the output angle speed is ω_m , gear ratio is i , the diameter of the wheel is D，the output power of the electric machine is：

$$P = T_d \cdot \omega_m = T_d \cdot i \cdot \frac{2}{D} \cdot V_d$$

(4.9)

：

$$T_d \cdot \omega_m = F(t) \cdot V_d = [m_1 h\ddot{\theta} + (m_1 + m_2)g\theta] \cdot V_d$$

(4.10)

the torque of three-phase asynchronous generator T_d is：

$T_d = K_m \cdot \phi \cdot I$ （K_m is torque constant of the electric machine, ϕ is magnetic flux, I is electricity current of rotor）

if $K_d = K_m \cdot \phi$, $T_d = K_d \cdot I$

（K_d is constant ）

formula（4.8）will be ：

$$[m_1 h\ddot{\theta} + (m_1 + m_2)g\theta] \cdot V_d = K_d \cdot I \cdot \omega_m$$

$$\therefore \qquad \ddot{\theta} + \frac{m_1 + m_2}{m_1 h} g\theta = \frac{K_d}{m_1 h} \cdot \frac{\omega_m}{V_d} \cdot I$$

(4.11)

through the Laplace convert, formula (4.11） will be:

$$S^2\theta(s) + \frac{m_1 + m_2}{m_1 h} g\theta(s) = \frac{K_d}{m_1 h} \cdot \frac{\omega_m}{V_d} \cdot I$$

(4.12)

so:

$$\frac{\theta(s)}{I}=\frac{\frac{K_d}{m_1h}\cdot\frac{\omega_m}{V_d}}{s^2+\frac{m_1+m_2}{m_1h}g}=\frac{K_d\cdot\frac{\omega_m}{V_d}}{m_1hS^2+(m_1+m_2)g}$$

$$=\frac{\frac{K_d}{m_1}\cdot\frac{\omega_m}{V_d}}{hS^2+\frac{m_1+m_2}{m_1}g} \qquad (4.13)$$

the transfer function $H(s)$ of the linear system expressed by the formula（3.13）is :

$$H(s)=\frac{K_1}{T_1S^2+T_2} \qquad (4.14)$$

4.3 Discussing the movement of the big (or small) vehicle

Observing the movement of the big(or small) vehicle in the braking process, the general force F（t）acted on the big(small) vehicle（includes the brake force F_Z an the resistance F_m）reverses the movement , from formula（4.3）and formula（4.6）,gives out:

$$\frac{x_2-x_1}{h}=\theta=\frac{F(t)}{(m_1+m_2)g}(1-\cos\omega_n\, t) \qquad (4.15)$$

introduce formula（3.15）into formula（4.1）, gives out:

$$m_1\ddot{x}_1-m_2g\,\frac{F(t)}{(m_1+m_2)g}(1-\cos\omega_n t)=-F(t)$$

（4.16）

so:

$$\ddot{x}_1=-\frac{F(t)}{(m_1+m_2)}(1+\frac{m_2}{m_1}\cos\omega_n t) \qquad (4.17)$$

Supposing that the brake torque is T_Z, gear ratio is i ,the diameter of the big(or small) vehicle is D，the transfer efficiency is η , from the relationship formula $F(t)=-(F_Z+F_m)$, gives out ：

$$F(t)=-(T_Z\cdot i\cdot\frac{2}{D}\cdot\eta+F_m)$$

If don't consider the resistance F_m from frication and wind,so $F(t)\approx -T_Z\cdot\frac{2i\eta}{D}$,introduces into formula（4.17）, gives out：

$$\ddot{x}_1=\frac{T_Z\cdot\frac{2i\eta}{D}}{m_1+m_2}(1+\frac{m_2}{m_1}\cos\omega_n t) \qquad (4.18)$$

however

$$T_Z=K_Z\cdot I_Z$$

（K_Z—the coefficient of brake，I_Z—the brake electricity current）

Laplace convert formula（3.18） into：

$$S^2x_1(s)=\frac{K_Z\cdot I_Z\cdot\frac{2i\eta}{D}}{m_1+m_2}(1+\frac{m_2}{m_1}\cdot\frac{S}{S^2+\omega_n{}^2}) \qquad (4.19)$$

$$\Longrightarrow$$

$$\frac{x_1(s)}{I_Z}=\frac{2i\eta K_Z}{(m_1+m_2)D}\left[\frac{1}{S^2}+\frac{m_2}{m_1}\cdot\frac{1}{S(S^2+\omega_n^2)}\right]$$

（4.20）

transfer function is ：

$$H(S)=\frac{2i\eta K_Z}{(m_1+m_2)D}\left[\frac{1}{S^2}+\frac{m_2}{m_1}\cdot\frac{1}{S(S^2+\omega_n^2)}\right]$$

$$=\frac{2i\eta K_Z}{(m_1+m_2)D}\cdot\frac{S^2+\frac{m_2}{m_1}S+\omega_n^2}{S^4+\omega_n^2S^2} \qquad (4.21)$$

write in the common formula：

$$H(S)=\frac{K_2S^2+K_1S+K_O}{T_2S^4+T_1S^2} \qquad (4.22)$$

the same , can infer the transfer function of the big(or small) vehicle when it starts as the above.

5. SIMULATION OF THE MODEL

In Matlab, Simulink realizes the model and the simulation of the system from the single fuzzy PID control to the whole fuzzy control, first, makes the visual model, sets up the visual model of the fuzzy control system, then simulating and adjusting.

This model includes no-linear link, calculus link, integral link and a common double dimensions fuzzy controller, opening the corresponding model warehouse in Simulink，introduce into Continuous、Nonlinear、Fuzzy Logic Controller.

The below gives three results respective about disturbance , no-disturbance and random-disturbance.（unit is second in horizon, uni is millimeter in vertical）.

transfer function of the controlled object $\frac{40S^2+4S+0.01}{350S^4+S^2}$ 。

5.1 Condition of no-disturbance

Ignoring the influence of the outside condition (speed ,direction of wind)，chart5.1 and chart 5.2 is the simulative result of the fuzzy control system.

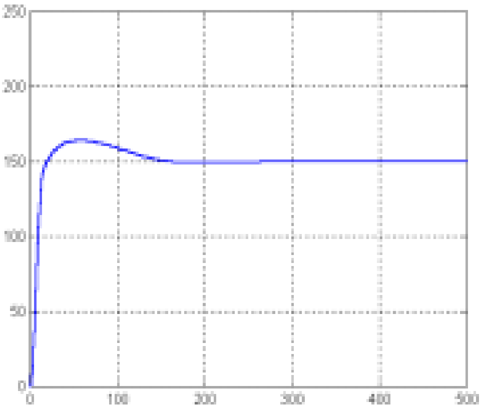


Chart 5.1 the curve of the big vehicle

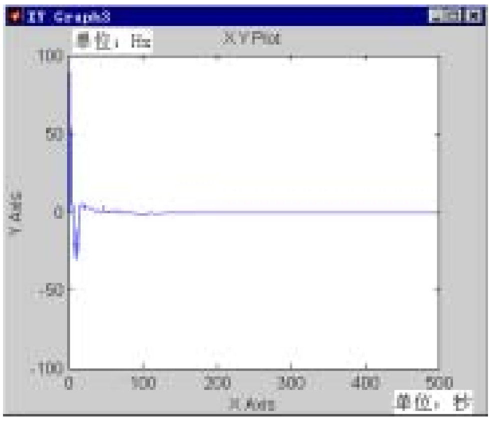


Chart 5.2 the output u curve of the fuzzy controller

5.2 Condition of disturbance

Considering the influence of the outside condition, when adds 10% disturbance into the fuzzy control system, the simulative result is express by the chart 5.3 and chart 5.4 at the condition of disturbance.

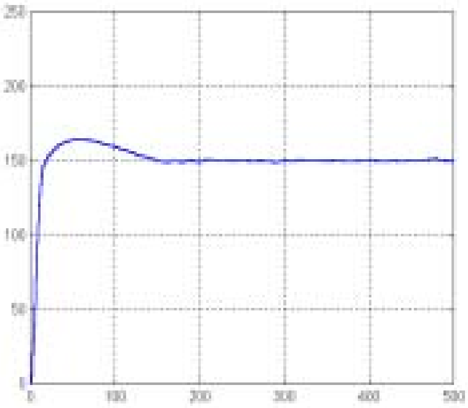


Chart 5.3 the curve of the big vehicle

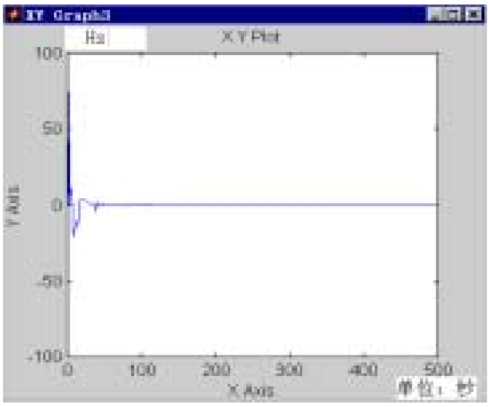


Chart 5.4 the output u curve of the fuzzy controller

5.3 Condition of random-disturbance

Considering the influence of the random-disturbance, adds random-disturbance into the fuzzy control system. chart 4.5 and chart 4.6 express the simulative result of the system at the condition of random-disturbance.

6. ANALYSIS AND STUDY OF THE EXPERIMENT

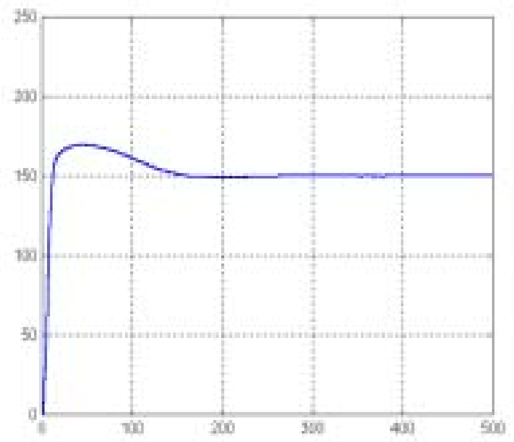


Chart 5.5 the curve of the big vehicle

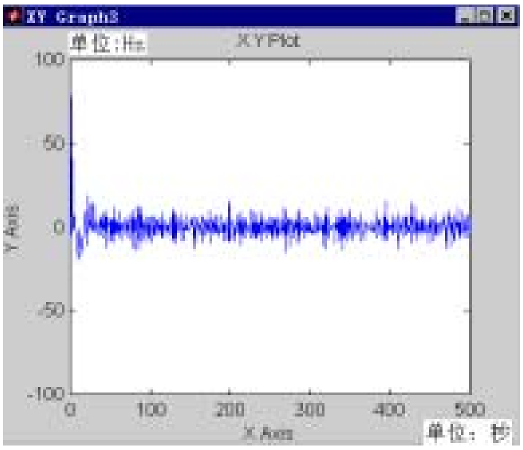


Chart 5.6 the output u curve of the fuzzy controller

(1) the structural parameter of the fuzzy controller has great influence on character of the controlled object, the variable of languages should be suitable. If the language is less, the density of the membership function is less in domain, the response is not sensitive probably. But if the language is too more, the calculation time will be longer more, and the control character can't reach best, when assumes the membership function, it should pay attention that the amplitude width make great influence on the character of the control, the little amplitude width membership function has a high control sensitive, and the great one has a flat control character and a high stability. If the system error is great, uses the little amplitude membership function, if not, uses the great one. When deciding the control rule, if error is great, the control quantity should reduce the error as possibly as it can, if error is little, the stability of the system should be considered, besides removing the error, in order to remove the overshoot and the vibration.

(2) the influence of quantizational factor and proportional factor on system, when fixing K_1、K_2 ,changing K_3, founds out that K_3 has a great influence on the output of the system, if K_3 become greater, the overshoot of the system will become greater, the situation process becomes quick, on the contrary, the overshoot will be littler, the adjustable range of K_3 is small, the situation process become slow, and the response of the system is sensitive to the change of K_3, K_1 对 has little influence on the output comparably. Fixing K_3、K_2, if K_1 become great, the stable error will reduce, on the contrary ,it will increase, if K_1 is too great, the system will vibrate, K_2 can be adjusted among large range, but it has little influence on the output of the system, it's the most insensitive to the change of the system.

(3) the choice of the integral parameter has great relationship with the time constant of the object, the constant is more, K_I should be littler, the effect of the integral is weak. If the object has hysteresis, the effect of the integral is more weak, this is according to the parameter adjustment rule of the PID controller.

7. CONCLUSION

From the result of the above simulative experiment, it gives that the fuzzy PID controller is satisfied with the requirement of the system, does't have great overshoot, response is quick, overcomes the disturbance, reaches the ideal control object, verifies the validity of the fuzzy control strategy and the feasibility of the fuzzy controller.

8. REFERENCES

[1] Deng xingzhong, Zhou zhude, Deng jian. Transmission control of the machine and electricity Wuhan. Publisher of huazhong university of technology. 1998

[2] H.Banndemer and S.Hartluann. A fuzzy approach tostability of fuzzy controllers, Fuzzy Sets and systems, 96(1998):161-172.

[3] Udo Kuln. Ein Praxisnahe Eistellregel fuel PID-Regler: Die T-Summen-Regel. Automatisier umgste chmische Praxis,1995,37(5)

[4] Hu zongwu, Yan yisong Dynamics of crane. Beijing Publisher of the machine industry. 1988

Codesign for Complex Hard Real-time Embedded Systems

Jin Yongxian
Institute of Computer Science studies, Zhejiang Normal University, Jinhua 321004, China
E-mail: jk78@mail.zjnu.edu.net.cn; jinyongxian@mail.zjnu.net.cn

ABSTRACT

Due to increased system complexity and pressure to reduce development times, the development of long lifetime hard real-time embedded systems is becoming increasingly difficult. This paper presents a hardware-software co design framework of three-phase process. The first phase develops constraints (in terms of time, resource usage etc.) that are placed upon a subsequent system implementation, produces non-functional design decisions using trade-off analysis method between different non-functional properties. Secondly, system functions are generated by use of high-level modeling tools (e.g. UML, Matlab) to increase automation and higher-level abstraction within the development process. Thirdly, low-level implementation performs relatively conventional hardware-software codesign to map the functions onto a platform until meeting the non-functional requirements. The method proposed by this paper takes advantage of increased automation for long-lifetime hard real-time embedded systems, until ensuring system timing predictability and amenability to change.

Keywords: High-level Design; Low-level Implementation; Real-time embedded systems; Codesign; Timing predictability; Constraints

1. INTRODUCTION

In terms of their functional and non-functional properties, real-time embedded systems are becoming increasingly complex, so making their design and implementation evermore difficult. However, systems need to be developed in shorter times, due to business requirements to reduce the time-to-market. Such conflicting pressures are often addressed by increasing the automation within the development process, e.g. by utilizing high-level modeling tools (UML, Matlab, MatrixX etc.) and utilizing the automatic software generation facilities within those tools for system software production. Effectively, the abstraction level at which most of the system is developed is raised from the software level to a modeling level.

In order to reduce system time-to-market, the general approach of high-level specification and process automation is extremely attractive. It can be seen in much hardware-software codesign research, which enables automatic derivation of a hardware architecture and application software from a high-level specification [1~3]. Such approaches are (1) limited in terms of the scale of the system that can be developed; (2) limited traceability from specification to final design (due to automation); (3) limited ability to change /update parts of the system with ease at some later date.

Many real-time embedded systems are developed for domains that have additional constraints than those assumed by traditional codesign work. A current research trend is concentrating upon the codesign of complex long-lifetime hard real-time embedded systems [4]. These systems have a number of important requirements: (1) timing predictability is a key. Failure to meet timing requirements (e.g. deadlines) can result in catastrophic failure of the system; (2) the system must be shown to be fit-for-purpose prior to use. Often some area (e.g. aerospace, nuclear, medical) requires documentary evidence that both the development process and the system are sufficiently robust and correct before the system can be used; (3) the system must be amenable to change /upgrade. This must be carried out in manner that minimizes the impact of the change upon the rest of the system, to simplify the process of convincing the regulators that the changed system has not introduced any unexpected problems.

This paper is organized as follows. Section 2 introduces important concepts. Section 3 presents development framework. Section 4 analyses and discusses characteristics of this method. Section 5 concludes the paper.

2. CONCEPTS

Traceability [5] — relates to "the feasibility of reviewing and identifying the source code and library component origin and development processes" thus facilitating verification and validation techniques, which are essential aids to ensuring program correctness.
Timing predictability — namely ability to forecast the execution time of a real-time program (or task), i.e. an estimation in advance satisfying time correctness.

WCET (worst-case execution times) — defined as "the maximum probable value of execution time for real time program (or task)".

WCET analysis[6]— computes upper bounds for the execution times of pieces of code for a given application, where the execution time of code is defined as the time it takes the processor to execute that piece of code.

Ada/ SPARK/ Ravenscar — Ada language facilitates the programming of real-time systems. It contains facilities for small programming or large programming, together with facilities for concurrent programming (i.e. tasks and inter-task communication). The SPARK [7] subset of Ada restricts the sequential part of the language. The SPARK Ada subset is consistent with the requirements for real-time system timing analysis in that all conforming programs are statically analyzable for their worst-case properties. The Ravenscar tasking profile [8] is a statically analyzable tasking subset. Ravenscar compliant code is predictable in its timing behavior and resource usage.

3. CONDESIGN FRAMEWORK OVERVIEW

This codesign framework of hard real-time embedded systems is broken into three phases, as shown in Fig.1.

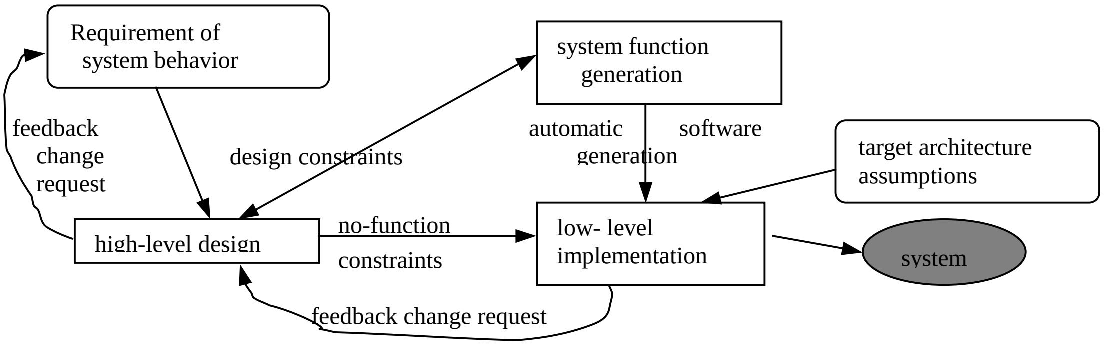


Fig.1 development framework

High-level Design and trade-off analysis captures design choices in a structured manner to aid traceability and provide supporting evidence of the system being fit-for-purpose, whilst providing automatic optimization of key non-functional properties to ensure that non-functional requirements will be met in the final system. This phase develops constraints (time, allocation, resource usage etc.) that are placed upon a subsequent implementation of the system.

1. System function generation utilizes appropriate modeling techniques (e.g. UML, MatrixX, Matlab) for the modeling of desired functional behaviors. Software to implement these functions can be automatically generated from these tools. Where the modeling techniques available are not sufficient to express required functions, manual software development can occur using a suitably rigorous software development process.

2. Low-Level Implementation produces a hardware architecture that supports the functions generated whilst meeting the non-functional constraints generated by the high-level design phase. Restricted codesign techniques for automatic hardware and software production are used. Non-functional requirements are considered by the high-level design process to order to establish a set of constraints and requirements that must be met by the low-level implementation. This process can feed back any recommendations for change to the requirements process to the high-level modeling process if it finds contradictions or inconsistencies. Fig.1 shows the system function generation phase that automatically generates the application functions in a manner that can then be taken by the low-level implementation phase.

3.1 High-level Design and Trade-off Analysis Process

The high-level design and trade-off analysis process shows in Fig.2.

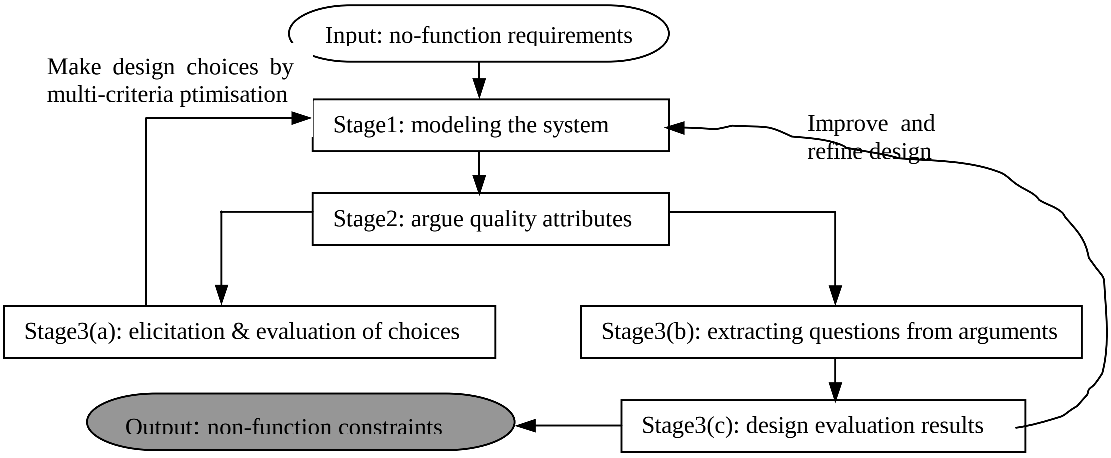

Input: no-function requirements
Make design choices by multi-criteria ptimisation
Stage1: modeling the system
Improve and refine design
Stage2: argue quality attributes
Stage3(a): elicitation & evaluation of choices
Stage3(b): extracting questions from arguments
Output: non-function constraints
Stage3(c): design evaluation results

Fig.2 high-level design and trade-off analysis method

Stage.1 is producing a model of the system to be assessed. This model should be decomposed to a uniform level of abstraction using UML, however it could be applied to any modeling approach that clearly identifies components and their couplings (A coupling is considered as a connection between components.). Arguments are then produced in stage.2 for each coupling to a corresponding abstraction level. The arguments are derived from the top-level properties of the particular system being developed. The properties often of interest are cost, dependability, and maintainability. Clearly these properties might be broken down further, e.g. dependability may be decomposed to reliability, safety, timing etc. Stage.3 then uses the information in the argument to derive options and evaluate particular solutions. Part of this activity uses representative scenarios (e.g. what happens when change X is performed) to evaluate the solutions.

Based on the findings of stage.3, the design is modified to fix problems that are identified – this may require stages.1~ stages.3 to be repeated to show the revised design is appropriate. When this is complete and all necessary design choices have been made, the process returns to stage.1 where the system is then decomposed to the next level of abstraction.

Components reused in other context could be incorporated as part of the decomposition. Only proceeding when design choices and problem fixing are complete is preferred to allowing trade-offs across components at different stages of decomposition because the abstractions and assumptions are consistent easing the multiple-criteria optimization problem.

The process allows the derivation of design choices, identifying where different solutions are available for satisfying a key system requirement, managing the sensitivities /dependencies between components and design decisions. The process also identifies the constraints that must be placed upon functional component design. Such constraints are passed to the system function generation process as they are found. The high-level design process is also the recipient of constraints from the system function generator. These constraints include the functional properties that must be considered by the high-level analysis process during the development of the non-functional design. For example, the number of processes and functions must be accounted for during timing analysis. Finally, the high-level analysis process collects all the design rationale (i.e. the design choices, sensitivities, dependencies and design decisions) into a repository to aid traceability. This is particularly important if changes to the system need to be made during the lifetime of the system, e.g. for planned updates or major revisions sometime after the system has been initially deployed.

The iterative nature of the high-level analysis process is used to develop the design, in terms of further decomposition of the design. Eventually, when sufficient design development has occurred, the low-level implementation phase can be utilized. A key aspect of the high-level design process is that many decisions made will constrain the eventual system implementation and architecture. For example, if during high-level design it is determined that redundancy is required to meet fault-tolerance requirements, and then this will be specified to the low-level implementation phase.

3.2 System Function Generation

The system function generation phase encompasses the mapping of functional requirements to implementation. This is usually achieved using appropriate modeling tools (e.g. UML, Matlab, MatrixX), that permit automatic generation of an implementation, as represented in a high-level language such as C, Ada, VHDL etc. The resultant “programs ”can be passed to the low-level implementation phase. Whilst the generation of the program is automatic, the use of the tools themselves is manual. Many of the modeling tools include model-level testing and simulation of the model (i.e. model execution), which aids verification that the model is meeting functional requirements.

The current realization of the system function generation phase is limited to tools that can produce Ada (including UML, MatrixX, Matlab). It is noted that the limitations on language are largely imposed by the current scope of the low-level implementation. A key part of this phase is the iteration of constraints with the high-level design phase, as described above. This phase is responsible for identifying functions that need to be executed (and perhaps upper and lower bounds on some timing properties); the high-level design phase is responsible for assigning execution times etc.

3.3 Low-Level Implementation

The low-level implementation phase takes as inputs the constraints and requirements established by the high-level phase, together with the application software. The compilation of Ada to binary (i.e. the software route) utilizes the GNAT Ada compiler [9]. The compilation to hardware is achieved using the hardware Ada compiler [10,11].SPARK /Ravenscar conformant Ada programs are ideal for direct compilation to hardware circuit. In [10,11] an Ada compilation process is described for such programs. Essentially, concurrency within Ada can be represented on hardware as truly parallel tasks. In terms of the Ravenscar tasking subset, the main implication is that task scheduling is no longer required.

The low-level implementation phase allows the mapping of functions expressed in Ada (developed by the system function generation phase), constrained by the non-functional requirements established by the high-level design phase, into an actual hardware and software implementation. Although a number of design decisions have been taken during the high-level design phase, there is still considerable freedom for the low-level implementation phase to search a wide range of potential solutions.

Currently, the physical target architecture assumed is that of a single Field Programmable Gate Array (FPGA), coupled to a number of RAM banks. Clearly, limiting the target architecture to a single FPGA restricts the solution space. However, the physical size of current high-end FPGAs is large, ensuring that substantial functionality can be achieved on a single device. Also, the presence of the RAM banks ensures that (parts of) the FPGA can be used for soft-core CPUs, further extending the size of the functionality that can be implemented upon the target.

The low-level implementation phase follows a timing analysis driven approach. The actual implementation of the system is then checked against the assumptions of the model. If the assumptions still hold (e.g. that the WCET of a software task is no more than some value), then the full implementation will meet its timing requirements.

The method is illustrated in Fig.3. It consists of an iterative process with two main parts. (1) Modelling and simulating the timing and interaction properties of the software. (2) Compilation to hardware circuit and CPU instructions of a given allocation of the software. These stages provide feedback in terms of timing characteristics of the actual software (e.g.WCET of software tasks, or circuit speed and size of an FPGA task). Perform timing analysis and simulation of the system. This is sufficient for the system configuration, in terms of the allocation of tasks to hardware or software, to be evaluated. As a consequence, a new allocation can be determined to further improve the system.

This phase is able to analyze the software for timing characteristics (e.g. WCET) and find a suitable platform on which the application software can execute to meet its timing requirements. Note that the low-level implementation phase is not necessarily restricted to implementations on CPU, it can also consider direct mapping to hardware (if permitted by the target architecture assumptions). The low-level phase is entirely automatic.

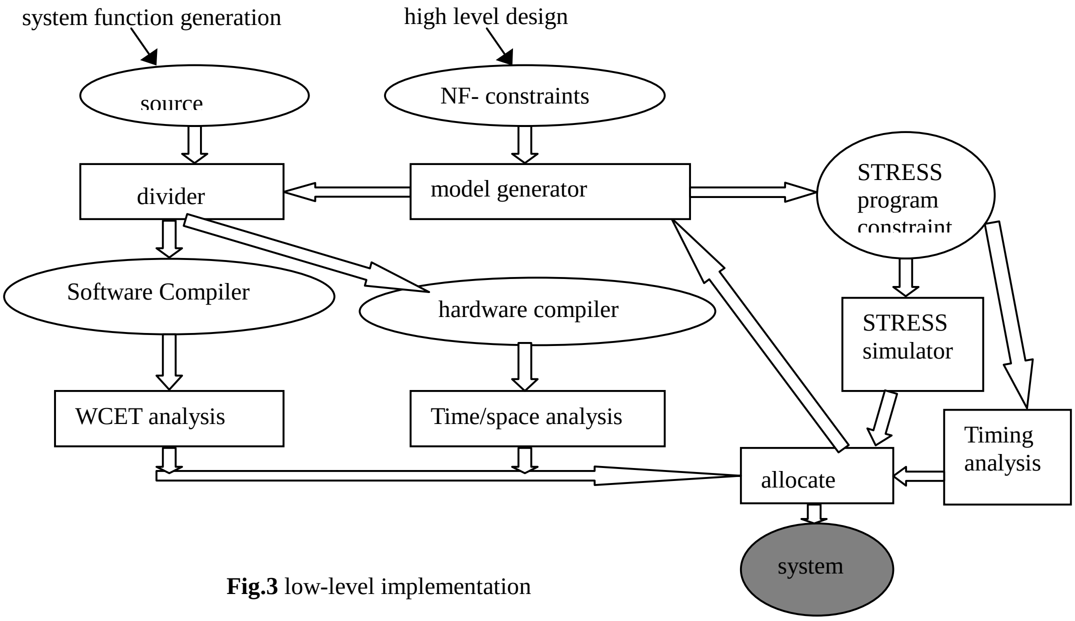


Fig.3 low-level implementation

4. ANALYSIS AND DISCUSION

Conventional codesign approaches assume that a complete specification is available prior to system generation. To some degree, this is also seen in the process given above, where a complete set of requirements is required prior to the commencement of design. However, in realistic large hard real-time embedded system developments, the precise specification is often not readily available until late in the development. Normally, the high-level design process and the modeling have started before a total specification is available.

Conventional codesign approaches usually assume a single process for development. This is not usually the case for large hard real-time embedded systems. It is important that the overall process described above is amenable for use by a subcontractor building part of a system (e.g. a sub-system), a prime contractor assembling the entire system, or a sub-contractor contributing either software (e.g. by some system model).

Conventional codesign approaches utilize an automatic partitioning of functionality between hardware and software implementation. This is adopted in the low-level implementation phase of the process outlined above. Here, functions expressed in a high level language (e.g. Ada) are mapped to a combination of logic and CPU, utilizing hardware compilers that map programs in a high-level language such as Ada, to circuit (i.e. FPGA).

Conventional codesign approaches assume a co-verification phase as part of the iterative search during system generation. In the overall process described above, verification occurs in many areas. As part of the high-level design process, key non-functional requirements are verified as part of iteration towards a design solution, e.g. timing. As part of the system function generation phase, functional properties will be verified. This occurs at the model level where appropriate. During low-level implementation, further verification of properties is performed during the iterative search for an implementation solution.

5. CONCLUSIONS

This paper has described a process, which utilizes codesign techniques within the development process for complex hard real-time embedded systems. The motivations for codesign techniques includes the structured capture of non-functional design decisions in a more structured and integrated manner than current practice suggests; technology independent design is encouraged, which postpones decisions regarding the target technology until late in the development process. Combined with automatic mapping of system functions to target architecture within the constraints imposed by the non-functional requirements, this provides a better process for complex hard real-time embedded system development. The process ensures that key non-functional properties are met by the design and eventual implementation. Importantly, the process is driven from a timing analysis perspective, closely integrating static timing analysis within the process. This imposes correctness by construction approach, rather than the build and test approach seen often in practical developments.

The key part of codesign that is adapted throughout the overall process given above is the automatic tradeoff design choices, particularly in the non-functional domain. This is seen in the high-level design and analysis phase where timing properties can be optimized via trade-off analysis. The low-level implementation phase automatically finds a solution to meet the non-functional and functional designs generated by the high-level design process and system function generation phases respectively.

6. REFERENCES

[1] K. Suzuki, A. Sangiovanni-Vincentelli. Efficient Software Performance Estimation Methods for Hardware /Software Codesign. In:Proc. Design Automation Conference,1996

[2] D. Gajski, F. Vahid, S.N. J. Chong. System-Level Exploration with SpecSyn. In:Proc. Design Automation

Conference, 812~817, 1998

[3] J. Henkel, R. Ernst. A Hardware/Software Paritioner Using a Dynamically Determined Granularity. In:Proc. Design Automation Conference, 691~696, 1997

[4] Michel Lemoine, Jack Foisseau. Modelling Long Lifetime Systems:Building a Referential with UML. http://www.cert.fr/fr/dprs/publications/tic.pdf, 2000.9

[5] J. Kwon, A. Wellings and S. King. Ravenscar-Java: A High Integrity Profile for Real-Time Java. Proceedings of the Joint ACM Java Grande - ISCOPE 2002 Conference

[6] I. Bate, G. Bernat and P. Puschner. Java Virtual-Machine Support for Portable Worst-Case Execution-Time Analysis. Fifth IEEE-ISORC'02, 83~90, May 2002

[7] J. Barnes. High Integrity Ada:The SPARK Approach.Addison-Wesley, 1997

[8] A. Burns, B. Dobbing and G. Romanski. The Ravenscar Tasking Profile for High Integrity Real-Time Programs. In: Reliable Software Technologies, Proceedings of the Ada Europe Conference, 263~275, 1998

[9] Ada Core Technologies:GNAT Ada Compiler. http://www.gnat.com, 2001

[10] M. Ward, N.C Audsley. Hardware Compilation of Sequential Ada.In:Proceedings of CASES, 99~107, 2001

[11] M. Ward, N.C Audsley. Language Issues of Compiling Ada to Hardware.In:Proceedings of Ada Europe 2002

Development of a Distributed Embedded Remote Control Monitor System Based on CAN Bus

Qin Juanying, Feng Xin, Wu Guoping
College of Automation, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: qinjuany@mail.whut.edu.cn, wgp62000@163.com **Tel:** +86 (0)27 87850114

ABSTRACT

This paper introduces a distributed embedded remote control monitor system based on CAN bus. A embedded operation system was ported in the control system. We used the embedded operation system to manage the switch between CAN bus protocol and TCP/IP protocol. The paper not only focus on the hardware design of the system, but also especially on the software design of the system. The detail of CAN bus protocol and the program flow diagram is exhibited, as well as the performance of the system working. The generalization of this system is finally outlined.

Keywords: Remote Control Monitor, Embedded System, Distributed, CAN Bus

1. INTRUDUCTION

With the rapid development of computer technology in hardware，software and integrated circuit, industrial control systems have become one of the activated branches and made huge progress in the field of computer technology applications. Because of the lofty requirements of system reliability and system flexibility, the development of industrial control system mainly represents: **control many-entity, system dispersion, namely load disperses、function disperses、risk disperses and terrene disperses**. Distributed industrial control systems just adapt such requirements Systems are the outcome using **micro-controller as kernel** and these 5C techniques—COMPUTER、CONTROL、COMMUNICATION、CRT and CHANGE are key issues for them.

Nowadays, the Internet has made progress at a high speed and become an important manner in the communication. And large-range monitor and control module has become more and more mature technology based on Web B/S (Browser/Server). The users can practice large-range monitor and control into industrial field equipment whenever the Internet has been switched-in if applying the module into industrial field control. The dominance speaks for itself.

In this paper, a distributed embedded large-range monitor and control system is introduced which is integrated by a network distributed control system based on CAN bus and a embedded system with the function of network. It not only can control the industrial field process real timely and safely but can connect industrial control field with management information system (MIS) wholly. It has been applied to the remote control monitor of the communication multi-inverters successfully.

2. EMBEDED SYSTEM

The Embedded System was defined: it is a special computer system that is as the center of application, as the base of COMPUTER, whose hardware and software can be reduced, which adapt the strict requirements of function、reliability、cost、capacity、power consumption. Today, embedded system has been used in every field that from automobile、housed microwave oven、PDA、television to industrial productive field、communication、device、instrument、automobile、ship craft、aviation、astronavigation、military accoutrement、consumer goods and so on .

The simple embedded system is not use operation system, which only includes some control flow. But the simple control flow cannot meet the system requirement when the functions have been flexed (such as image consumer interface and network supporting) that are supplied by embedded system.。Then we use operation system as system software. So the Embedded Operation System emerges as the requirement of times.

In this system, we extend CAN bus interface based on Embedded kernel plate. The whole system is called Embedded long-range monitor and control system (uCremm). It is based on the peripheral device interface such as Motorola company's ColdFire5272CPU, 2MFLASHMEMORY 、 32MDRAM 、 UART and ETHERNET network card .CAN bus control slug uses Infineon 82C900. SPI interface supports CAN2.0B.

Embedded operation system uCliunx reduces Memory Management Unit (MMU) based on Linux 2.0 version. Moreover we transplant and debug the HTTP server software BOA that supports CGI (Common Gateway Interface).

The monitor and control system plate connects PC with RS232 series line and twisted pair line, which makes up the circumstance of development that can cross compile. Its compiler language is standard C. The progresses of development are compiling uCilnux kernel in host PC and loading it into uCremm plate, through uCremm MOUNT host 's application development catalogue. Then we can develop embedded application in the circumstance of PC Linux. At last we compile the kernel including application and erase and write to uCremm's FLASH MEMORY. By now we complete the monitor and control system's software development.

3. CAN BUS

CAN bus is a serial data communication protocol that was developed by German BOSCH company to settle numerous data exchange between control and testing instrument of modern automobile. It is a multi-bus. Its communication can be twisted pair line、coax or optical fiber. Its communication rate can be 1MBPS. CAN bus communication interface integrate the CAN protocol functions of physics and data link layer. It can complete the frame process of communication data including bit fill、data block compiling、cyclic redundancy check、priority discrimination and so on

The most merit of CAN protocol is that it abolishes the conventional station address code and use communication data block code as substitute. The merit of this method is that the number of node is limitedness in the network in theory. The identification code of data block can be made up of 11 bits or 29 bits binary number, so we can define 2^{11} or 2^{29} differ data blocks. This method of data block code is very useful in distributed control system because it can make the different nodes receive the same data at the same time. The length of the data segment can satisfied with the common requirement of control command、active state and test data in the usual industrial field that has 8 bytes at the most. At the same time, the 8 bytes may not hold the bus too long, so it can ensure the communication real timely. CAN protocol adopts CRC check and apply corresponding error handling function, which ensure data communication reliability. The excellent characteristic 、higher reliability and unique design of CAN bus make it suits the interconnection of industrial procedure monitor and control instruments. So CAN bus gets more and more regards in industry and is thought of one of the most outlook bus.

Furthermore, CAN bus adopt multi-competitive bus structure, which makes it has the characteristics of multi-operation and distributed arbitrage serial line and broadcast communication. The random node in CAN bus can send information to other network nodes without primary and secondary at any time, which can realize free communication in every node. CAN bus has been authenticated by International Organization for Standardization. Its technique is relative mature .Its control slug has commercialized performance/price ratio. All the characteristics make it appropriate for data communication among the distributed testing system purposes.

4. SYSTEM CONFIGURATIONS

The chart of the system structure is shown in Figure1.

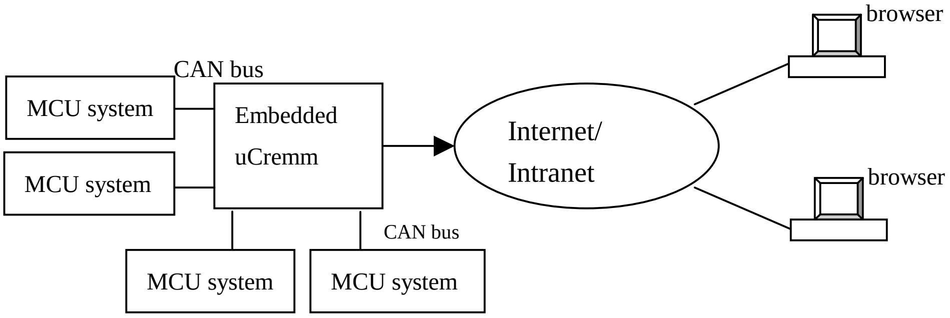


Figure1. System structure

Embedded uCremm and field data gathering node make up of the relationship of Master and Slave. The master station sends out inquiry information to authorized MCU unit according to authorized user in order to realize the inquire function of industrial control field node. The data structure reference CAN bus protocol. Every CAN node monitors bus at any time. When the node discoveries there are the same frames in bus address field as its address, it will judge these are either remote frames or data frames then. CAN node will delivery data to bus according to the protocol if these are remote frames. Otherwise Can node will receive data from CAN bus. This is the understructure flow. The network server of embedded system completes the functions of the interconnection of embedded uCremm and Internet and the interaction with user. Embedded uCremm allocates Ethernet interface RJ45. The authorized users can knock-in IP address in the browser of any computer in Internet and call on the principal branch that stored in Embedded uCremm after uCremm and Internet have been linked by gauze wire, allocating IP address and operating network server.

The interactive between users and webserver is fulfilled by CGI. CGI (Common Gateway Interface) is considered simply as a program running on the webserver. It is activated by browser. The CGI scenarios are a bridge between the client and the program e.g. Database of the server. Customers can inquire the data from the control field for instances the voltage, current and power of each inverter module or transmit the command to set the parameters of the max input DC and output AC value under hitting the coherent homepage.

5. THE LOWER LEVEL CAN BUS PROTOCOL AND THE PROGRAM FLOW DIAGRAM

There are currently three CAN protocols-CAN2.0A， CAN2.0B, and CAN2.0B passive. The Infineon device supports CAN2.0B.The difference between these protocols lies in the length of message identifier they can transmit and receive in a message frame. A CAN2.0B controller can transmit standard frames and extended frames with 29-bit identifiers. Finally, CAN2.0B passive controllers can only transmit standard frames but can receive both standard and extended frames. For the majority of today’s applications CAN2.0B is considered standard, with system designers often requiring the extended 29-bit identifier to relieve them from compromises with respect to defining well-structured naming schemes. The backward-compatible nature of the CAN protocol ensures the Infineon device can also handle messages with the standard frame format.

During the system design the data packages transferring on CAN bus are defined standard frames while extended frames are not handled. Standard frames consist of two types, one is remote frames the other is data frames. The difference between them is whether the RTR bit is set.

The data of structure of standard frame are presented in Table1.

The bottom CAN bus module is in charge of receiving the data from the distributed DC-to-AC inverters and transmitting the data set the each inverter parameters. When CAN controller transmits a remote frame on CAN bus commanding the inverter upper transmitting the parameters, the five bits of high order in the 11 bits ID of remote frames are defined the number of each inverter, similarly the MsgID register of CAN controller of each inverter must be also set correspondingly. In addition the RTR of frame format is set and the DLC is set “0”.

A brief example of remote frame is presented in Table2.

Table1. Structure of standard frame

Start bit	Arbitration region	Control region	Data region	Parity region	End bit
	11bits ID	Remote transmit request bit(RTR)and 4bits data length code(DLC)	8Bytes data	16bitsCRC	

Table2. An example of remote frame

Arbitration region（D15~D5）	Control region(D4~D0)	Command code	Inverter Module number
00001000000	10000	0810H	01
11110000000	10000	F010H	30

Table3. A data structure example of inverter parameters

Descriptor(arbitration region＋control region)	Data region(8 bytes D7…D0)	Inverter module number
0808H(0000100000001000)	D7:Input max DC voltage D6:Input min DC voltage D5、D4:Output max AC voltage D3、D2:Output min AC voltage D1:Output max AC D0：module temperature	01
F008H(1111000000001000)	……	30

When CAN controller transmits a data frame on CAN bus to set the inverter parameters, except of the ID set the RTR of frame format is set “0” and the DLC is set”1000” indicating a data frame have eight bytes data in data region.

Table3 is a brief example shown the data structure when CAN controller transmits the inverter parameters.
When CAN controller transmits data frames to set the on-off status of inverter modules, the information of on-off status is expressed by a bit following the command code. When data value is equal to “0FH” is on status; when data value is equal to “00H”is off status.

Besides transmitting data information control system also receives the data information from each inverter module including warning message frames and the data frames of the inverter parameters.

Figure2 shows the program flow of the main station (uCremm) (SHOW IN NEXT PAGE).

6. WEB SEVER CONFIGURATIONS AND HOMEPAGE DESIGN

After the uClinux and HTTP server software-BOA portal on the embedded uCremm. These steps as follows are made. First, the configuration file of BOA is configured. Second, the homepages and CGI programs used by the control monitor system are located at the specified path of BOA. The CGI program is written by C language, in fact these are some applied programs which function are transferring CAN communication data etc. Finally, the application programs are activated by means of hitting the homepages showing in browsers, thereby the browser and server can get the interactive function. The work principle of the whole system is showed in Figure3. (SHOW IN NEXT PAGE).

After the control monitor system is set up successfully customers are able to hit the address of the embedded uCremm IP in the site field of browsers and the system returns the interface shown in Fighure4. One of control monitor interfaces is showed in Figure5 until the user name and user password is correctly filled in.

Making use of homepages to show the control monitor interface changes the conventional monitor method-C/S (Client/Server) to the advanced monitor method-B/S (Browser/Server), which not only reduces the expenditure of purchase or exploit the software of superior PCs but also make the remote control monitor cast off the limitation of terrene to be a absolute "remote" monitor. In a sense the B/S method will have enormous commercial value with the Internet technology overreach.

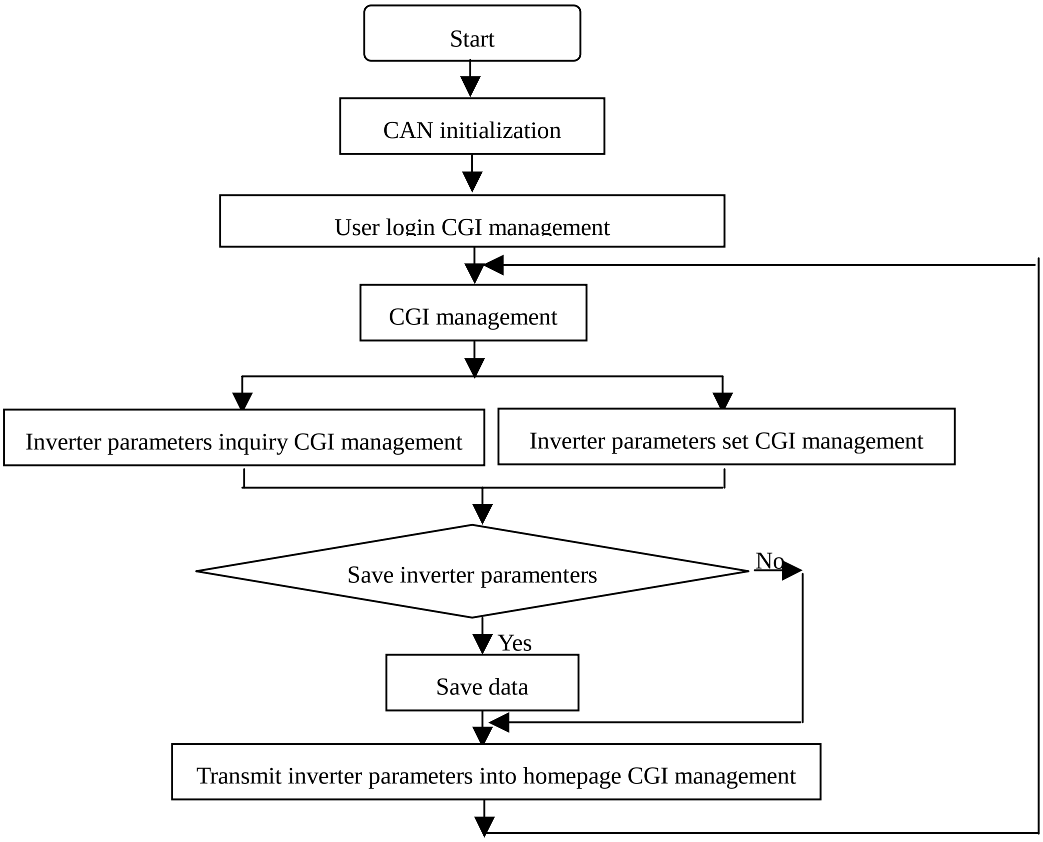

Start
CAN initialization
User login CGI management
CGI management
Inverter parameters inquiry CGI management
Inverter parameters set CGI management
Save inverter paramenters
No
Yes
Save data
Transmit inverter parameters into homepage CGI management

Figure 2. Program flow diagram of the main station

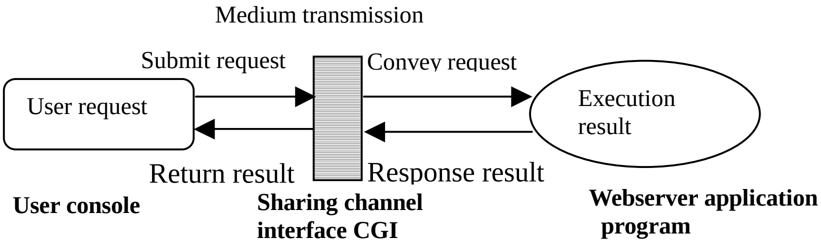
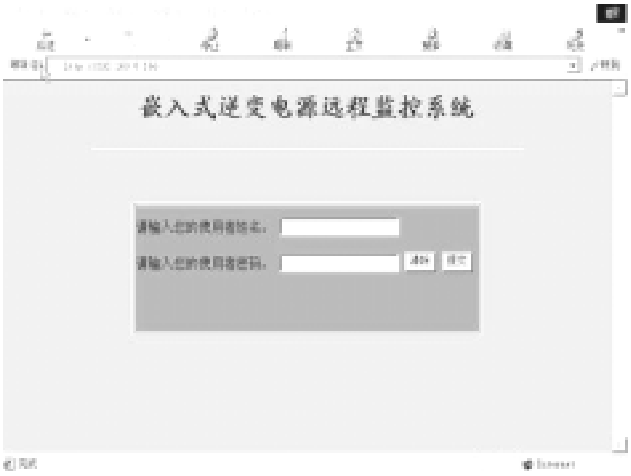


Figure 3. Work principle of the whole system

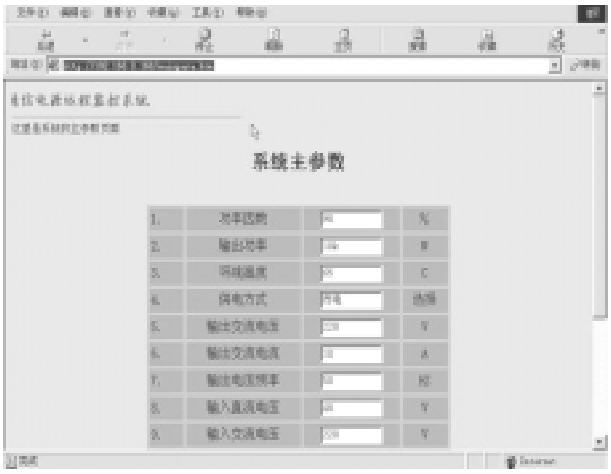


嵌入式逆变电源远程监控系统

请输入您的使用者姓名：

请输入您的使用者密码：

Figure 4.Login page of the system



系统主参数

1.	功率因数	[illegible]	%
2.	输出功率	[illegible]	W
3.	环境温度	[illegible]	C
4.	供电方式	市电	选择
5.	输出交流电压	220	V
6.	输出交流电流	[illegible]	A
7.	输出电压频率	[illegible]	Hz
8.	输入直流电压	[illegible]	V
9.	输入交流电压	220	V

Figure 5. Control monitor page of the system

7. CONCLUSION

It is the market requirement that the traditional control domain should be and has being innovated by the field bus and Internet technology. Embedded systems, as a representative will play an important role for its reducibility, mini-capacity, management of memory and process based on operation system, network support etc. Moreover, the technology on field bus and Internet are bonding. This will a new incremental point in the back PC times.

8. REFERENCES

[1] Li Jia *et al*. Development of Embedded Remote Control Monitor System. Embedded development web

[2] Huang Ying Xiao Xu. Development of Remote Control and Monitor System Based on Embedded Linux. Electronic Engineer, 2002.4

[3] He Wen. CAN Bus Summarize

[4] Wu Kuanming. Principle of CAN Bus and Application System Design .Beihang University Publish Press,1996

[5] HuaHeng Science and Technology. HHCF5272-R1Techno-mannual

[6] MCF5272 User Manual. e-www.motorola.com

[7] Rao Yuntao etc. Princinple of CAN Field Bus and Application technology. Beihang University Publish Press, 2003.6

[8] Manual about BOA,www.boa.org

[9] David A Rusting. Linux Kernel

[10] Linux Device Drivers. O'reilly Publish Press

[11] uClinux ports on MCF5272 manual. www.uClinux.org

QinJuanying was born in 1944. She is a Full Professor in college of automation, Wuhan University of Technology (WHUT). She graduated from University of Electronic Science and Technology of China in 1969. Her research interests are in Control Theoty and Engineering, Computer Control.

Feng Xin was born in 1977. He is studying in School of Automation at Wuhan University of Technology(WHUT). His research interests are computer control and net control.

Wu Guoping is a MS candidate of Control Theory and Engineering，college of automation, Wuhan University of Technology (WHUT).

Study on Distributed Control System of Fuel Cell Electrical Vehicle

Wu Youyu Yang Jufang Xie Changjun
School of Information Engineering，WuHan University of Technology
WuHan，HuBei Province，China
Email：wuyouyu1@sina.com **Tel：**13349995518

ABSTRACT

This paper introduces the tasks and constitution of distributed control system in fuel cell electrical vehicle (FCEV). The function of each ECUs is analyzed. Based on optical fiber CAN, the distributed communication network is designed in vehicle. Using the active star coupler, the optical fiber CAN with star topology is put forward. The active star coupler is designed based on CPLD, and it has CANH and CANL communication interface. The transmission protocol of the distributed control system is introduced briefly. The road experiment of the vehicle shows the design of distributed control system is reasonable and the effect of control is satisfied also.

Keywords: fuel cell electrical vehicle，distributed，control system，optical fiber CAN，active star coupler.

1. INTRODUCTION

Facing the double crisis of energy and environment at the present time, the automotive industry raises the need to improve the energy efficiency of the automobile, and to reduce the emission pollution. But it is difficult to solve the problem only by improving the performance of internal combustion engine. Developing EV (Electric Vehicle) is one of the effective ways to solve the problem. At present EV is divided into three kinds: PEV (Electric Vehicle), FCEV (fuel cell Electric Vehicle) and HEV (Hybrid Electric Vehicle). For the PEV the energy is a storage battery on vehicle, and the vehicle is driven by electromotor. The advantage of PEV lies in its no emission, but the distance covered is short after each charged. If the waste battery is disposed improperly, it will cause more serious pollution. For the HEV the car-mounted energy is provided by gasoline and battery. The advantage of HEV is economy in fuel, but zero emission can not be achieved. For the FCEV the car-mounted main energy is a fuel cell and the auxiliary energy is a storage battery. So far, Fuel cell is very likely to solve the energy problem for the vehicles.

FCEV is characterized by its high efficiency and no pollution. Its emission is water, so it brings no pollution to the environment and it is used more widely than PEV. It compares favorably with the internal-combustion engine, but its cost is too high. With the approaching energy shortage and the increasing concern for environmental protection, FCEV is drawing the attention of the world. Many automotive companies focus their personal and resource on FCEV and its relevant technology.

2. THE TASKS OF FCEV CONTROL SYSTEM

The electric vehicle which is driven by fuel cell and battery is called fuel cell electronic vehicle [1][2](FCEV. fuel cell electric vehicle). On FCEV, the fuel cell is the main power supply, and the battery is auxiliary power. The tasks of the control system for the vehicle are：① the two energy quotas are assigned rationally, and the dynamical, economy, security and comfortableness of the vehicle are improved. ② In order to ensure the operation of fuel cell stack is normal with high efficiency, the controlling for the fuel cell stack is implemented by the fuel cell ECU. The ECU controls the circulating water of the fuel cell, air blower, hydrogen supply and produces various alarm signals. ③ According to the vehicle operation status, gear location and position of the pedal, the electromotor is controlled by the driving current. ④ According to the quota of power flows, the output power of DC/DC is controlled. ⑤ For the auxiliary energy on vehicle (NiH battery) -----the voltage and temperature of each monomer batteries is stakeout, and the equalization charge management is implemented for each battery group(consisted of 10 monomer batteries in series). ⑥ The original instrument panel should be resumed and adjusted, the odometer should be resumed, some function of the panel should be regulated to instruct the special information of FCEV that is very important , such as the SOC of NiH battery, the entrance pressure of the stack , the water temperature out of stack and various alarm signals. ⑦During the steering many kinds of information should be watched and stored in real-time, it is convenient to realize vehicle operation status in time and guarantee the steering security; Store the operation information for the failure analysis and the debugging on vehicle.

3. THE COMPOSING OF DISTRIBUTE CONTROL NETWORK IN FCEV

As mention above, there are many tasks in FCEV, and it is very complicated to control the process. If the centralized control system is adopted in FCEV, it will bring the insufficient process ability of the central ECU. At the same time complicated and long physical wires of the system will lead to instability because of interference. As the interference is easily induced, it will cause the instability of the overall control system. And the extremely trouble will be bought in the mount, debugging and maintenance. To solve above problems, distributed control system basing on CAN Bus is introduced in the FCEV’s control system. All the tasks are divided into several modules. Each module will be managed and controlled by each CPU. Among the modules the data exchange and control command transmission is implemented by CAN bus.

The Controller Area Network (CAN) is an advanced serial communications protocol [3] which was proposed by German Bosch Company. Used for solving the numerous data exchange in the automobile. The CAN network transfer rate can reach up to 1 Mbit/s. The transmission medium of CAN bus can be twisted-pair, optic fiber and coaxial cable. Though the twisted-pair CAN network is realized easily in technically, and have certain ability of resisting electromagnetic interference, but it cannot work well under awful industrial field. The field situation of FCEV is comparatively complicated, because of the strong electromagnetic interference is produced by various electric equipments on vehicle, the CAN network which employed twisted pair can't work normally sometimes. The demand for higher anti-interferential ability and reliability is raised to CAN bus network transfer medium.

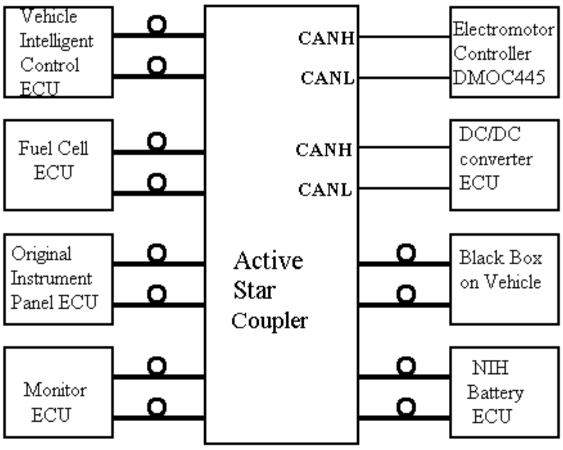


Fig.1 Fiber CAN Distribute Control Network Structure

In this paper, optic fiber CAN bus is engaged in the distributed control network for the FCEV. The network architecture is a star topological [4] based on the active star coupler, as Fig.1 shows. The optic fiber CAN network was composed of various nodes on vehicle and the center active star coupler. According to the function of tasks on vehicle, it is classified into 8 nodes. They are: vehicle intelligent control ECU, the fuel cell ECU, original instrument panel ECU, monitor ECU, the electromotor controller, the NiH BMS ECU, DC/DC converter ECU and black box on vehicle.

The fig.2 shows the blocks diagram of the distributed control system of FCEV. The diagram illustrates the function of each ECU. The whole distributed control system consists of the vehicle intelligent control ECU, the fuel cell engine, the NiH battery pack, the propulsion system and the wheels, etc.

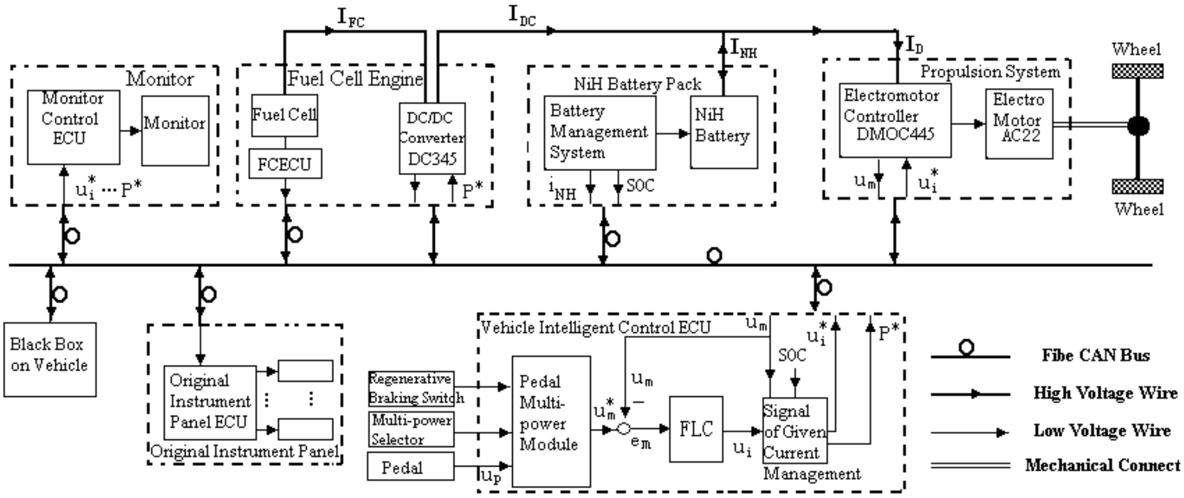


Fig.2 Blocks diagram of the control system of FCEV

The fuel cell engine supplies the main power for the FCEV [5]. It is made up of the noumenon of the fuel cell and DC/DC converter. The output voltage of the fuel cell noumenon is a widely range of 125-250V.Its output characteristic is unsteady, which means the output voltage is changed with its load; In order to gain a stability DC voltage of 288V for Hi-voltage bus of FCEV, a high power DC/DC converter (DC345) is needed to perform the conversion of the DC voltage. The DC output of DC/DC converter is controlled by the means of constant power controlling. According to the speed shifted by the driver, the vehicle intelligent controller controls the product of the fuel cell engines voltage and current.

The NiH battery pack is engaged as the auxiliary power of FCEV. It is provided the Hi-voltage bus standard, and serves as 'the reservoir' of FCEV Hi-voltage power, such as supplementing the energy shortage of FCEV while steering; or storing the excessive energy of FCEV while the vehicle loads lightly. The function rules and regulations is mentioned as follows: ① The electrical power is provided when the fuel cell is started; ② The electrical energy is supplied during the climbing and accelerating of FCEV ; ③ Reclaim the feedback energy [6] when regenerating braking in FCEV; ④ The electrical energy is provided for the electric equipment on vehicle such as the control system, the lighting system and etc.

The fuel cell engine and NiH battery pack are in parallel connection to produce DC hi-voltage power of 288V. The driving system is consisted of the electromotor (including the decelerator) and the electromotor controller DMOC445.The function of driving system is similar to the traditional engine on vehicle. DMOC445 is programmed as torque controlled. The driving torque of electromotor is controlled by the driving current provided by the fuel cell engine and NiH battery.

The vehicle intelligent control ECU is the heart of the distributed control system of FCEV. Its tasks are : ① According to the driver's manipulation of the multi-power selector(for shifting accelerative power), the regenerative braking switch and the pedal, as well as the driver's habit and the requirement of vehicle steering, the different mode and style are selected (such as reverse, cruise, direction machine, man-steering mode, lady-steering mode etc.) to control the vehicle steering; The control information of driving torque of motor is calculated and sent to DMOC by the vehicle intelligent control ECU. ② According to the load of vehicle, the status of the fuel cell engine and the NiH battery pack, the direction of power flowing and the distributive ratio of power flowing is controlled by the vehicle intelligent control ECU. It is implemented by the manipulation of the output power of fuel cell engine.

The operation principle of the vehicle intelligent control ECU is: after being processed by the pedal multi-power module, the pedal input u_p is converted into u_m* that is the given parameter of driving torque of the system. The difference e_m between u_m* and the practice torque u_m* is sent into FLC (fuzzy logic controller). FLC is the driving strategy controller. Originally, its output can control and drive the torque of motor directly. Because the fuel cells and the NiH battery pack supply power on vehicle, according to the value of u_i and the status of the two powers at present, the direction of power flowing and the distributive ratio of power flowing must be determined, and the distributive ratio control is implemented. The task is completed by the power flow module. The output of the power flow module u_i* is regarded as the signal of given current to control the driving current of motor (the driving current controls the torque of the driving system), until the torque difference e_m is equal to 0.

Stakeout and diagnosis of failure is the task of monitor system on vehicle in the real-time. Some important data are displayed selectively. These data are the output current and voltage of fuel cell noumenon, the hydric pressure and the flow of fuel cell, the temperature of water, the rotational speed of the air blower, the output current and voltage of DC/DC converter, the output current and current direction of the NiH battery pack, and the voltage and temperature for each NiH battery group (the auxiliary battery is consisted of 24 groups), and driving current of motor.

The original instrument panel ECU implements the recovery and improvement of the function for the original instrument panel. The data is shown as follows: the speed, the total mileage, the daily mileage, the SOC of the NiH battery pack, the entrance pressure of stack, the water's temperature out of the stack and various alarm signals.

The black box on vehicle records the key datum during the steering of vehicle. And it is convenient to debug the performance of the vehicle or find out the reason of failure. According to the requirement, the recorded data can be adjusted in time for the black box. The minimum time of continuous recording is more than 24 hours for the black box.

4. THE DESIGN OF THE OPTIC FIBER CAN NETWORK

According to the material, fiber optic can be divided into silica fiber and plastic of fiber (POF) by the material. At present, the silica fiber is widely used in modern communication networks. Silica fiber technology is mature. Silica fiber has many advantages such as lower attenuation, big bandwidth, and anti-electromagnetic interference and easy cabling. But silica fiber has small fiber core (less than 10μm). And it requires higher alignment precision, which means higher alignment cost. Silica fiber can be divided into singlemode fiber and multimode fiber. POF has been widely applied to high speed, short distance communication network. Its cost is similar to electric cables. Its diameter can reach 0.5-1 mm. The alignment is easy. The low cost plastic connector can be used. The connection cost is low. Plastic fiber also weighs light. Its loss can be reduced as low as 20dB/km.

In the design of fiber CAN network on vehicle, the communication distance is usually short. We can select Agilent's high performance ultra-low loss POF, and its typical loss is 0.19dB/m. POF's loss is minimum at 650nm. Agilent's photoelectronic transceiver HFBR-0508 series is chosen. The transmitter is HFBR-1528, and the receiver is HFBR-2528. Their operating wavelength is 650nm. The operating rate of HFBR-0508 series can reach 10Mbit/s. The transmission delay of each module is not more than 13ns. The transmission distance is 30m.

The active star coupler is the core of fiber CAN network. In order to design the active star coupler, the transmitting and receiving protocol of the CAN bus should be understand at first. For CAN bus, when a certain node acquires the control authority of the bus and becomes the transmitting node (the master node), this node completes two tasks: transmit data and monitor the bus data transmission and receiving. The specific process is: the host node's Tx0 port transmits data; its Rx0 port first receives date from CAN bus and compares it with the data it sends by itself. If RX0=TX0, the data transmission is correct. The transmitting node RX0 continues to monitor the bus data. If the acknowledgment slot of the data frame acknowledgment field is low level 0, it regards the data receiving to be correct. As to the receiving nodes in the bus, it should also fulfill two tasks: receive data and transmit ACK signals. The receiving node RX0 first receives data. When it receives the matched CRC, it regards the data receiving to be correct; and then it transmits acknowledging signals by TX0 (a low level), overwriting the data frame's ACK slot to show that the data has been correctly received. It should be noted that no matter the transmitting nodes or the receiving nodes, all nodes RX0 receives the same signal.

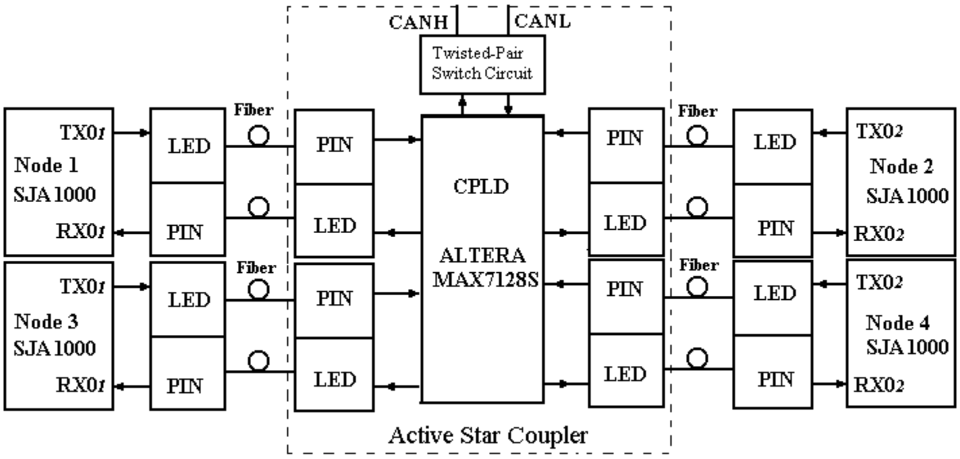


Fig.3 The structure of Active Star Coupler

Following the CAN transmitting protocol, the design of active star coupler is shown in Fig.3. It is consisted of CPLD and photo electronic transceiver modules; CPLD implements CAN bus transmission and receiving protocol. It must be ensured that transmitting node can receive the data sent by itself, and transmit an ACK signal to the transmitting node after receiving one frame message correctly. The chip of CPLD is MAX7128SLC84-15 made by Altera corp., and it is programmed by VHDL language.

In the network on vehicle, DMOC and DC345 are the important components, which are made by Solectria corp. of America. Unfortunately, and only twisted-pair communication port (CANH、CANL) is provided. In order to communicate with fiber CAN network on vehicle，CANH and CANL ports are extended by special twisted-pair convertor in active star coupler. By this means, DMOC445 and DC345 can communicate with the fiber CAN network without special gateway.

Table 1. ID assignment and update rate of each ECU node for FCEV

Node	Send	The Turnover Rate
Vehicle intelligent control ECU	0x013:The statue of motor, fuel cell statue establishing 0x007: the order of Shake hands with fuel cell and other ECU 0x022:DC/DC changer and DC345 output power establishing 0x709:Electrical machinery controller DMOC445 urges t he torsion to establish	50ms 50ms 50ms 50ms
Fuel cell ECU	0x012:the order of Shake hands with completed car 0x032: Send the voltage of the electricity and pile, the temperature of electric current pile input and output of the pile, Water pressure of the cooling water, flow of hydrogen Various kinds of warning information	50ms 1s
NiH battery ECU	0x024: The battery charters SOC state, total voltage and total electric current 0x014:Send group's voltage, temperature per battery 0x034:Send three groups of most abominable battery group's information	50ms 1s 1s
Monitor ECU	0x015:Shake hands with The completed car controls ECU	50ms
Original instrument panel ECU	0x027:Shake hands with The completed car controls ECU	50ms
DMOC445	0x611:Temperature of the radiator and electrical machinery 0x601:Actual rotational speed , real torsion , voltage of bus bar , state and trouble	50ms 50ms
DC345	0x020:the high and low terminal voiltage and current 0x021:Temperature of the radiator	100ms 100ms

5. PROTOCOL OF NETWORK

Because Solectria's motor controller DMOC445 and DC convertor DC345 adopt protocol of CAN2.0B with 11 bits ID and with the transmitting rate at 125Kbit/s. In order to get rid of special CAN gateway to transform protocol and simplify CAN network. The rate of transmission of fiber CAN network is 125Kbit/s. The CAN2.0B protocol with 11 bit ID is adopted. ID assignment and update rate of each ECU node shows in

Table 1.

6． CONCLUSIONS

In the design of EV completed car control system, we divide the task into 8 sub functional blocks. The management of every sub functional block and task of controlling are undertaken by one's own CPU. The speed of optic fiber CAN network location of control system of completed car is established as 125kbit/s. its frame cycle should be 8.56*10-4s. The communication protocol that this FCEV are made follows CAN2.0B edition; Select standard frames of 11 ID to convey, and every frame owns 107bits in common, on terms that meet controlling and data transmission demand, compress every node send data, control all nodal datum send frequency as much as possible. The optic fiber CAN bus occupation rate of the distributed system designed is 19%. The occupation rate of the network is 15%. Distributed completed car control system this after loading, according to function that design, realize fuel cell normal various kinds of functions needed to go of EV, by the data analysis that the car loads the black box, according the demand of design, prove that controller control strategy of completed car be able to realize energy flows assignment, Repay and apply the brake in energy, and Exercise intellectual control, etc. energy is it apply the brake, exercise intellectual control, etc. The distributed optic fiber CAN network designed not only can realize the normal conveyance of the data, still can stand on various kinds of electric interference. This distributed completed car control system operates normally, and meet the designing requirement.

7. REFERENCES

[1] Sun Liqing, Sun Fengchun, Chen Yong．The Design and Trial Manufacture of Fuel Cell Car Prototype [A]． The 19th International Battery、Hybrid Fuel Cell Electric Vehicles Symposium & Exhibition[C]．Korea: IJAT (International Journal of Automotive Technology), 2002, 72-79

[2] Scott Paul B, Mazaika David M． Hybrid-electric Fuel Cell Bus Test and Demonstration [A]． The 19th International Battery、Hybrid Fuel Cell Electric Vehicles Symposium & Exhibition[C]．Korea: IJAT (International Journal of Automotive Technology), 2002, 56-63

[3] 邬宽明．CAN 总线原理和应用系统设计[M]．北京：北京航空航天大学出版社，1996

[4] 吴友宇，梁红．星型光纤 CAN 网络研究[J] 计算机测量与控制， 2004.5

[5] 胡骅，宋慧．电动车[M]．北京：人民交通出版社，2003

[6] Yeo Hoon, Kim Talchol, Kim Chulsoo, Kim Hyunsoo． Performance Analysis of Regenerative Braking for Parallel Hybrid Electric Vehicle using HILS[A]． The 19th International Battery、Hybrid Fuel Cell Electric Vehicles Symposium & Exhibition[C]．Korea: IJAT (International Journal of Automotive Technology), 2002, 438-448

Wu Youyu is a full professor and a header of Intelligent Control Lab. She received the master degree in Solid-state Electronics Department, Huazhong University of Science and Technology, China, in 1990. In 1998, she attended in advanced studies, Institute National Polytechnique de Lorraine, France, doing researches on electronic control ignition for Engine. In 2000, she joined the research center of Intelligent Control, where she is currently a senior researcher. She has written a larger number of research papers in areas of FIBER-CAN. Her current research interests include FIBER-CAN, WLAN, and Distributed Control System of Fuel Cell Electrical Vehicle.

Yang Jufang graduated from the Wuhan University of Technology, China, in 2002. And then she keeps on sduding there for a master degree. She joined the research center of Intelligent Control.She has worked at the FIBER-CAN network project divided from Distributed Control System of Fuel Cell Electrical Vehicle. Her research interests include FIBER-CAN network, and wireless LANs.

Design of the Distributed Long Distance Water Supply Control System With Process Field Bus Technology

Meng Hua, Yan Cuiying, Jia Huiren, Wu Xueli
College of Electric Engineering and Information Science, Hebei University of Science and Technology
Shijiazhuang Hebei 050054, China
Email: menghua0311@eyou.com **Tel**.: 0311-3913503

ABSTRACT

To meet the need of long distance water supply control system of a city, make full use of the characteristics of opening, reliability and good anti-interference of the PROFIBUS technology, make up the distributed measurement and control network. It realizes the long-distance control of well swarm, intelligent dispatch of water supply and synthetically network of pipes. Adopting distributed control structure, each control station can separate itself from the network and work independently. Being connected by Field Network Bus, they can communicate and deliver information with each other, which can greatly improve the automation and quality of the long distance water supply, realize the intelligent monitor and science management. It has the advantages of convenient operation, good reliability, prompt and accurate supervision and control etc, and has important realistic meaning for saving energy, reducing consume and improving economic performance of water-supply field.

Keywords: Process Field Bus, Distributed Measurement and Control Network, Long-distance Monitor, Network Transmission

1. INTRODUCTION

Water is the source of human and plays an important role in our lives. The serious problem that we must face is the lacking of water resource day by day, but the traditional water supply system can't make good use of it. Its management is distributed, lack of science, and poor in supply quality. Therefore it is a urgent demand for us to design a new, scientific water supply system, which can realize intelligent supervision and optimum dispatch of water consume of a city. Now the most common used forms of the automatic control system by local and international water factory include SCADA system, DCS system, IPC+PLC system, and a system composed of bus industry control computer, and so on. The main problems of those are bad opening, lack of distribution, and impossible to supervise and control the spot equipment etc. This paper present the method of designing the distributed intelligent monitor of long distance water supply control system with PROFIBUS technology, which is popular now, and can get good result of control.

2. CONTROL REQUIREMENT OF THE WATER SUPPLY SYSTEM OF A CITY

In a water supply system of a city, the groundwater is 25km far from the center, and there are 6 deep wells, which covers an area of 1.5km. Being gathered by the output pipe, the water is first sent to the reservoir of water factory, then to the customers. It demands that the level of the reservoir must be maintained at a certain set point, so the final controlled variable of the system must be the level. The network communication can realize the data communication among the water resource, water factory and the dispatch's office of water supplying company. The real-time control for the pump in the resource area can change the quantity of water supply and satisfy the level requirement, accomplish the sufficiently closed-loop control and attain long distance intelligent supervision of the city water supply.

3. FILED BUS TECHNOLOGY

Filed Bus technology is a real-time control communication network, which connects the spot controller at the lowest profile of automation and the spot intelligent equipment with each other. The FCS has the characters of opening, compatibility, maneuverability, integration and easy maintenance etc, it has overcome the disadvantages in the traditional industry process control system such as high cost, poor transmission accuracy and weak anti-interference etc, and strengthens the integrated ability of spot class information. It is a digital communication system with complete digital, real-time，double-directions, multistage, and is used on the production spot. FCS put appropriative CPU to traditional control equipments, which makes them have the characteristics of intelligence and network. At present, the PROFIBUS technology of Germany is one of the most influential types. Field Bus has simplified network system construction, 3 layer agreement, physical layer，data layer and application layer， which are similar to the OSI models, or has 4 layer agreement further including network layer or transmission layer. The agreement construction of PROFIBUS is established according to the international standard in ISO7498 and takes the (Open System Interconnection) OSI as reference model. The relation between the Field Bus model and OSI model is illustrated in Fig I. Thus PROFIBUS provides an all-directions transparent network from spot sensors to production management layer，And with its strict definition, perfect function and standard general technique, it becomes the model of opening distributed control system.

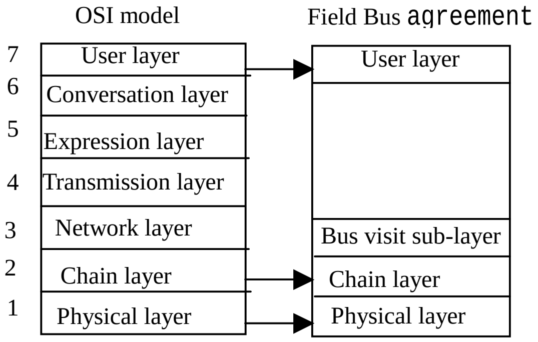


Fig.1 The relationship between Field Bus model and OSI model

4. CONTROL PROJECT AND SYSTEM CONSTITUTES

Systems based on Field Bus improves the autonomy and reliability, has the characteristics of opening, complete distribution, maneuverability and good adaptability in the spot etc. It forms all digital communication networks from measurement and control equipment to supervision and control computer, and adjusts the development of control network. According to the character and technology behavior demand of city water supply system, it takes the PROFIBUS technology to accomplish the communication and control among the plants, and takes this as the foundation to set up the whole control system. Considering the entirety, consistency, function economy and other parameters, we take the SIEMENS products to design the system. The whole supervision and control system is made up of spot monitored layer, data-handle layer, surveillance/supervise and control layer and center management layer and so on. Having the advantages of reliability, flexibility and good function, PLC is used for the spot monitored layer to collect the spot input and output signal and observes the operation of every spot equipment; the data handle layer decides control mode and data's returning to the file by PLC according to the user program and the spot equipments operation.

The surveillance/supervise and control layer directly communicates with PLC by the main computer which carries the software interface of I/O drive machine, and provides sketch which can be surveyed by spot operator, at the same time, the main computer provides an interface for the spot operator, so that he can complete the manual control to PLC. network, the center management layer can not only examine the node data of the whole system, analysis and print the data, make the statement, but also transmit them by wireless MODEM in specified rules to the water factory's center dispatcher office, where can give unitary dispatching to the system. The Topological Structure of Network is illustrated in Fig. 2

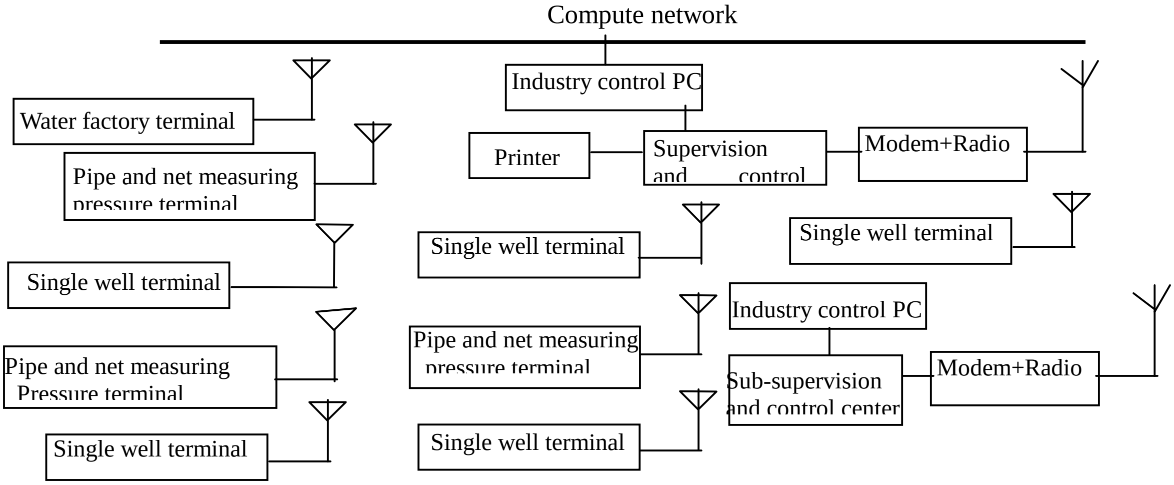


Fig. 2 The Topological Structure of network

Thus, the design sequence for the supervised and control system should include:

- Design of the pump station control system;
- Design of the central control center for pumps;
- Design of the level control system for reservoir of water factory;
- Design of Long-distance supervision and control dispatcher's office of water supplying company;
- Design of network transmission system.

This system is a typical distributed control system, the hardware install mainly includes industry control computer, SIEMENS wireless receiving, sending device, PLC, frequency converter and so on. The information delivering adopts two of the most advance communication networks: 1.The spot pump swarm in a close quarter adopts the wired network PROFIBUS of German SIEMENS; 2.Wireless transmission network is used for the long distance transmission among the spot central control room, dispatcher's office of water conservancy bureau and water factory. The structure of the system is illustrated in Fig.3.

3.1 Design of the pump station control system

This system collects signals related to water supply, it is primarily make up of frequency converter, PLC network interface and some common control equipments. S7-315 is the control center of the whole water resource, it can receive all kinds of signals from 6 pump station, water factory and water supplying company, and correspondingly sends out signal to S7-215 of related pump station, S7-215 control the frequency converter to adjust the speed of the motor, and accomplish the degrees of long-distance water supply. The data communication interface of S7-315, which can be connected with the control center of water factory and the computer network of dispatcher's office by wired and wireless, carries out real-time data transmission, transmits water-supply information and operation of the whole pump swarm to the honor calculator in control center to complete lumped supervise, S7-215 is the only one which has the PROFIBUS-DP interface in 200 series, it receives control signal from S7-315 to control frequency converter etc and collects signals of pressure, flow and super-speed and data transmission. Its interface can be connected to PROFIBUS to go forward super-speed data transmission. Thus, the PLC in

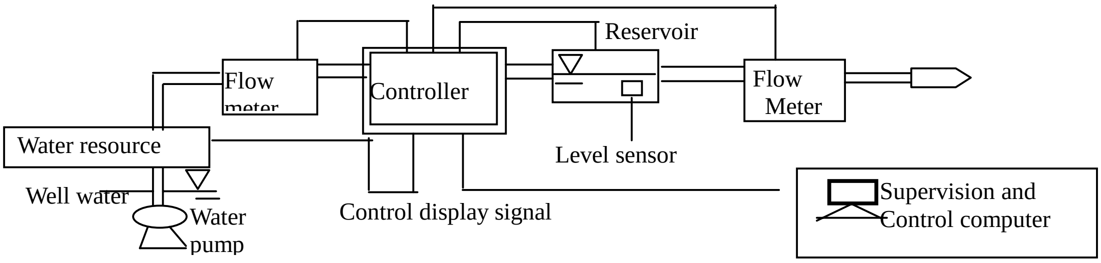


Fig.3 the Structure of System

pump room can complete measurement and control every signal and electric equipment, and enter center control room by network bus, effectively achieves the share of information and resources. The control structure of water pump control system is as follows.

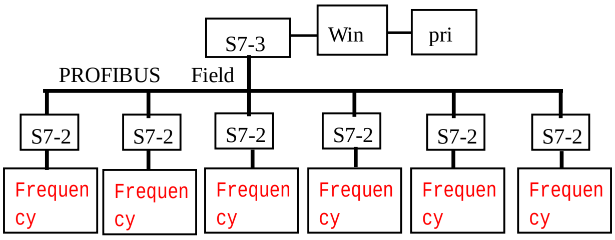


Fig.4 The Structure of Water Pump Center Control System

3.2 Level measurement and control system of water factory

When the level of water reservoir attains or exceeds the upper limit of set point, we can turn off several pumps manually by communication system, or send signal to s7-215 in automation form, which can control frequency conversion device to change the frequency. When the level attains or lower than the low limit of set point, we can turn on water pumps manually or control frequency conversion device by s7-215 in automation form. The opening or stopping operation of water pump can be completed automatically according to procedure by system or completed manually by operator, or by order sent out by supervisor and control room. The principle of the level closed-loop system is as Fig. 5.

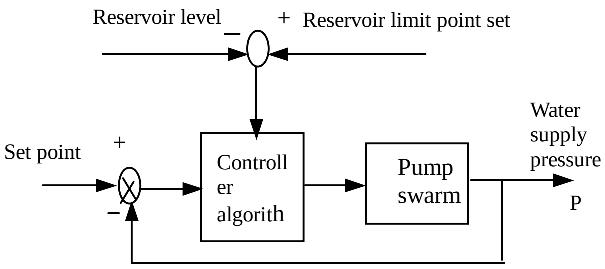


Fig.5 The principle of the level control system

3.3 Dispatcher's office of water-supply company

First, dispatcher's office is the water-supply dynamic information center, responsible to water-supply in superior quality, safely and low cost, It displays prints and saves the operation data of water factory, pipe and net and other related segment in time, then makes the optimum dispatch project, Second, it make reasonably allocations for water-supply, send out production order or directly operate the final control element to control the operation of the pumps and values. As the water-supply company is far away from the resource. We adopt the option of wireless transmission in signal transmission. Or we can make up a LAN by one or several computers when necessary. In this system design, the place of honor calculator uses the SIEMENS WinCC configuration software.

3.4 Design of network transmission system

In this system, to realize the coordination, optimization and dispatch of the pumps, a network system which can connect controllers, frequency conversion device of each pump station together is needed, so as to accomplish information exchange transmission, lumped display and control. In this project, it adopts the method of combination between wired transmission and wireless transmission. Considering that the system operates outdoors, the information transmission from spot pump to control center of center power distribution adopts reliable fast wired transmission network, further more, the PLC are all product of SIEMENS, they have all integrated PTOFIBUS interface, so we choose industry spot bus net Profibus released by SIEMENS, which is the most advanced in technology and successful in application. In addition, the level signal and flow signal of water reservoir should be transmitted to spot, and the spot signal should be transmitted to supervisor and control computer of water-supply company. Because wireless transmission is a valid and feasible method to complete long-distance data transmission in modern control fields, it can guarantee the reliability of data transmission. This part adopts wireless transmission form. The allotment between the primary and inferior stand is as follows:

①This system chooses PLC with PROFIBUS interface as primary station, it allots an address for each station, collects information such as the appearance condition station of vary-frequency and condition, value, pressure, flow, level and other parameters of the valves in each information cycle, sends them to PLC, at the same time, it sends optimum signal from PLC to each inferior spot, control the operation, run and stop and rotational speed of the water pump, match value's control get the purpose of optimum.
②The place of honor PC in this system is the second kind primary station, it chooses industry control computer with PROFIBUS. Based on the flow of water factory attained by from the wireless network, it get optimum control, sends it to PLC by PROFIBUS. Then PLC controls frequency conversion device's operation based on optimum result.
③Six S7-215 of water resource are regarded as inferior station, they receive orders of PLC to control frequency conversion device's operation, collect information from all pump stations.
④Communication between PC and PLC is realized by data exchanging.

5. CONCLUSION

The method of realizing the intelligent supervision of city water supply in this paper, makes full use of the opening, reliability and good anti-interference of Field Bus, Realizes the long-distance control of well swarm, intelligent dispatch of water-supply and synthetically network of pipe. Adopting distributed control structure; each control station can separate itself from the network and work independently. Being connected by field network bus, they can communicate and deliver information, improve the operating efficiency of water pump, has the advantages of convenient operation, good reliability, prompt and accurate supervision and control etc., it has important realistic meaning for saving energy, declining consume and improving economic performance of water-supply field.

6. REFERENCES

[1] Bai Yan, Wu Hong, Yang Guotian. distributed control system and field bus control system, electric power press of China, 2000.

[2] Lee Hongbin, Zhang Chenghui etc. Long-distance frequent conversion speed governor computer control system design, electric power drive, 2002.

[3] Wang shiping, Zhong junwei, Xia anbang, Toy building blocks model forming technology and application in city water-supply dispatch. Automatic learned journal. 1992, 18(b): 367-370

[4] Zhang shaohou, Lu Xihua, Talk about the form structure and function of automatic control system in water factory. Water-supply and water-drain, 1996, 5

The Research and Application of Real-time Monitor System Based on CAN Bus Network

Tao Dexin, Cao Xiaohua, Mo Lili
Logistics Department of Wuhan University of Technology
Wuhan,Hubei,430063,China
Email: DeXinTao@mail.whut.edu.cn **Tel:**13907133072

ABSTRACT

This paper analyses problems exist in traditional DCS industry monitor system and brings forward a real-time monitor system model which is based on CAN bus network. It also expatiates on the design method of the model's hardware and software. At last, it introduces the application result of the model in brief.

Keywords: CAN, Real-time monitor, Network Communication , Communication Protocol

1. INTRODUCTION

At present, the industry monitor system which is universally used in the world is Distributed Control System (it is called DCS for short).Its main characteristic is that the connection between devices in field and controllers is one-to-one(an I/O point is corresponding with a measuring point of the instrument) and signals transmitted are 4~20mA analog or DC24V switch message. This characteristic of the technology will cause problems as follow: ①message integration is not great. It can't satisfy whole industry monitor system's all requests to bottom data.②It is hard to ensure the reliability. Today control system's range is wide, I/O points are more than before, analog signals are easier to be interfered, all of which enhance the system's unreliability. ③The system's compatibility is bad. The types of products of DCS system are diverse and the products lack of manipulation and interchangeability each other. The system has no expansive space, so it is difficult to share messages in a wide scope.④ Real-time is poor. Because data transmission speed of traditional exchange technology is slow and it can't well deal with the problem in medium access, so it is hard to satisfy system's real-time.⑤The economic results are poor. Using traditional monitor system needs a large number of cables and its installation engineering is huge, so it increases the cost of production. New pattern of real-time monitor system based on CAN(Controller Area Network) bus network can overcome the defects in traditional system. It has the advantages such as higher communication speed, greater real-time, higher reliability, easier connection and higher ratio of function to price. So it will be widely used in field of real-time monitor and it will be a developmental direction of the monitor system.

The main monitor nodes are located in center room. They can not only monitor the running situation of every slave node and master the message about alarming fault, but also store all kinds of history records. Moreover, main monitor nodes can send different demands to every slave node at random, so they can realize the remote control to slave nodes.

The intelligent devices in field are installed near monitored machines. They mainly accomplish two tasks: one is to acquire and analyze data and control machines' running. The other is to receive commands from main nodes and accomplish the corresponding operation. The monitor is only used as supervision. It can't be used to send any data or command to any other node in the network, just can be used to receive the diverse real-time messages transmitted by broadcasting. So it can master the real-time produce situation in field. Whether monitor working or not will not affect the running of the whole monitor system. According to the actual request, we can increase or decrease the monitors.

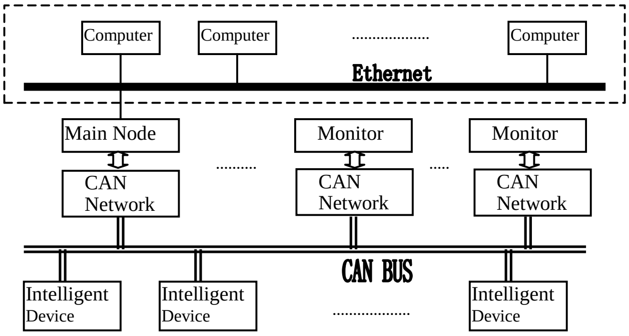


Figure.1 Whole Structure of CAN bus monitor system

In this kind of system model, all devices of the whole monitor system are connected only by a twisted pair of wires. They not only decrease the number of hiberarchy of traditional monitor network, but also can be more economical and credible than the traditional DCS system. At the same time, configuration of hardware devices in system is more flexible.

2. WHOLE DESIGN OF CAN BUS MONITOR SYSTEM

We use CAN bus network technology to establish a real-time distributed control system. It is illustrated in Figure.1. It consists of one or many main monitor nodes (they are also called main nodes), a number of monitors and intelligent devices (they are also called slave nodes). All these nodes are connected to CAN bus by a twisted pair of wires. (Area in the broken line is the enterprise information network and it is not studied here).

3. HARDWARE DESIGN OF CAN BUS MONITOR SYSTEM

Hardware configuration of main nodes and monitors

For the importance of main nodes, we may choose the computer (IPC above 486) that is designed according to the industry standard and is installed with CAN bus network card. From its components' selection to heat emission, quakeproof, dustproof, electromagnetic anti-interference is designed according to request of industry environment，so it has high reliability. The monitor can be made up of a commercial PC with a CAN network card. CAN network card consists of address coding logic, interrupt logic, CAN protocol controller, CAN drivers, photoelectric insulation circuit and so on. It mainly accomplishes the function of monitoring physical layer and data link layer of system network and realizes data exchange in bottom layer.

Hardware design of intelligent devices in field

The principle structure chart of CAN bus intelligent devices in field is illustrated in Figure.2.

In Figure.2, intelligent devices' host computer section consists of single chip microcomputer 8031, 16KB EPROM 27128, 16KB RAM 62128, latch 74LS373 and other assistant circuits. It realizes data analysis, data handling and other control functions. CAN bus units and interface sections consist of CAN protocol controller SJA1000 (supports the protocol of CAN2.0A and CAN2.0B), photoelectric insulation circuit and CAN driver 82C250. It accomplishes data exchanges with other nodes. Analog input circuit consists of I/V converter, multiple switch, amplifying circuit and A/D converter and so on. The multiple switch and A/D translation circuit can be selected according to concrete requests such as the number of analog input channels and the precision of A/D converter. This paper chooses 1/8 multiple switch MAX354, high speed A/D converter AD1671 which has 12 bits. It can realize acquirement of analog value in field. Analog output circuit consists of D/A converter, voltage converter and V/I converter AD694JN and so on. Its analog output channels' selection accords to system requests. D/A converter can select MAX530 which has 12 parallel input channels and low energy loss. Switch value's input and output channels are 4-input and 4-output switch circuits that consist of photoelectric insulation and 8 bits bus driver SN74LS245. Setting the logic of nodes address and Baud rate is to achieve initialization of CAN communication parameters.

Intelligent device in field is one node of CAN bus network. It not only has analog input and output channels, but also has switch input and output channels. It is convenient to acquire data and control process in the course of industry production. Because installed in field, it can decrease interference from transmission channels and enhance the reliability of system. At the same time, it can utilize communication advantages of CAN bus network to make communication with other nodes more flexible, communication speed faster, correcting function more strong. So it can radically ensure whole monitor system's reliability and real-time.

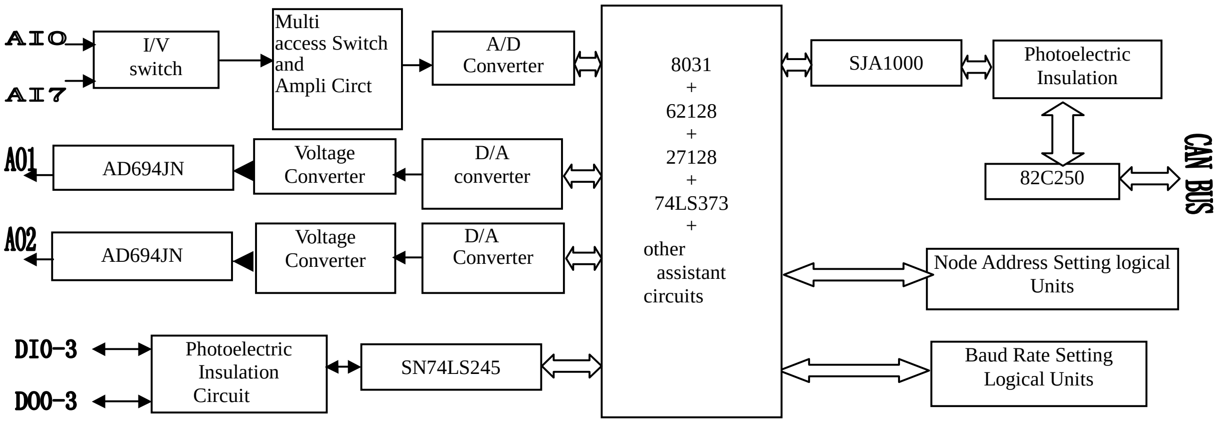

AI0
AI7
I/V switch
Multi access Switch and Ampli Circt
A/D Converter
8031 + 62128 + 27128 + 74LS373 + other assistant circuits
SJA1000
Photoelectric Insulation
82C250
CAN BUS
A01
AD694JN
Voltage Converter
D/A converter
A02
AD694JN
Voltage Converter
D/A Converter
Node Address Setting logical Units
DI0-3
DO0-3
Photoelectric Insulation Circuit
SN74LS245
Baud Rate Setting Logical Units

Figure.2 Principle-structure chart of CAN bus intelligent devices in field

4. SOFTWARE DESIGN OF CAN BUS MONITOR SYSTEM

Design user layer's exchange protocol

According to technology convention of CAN bus and standard of ISO11898, what we need to do now is to design user layer's protocol because hardware design we have mentioned above has achieved the function of physical layer and data link layer. Frame definition of user layer's protocol is illustrated in Figure.3. Node-to-node communication in system is based on this protocol to exchange data.

DIR: Direction bit. It is 0 when message is sent from main nodes to slave nodes and is 1 when it is reversed.

Address: Address domain. When message is sent from main nodes, this domain is destination address. When message is sent from slave nodes, this domain is source address. This domain has no meaning when broadcasting.

TYPE: Type of frame. Communication mode is called single frame broadcast when it is 100 and is called single frame point -to-point when it is 000. When message is sent from main nodes, the type of frame is the single frame point -to-point.

7	6	5	4	3	2	1	0
DIR	Address						
TYPE			0	DLC(0-8)			
Cmd 0							
Cmd 1							
Serial number of physical channels							
Data (0~4) (5 bytes)							

Figure.3 Frame definition of user layer

When message is sent from slave nodes, the type of frame is the single frame broadcast, in which way all nodes in bus network can receive message.

DLC: length of message. It defines the length of data (0 ~ 8 bytes).

Cmd0, Cmd1: Command function symbol. It has varieties of definitions according to the concrete situation. This paper gives some definitions in common use .It is illustrated in Table.1.

Table.1

Cmd0	Cmd1	Function Definition
“O”	“K”	Success of Initi
“S”	“T”	Node Initialization
“A”	“R”	Analog Request
“A”	“I”	Analog Input
“A”	“O”	Analog Output
“D”	“I”	Digital Input
“D”	“R”	Digital Request
“D”	“O”	Digital Output

Serial number of physical channels: Serial number of input and output channels set by user (0~255).

Data(0-4): Data which length is 5 bytes.

The length of protocol frame in user layer we have designed above is short. It not only satisfies message exchange requests of control system, but also conduces to reduce message conflict ratio when communicating and it can increase system’s communication speed and enhance real-time performance.

Design of sending and receiving program of main nodes

According to interface function offered by CAN network card’s drivers, user can use the languages such as C, C+, VC++ to design sending and receiving program of main nodes. The program adopts active way to send commands and adopts interrupt way to receive message. The design steps are as follow: Firstly, open CAN network card and make it initialization. Then use sending function in driver program of CAN network card to sent the system’s control commands or data request commands. The concrete method is to send a data-request command every 500ms and varieties of control commands at random. Utilize receiving function to receive different real-time message sent from devices in field and do corresponding treatments, display and store. In the end of the program, use close function to close CAN network card. It is not to give unnecessary details about realization of concrete codes here.

Software design of intelligent devices in field

Intelligent devices in field (slave nodes) mainly achieve two tasks: one is to acquire and handle data, and control the machine; the other is to receive the commands from main nodes and do corresponding operations. The first task can be defined in main program while the second task is defined in the interrupt service program. The program can use the assemble language C51 or ASM51 to design. The flowchart of program is illustrated in Figure.4 and Figure.5.

5. APPLICATION AND CONCLUSION

This system applies in monitor system of metallurgical industry’s water treatment equipments. This system used traditional DCS system before. It has many disadvantages. For example, its installation is complex, its cost is high and its data transmission speed is slow, real-time update is slow. It frequently appears the trace-lagging phenomenon and affect reliability of system’s running. When it becomes worse, equipments in field will be burnt down. After utilizing CAN bus real-time monitor system that brings forward in this paper, it not only overcomes disadvantages in former system, but also enhances system’s anti-interference and integration of system message. It also has flexible function in configuration of hardware and software.

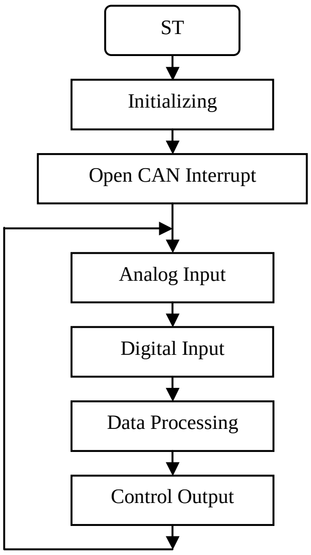


Figure.4 Main flowchart of slave node

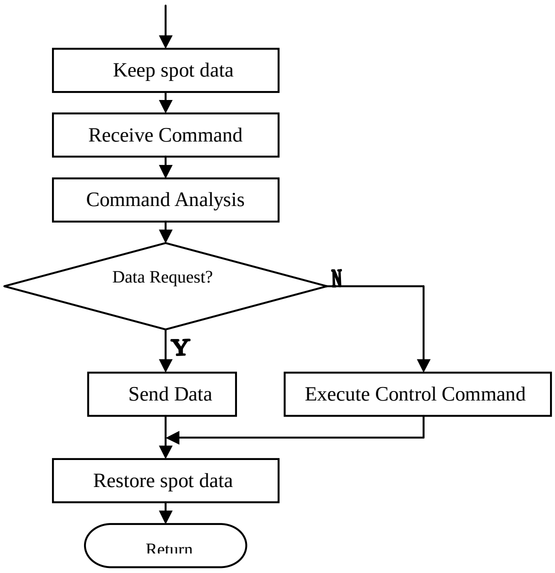


Figure.5 Interrupt service flowchart of slave node

It is proved from practice that real-time monitor system model based on CAN bus network brings forward in this paper is feasible. We use CAN bus network technology which has developmental outlook to research real-time monitor system. It also accords with developmental trend of monitor technology, so it will certainly replace traditional monitor system such as DCS and become mainstream technology.

6. REFERENCES

[1]. Wu Kuanming. Theory of CAN bus and design of applied system. Beijing. Beijing:university of

[2]. aeronautics and astronautics press, 1996.56—90.

[3]. Yang Xianhui, Wei Qingfu. Bus technology and application used in spot. Beijing:Qinghua university press,1999.24—56.

[4]. Shui Aishe,Liu Weihua, Zhou Ahaoqi etc. Intelligent monitor system of canned oil based on CAN bus. Beijing:Automation and instrument, 2000,6:24~25.

[5]. The CAN Protocol,

[6]. http:/ /www. kvaser.se/ can/protocol/index.htm.

r

Tao Dexin 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.

Semi-active Logic Control Algorithm for MR Dampers Using Accelerations Feedback

Chen Jing[1], Qu Weilian[2], Xu Youlin[3]
[1]College of Automation, Wuhan University of Technology, Wuhan 430070, P.R. China
Email: jingchen680@163.com Tel.: +86 (0) 27-62356851
[2]College of Engineering and architecture, Wuhan University of Technology, Wuhan 430070, P.R. China
Email: qwlian@public.wh.hb.cn **Tel.**: + 86 (0) 27-62680906
[3]Department of Civil and Structural Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Email: xceylxu@polyu.edu.hk **Tel.:** + 0852-27666050

ABSTRACT

A tall building with a large podium structure under earthquake excitation may suffer from a whipping effect due to the sudden change of building lateral stiffness and mass at the top of the podium structure. A comprehensive experimental investigation was thus carried out to explore the possibility of using magnetorheological (MR) dampers to connect the podium structure to the tall building to prevent the whipping effect. The tall building was constructed as a slender 12-storey building model whereas the podium structure was built as a relatively stiff 3-storey building model. Three MR dampers together with three current controllers manufactured by the Lord Corporation, USA, were used to link the 3-storey building to the 12-storey building. The two building models connected by the MR damper manipulated by a semi-active logic control algorithm using accelerations were tested under the specified ground motion. Experimental investigations were performed in the Structural Dynamics Laboratory in the Hong Kong Polytechnic University. The experimental results show that the MR damper with the semi-active logic control algorithm could significantly mitigate the seismic whipping effect and reduce the seismic responses of both the tall building and podium structure.

Keywords: semi-active control, multilevel logic control, MR damper, seismic response.

1. INTRODUCTION

Magnetorheological (MR) damper is deemed to be a smart fluid damper because the device force can be controlled through the on-line adjustment of magnetic field exerting on MR fluids to response the change in either structural motion or external excitation. MR dampers have the capability to provide large controllable forces and possess several attractive features such as simplicity, reliability, stability, and small power requirement. Additionally, the fluid itself responds in milliseconds, which allows for the development of devices with a high bandwidth. Their insensitivity to temperature fluctuations and contaminations makes them suitable for both indoor and outdoor applications (Carlson et al 1996; Spencer and Sain 1997). However, MR dampers are highly nonlinear control devices, and their control performance and real application are greatly dependent on control algorithm used. The proper selection of a control algorithm should consider many factors such as available feedback measurements, the number of control devices, and the control system reliability, in addition to intrinsic nonlinear behavior of MR dampers. It is also noted that it is not possible to directly command a MR damper to generate a specified control force because the response of the MR damper is dependent on the local motion of the structure where the MR damper is attached. In the past several years, intensive researches, both theoretically and experimentally, on the semi-active control algorithm for MR dampers for reduction of seismic response of civil structures have been undertaken. Jansen and Dyke (2000) carried out a numerical comparison of five semi-active control algorithm for MR dampers, which include the Lyapunov controller (Leitmann, 1994), the decentralized bang-bang controller (McClamroch and Gavin, 1995), the maximum energy dissipation controller (McClamroch and Gavin, 1995), the clipped-optimal controller (Dyke et al, 1996), and the modulated homogeneous friction controller (Inaudi 1997). Yi et al (2001) performed an experimental comparison of the Lyapunov controller and the clipped-optimal[2] controller for multiple MR dampers for seismic control of a six-story building model.

In consideration of previous studies, a novel semi-active logic control algorithm using accelerations[3] is proposed for MR dampers in this paper. The logic control algorithm are basically developed from the Boolean algebra and represented by multilevel logic control rules that relate the state variables of a multivariable input system to the state variables of a multivariable output system for the MR dampers. The logic control algorithm can be easily implemented in practice while making the best use of nonlinear properties of MR dampers. One practical logic controller is designed for a tall building with a podium structure linked by MR dampers, without requiring an accurate mathematical model of the control system and control device. The feedback information required in the proposed logic control algorithm is only the accelerations of dampers. To verify the proposed logic control algorithm, a series of experiments using a seismic simulator are performed on seismic response control of a slender 12-story building model and a relatively stiff 3-story building model linked by three MR dampers. Experimental investigations were performed in the Structural Dynamics Laboratory in the Hong Kong Polytechnic University.

2. SEMI-ACTIVE LOGIC CONTROL ALGORITHM USING ACCELERATIONS

Logic control has become a branch of knowledge that is composed of the mathematical logic and practical control technology, by which the accomplishment of logic thinking of human being into logic control model for a system is more straightforward. Semi-active logic control algorithm using accelerations developed in this study is based on the multilevel control rules, the theory of Boolean algebra, and the rules for multiple state variables. The multi-level control rules are established based on the practical control technology and the problem concerned, and they are used to relate the state variables of the multi-variable input system to the state variables of the multi-variable output system.

For a large civil structure subject to earthquake excitation, the application of semi-active control technology aims at reducing structural vibration significantly so as to enable the structure to go back to its static equilibrium position as far as possible. If semi-active control devices dissipate most of the structural vibration energy and hence mitigate the structural seismic response in turn, the structure can be maintained within a small region around its static equilibrium position. Thus, the basic principle for setting up the multilevel control rules is that the large damper force should be applied if the larger magnitude of deviation of the structure from the static equilibrium position occurs or tends to occur. If the structure vibrates within a small region around its static equilibration position, the damper force can be maintained at the present level. The core concept of semi-active multilevel logic control algorithm is then to switch the control device to the pre-specified force level (a multivariable output system) according to the feedback responses (a multivariable input system) and the multilevel logic control rules.

Development of logic controller based on acceleration feedback is an ideal solution to the problem. The acceleration $\&\&x_r$ of the *r*th MR damper is chosen to be control feedback signal[4][5]. The equilibrium range of the structure is defined by an index a_{r0}, which is a real number with value slightly greater than zero. In consideration of the large difference of dynamic properties between the two buildings used in the experiment, the two buildings are deemed to be in static equilibrium when the condition $\&\&x_r \le a_{r0}$ is satisfied. By introducing a set of acceleration index a_{rz} , for(z=0 ,1 ,Λ q | $a_{rz}<$ a_{rz+1})for a rigorous structural vibration control, the control force adjustment can be determined. In this experimental study, q is selected as 3. Therefore there are 5 (2q-1 = 5) zones for the acceleration $\&\&x_r$. These zones can be interpreted as follows:

(1) Zone O ($|\&\&x_r| \le a_{r0}$): it is specified as quasi-static equilibrium region and no additional current (damper force) should be supplied (NC*r*);

(2) Zone I ($a_{r0} < |\&\&x_r| \le a_{r1}$): the structure slightly departs from the equilibrium region, and the small control force should be applied by supplying low current (LC*r*);

(3) Zone II ($a_{r1} < |\&\&x_r| \le a_{r2}$): mild vibration is expected. The control force should be augmented by moderately increasing current supply (MC*r*);

(4) Zone III ($a_{r2} < |\&\&x_r| \le a_{r3}$): the structure is undergoing a large vibration, and therefore a medium high current (MHC*r*) is supplied to produce a large control force; (5) Zone IV ($|\&\&x_r| > a_{r3}$): the highest current supply (HC*r*) is required to generate the largest control force in order to enable the acceleration $\&\&x_r$ back to a lower level. In consequence, there are totally five states of control force (or current), which correlate with high current supply (HC*r*), medium high current supply (MHC*r*), medium current supply (MC*r*), low current supply (LC*r*) and no current supply (NC*r*).

The aforementioned control rules can be represented in terms of the input system variables, output system variables, and their state variables using Boolean algebra and the rules for multiple state variables in order to facilitate the realization of logic controller. The acceleration $\&\&x_r$ of *r*th MR damper is denoted by the input system variable X_r , and its five state variables X_r^1, X_r^2 X_r^3 X_r^4 and X_r^5 stand for $|\&\&x_r| \le a_{r0}$, $a_{r0} < |\&\&x_r| \le a_{r1}$, $a_{r1} < |\&\&x_r| \le a_{r2}$, $a_{r2} < |\&\&x_r| \le a_{r3}$ and $|\&\&x_r| > a_{r3}$. For the control output, Y_r is adopted as the output system variable. The state variables of the output system variable Yr are Yr^1 Yr^2 Yr^3 Yr^4 and Yr^5, referring to the five subdivided current supply HC*r*, MHC*r*, MC*r*, LC*r* and NC*r*, respectively. The logic control model of the semi-active logic controller is given by

$$Yr^1 = X_r^5;\ Yr^2 = X_r^4;\ Yr^3 = X_r^3; Yr^4 = X_r^2;\ Yr^5 = X_r^1 \quad (1)$$

When one of the output state variables Y^t_r for $t = 1, Λ ,5$) is 1 (true), the rest of its state variables will be 0 (false). The current supply corresponding to the true variable will be activated and the proper current will be set to the MR damper through the current controller.

3. EXPERIMAENTAL VERIFICATION

To apply and verify the concept of semi-active logic control using accelerations, one practical logic controller was designed for a tall building with a podium structure linked by MR dampers, and a series of experiments using a seismic simulator were performed on seismic response control of a slender 12-story building model and a relatively stiff 3-story building model linked by three MR dampers. The components of the experiment setup[1] include building models, MR dampers and current controllers, seismic simulator, instrumentation, and real-time control system.

Real-time Control System

The signals from the signal conditioners were passed to the dSPACE real-time simulator system. The real-time simulator system comprised a DS1005 PPC processor board (PowerPC controller), a DS2003 16-bit 32 channels A/D board, a DS2102 16-bit 6 channels D/A board, and a DS4003 digital I/O board. The DS1005 PPC processor board, which embraced a Motorola PowerPC 750 microprocessor running at 480 MHz as its central processing unit (CPU), functions as an interface to the I/O boards and the host PC. The control of the dSPACE CPU and the access to its memory were executed by the main program ControlDesk of dSPACE, which offered an automatic implementation of the MATLAB/Simulink block program on the host computer via Real Time Interface (RTI) and provided a real time interactive data display and visualization.

For the semi-active control of the two buildings linked by MR dampers, the motions of the dampers were determined based on the digital signals related to the structural responses of the two buildings at the common floors. The damper motions were then analyzed based on the semi-active logic control algorithm, from which control signals (command signal) were obtained and converted into analogical signals by the DS2102 D/A board in the dSPACE to apply the control voltage signal to the MR damper via the current controller so as to form a closed loop semi-active control system. The semi-active logic control algorithm was implemented by the PowerPC controller via the MATLAB/Simulink program. A block diagram for real-time control system is depicted in Fig.1.

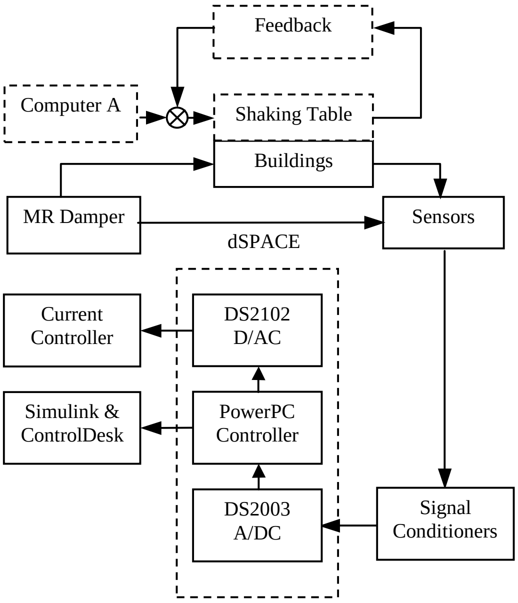


Fig. 1 Block Diagram of Real-Time Control System

Input Ground Excitations

In the verification of control performance of the semi-active logic controllers, the EL Centro 1940 earthquake ground acceleration (N-S component) was used as input excitation. In consideration of the typical parameters of prototypes, the simulation requirements, and the moderate seismic zone, the original El Centro N-S ground acceleration was scaled down with a time scale of 1:3 and a peak acceleration of 0.13g. The duration and sampling frequency of each ground acceleration time history, which was measured directly from the table surface of the simulator, were 30 seconds and 1000 Hz, respectively. The time histories of four types of ground acceleration excitation measured directly from the table surface of the simulator are shown in Fig.2.

Experimental Results

The proposed semi-active logic controller was first programmed by the MATLAB/Simulink and then converted to the executive program in the dSPACE via RTI, which consists of 3 major parts:

(1) Converting the measured voltages signals of the responses of MR damper to digital signals as feedback;

(2) symbolizing the feedback signals into the state variables of the input system, that are readable by the PowerPC controller to obtain the state variables of the output system according to equation (1);

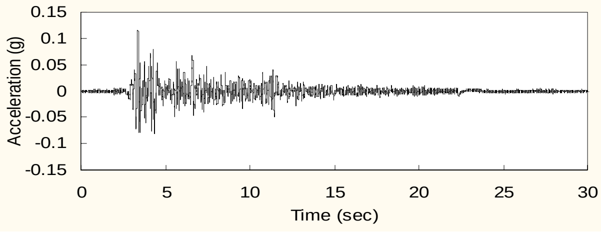


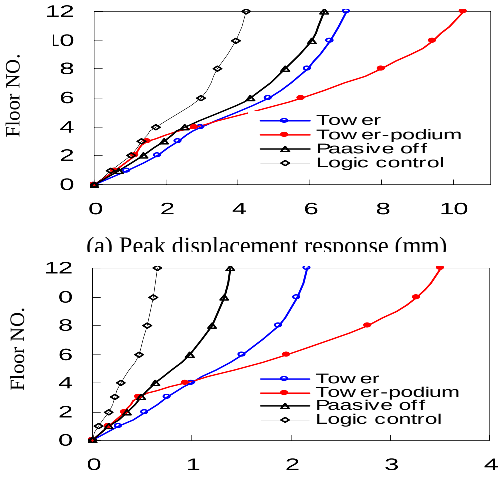
Fig. 2 Time Histories of Four Types of Ground

(3) sending a command voltage signal to the current controller to adjust the control force based on the result obtained from (2). The index of equilibrium state and the states of applied control current used in the semi-active controllers are shown in Table 1.

Table 1 Static Equilibrium States and Current Supply States

Equilibrium State(cm/s^2)				State of Applied Control Current (A)				
a_{10}	a_{11}	a_{12}	a_{13}	HC$_1$	MHC$_1$	MC$_1$	LC$_1$	NC$_1$
3	6	10	20	0.57	0.4	0.3	0.2	0.0
a_{20}	a_{21}	a_{22}	a_{23}	HC$_2$	MHC$_2$	MC$_2$	LC$_2$	NC$_2$
4	8	13	25	0.49	0.36	0.3	0.2	0.0
a_{30}	a_{31}	a_{32}	a_{33}	HC$_3$	MHC$_3$	MC$_3$	LC$_3$	NC$_3$
5	10	15	30	0.59	0.4	0.3	0.2	0.0

For comparisons, the passive off mode in which the applied current was zero to the MR damper (Passive off), separate tower mode (Tower), the case in which the two buildings were rigidly connected were also investigated (Tower-podium), and the MR damper in the passive off mode provided the information on minimum effectiveness of the damper if the external power was cut off.



(a) Peak displacement response (mm)

(b) RMS displacement response (mm)

Fig.3 Comparison of Displacement Responses of 12-Story Building

Figures 3a and 3b show the relative peak and RMS displacement responses of the 12-story building to the ground under the excitation for four cases: Tower, Tower-podium, Passive off, and in the semi-active controller mode (Logic control). It is seen that the responses of two buildings rigidly connected clearly exhibit the whipping effect as discussed by Qu and Xu (2001). The MR damper in the passive off mode totally eliminates the whipping effect that exists in the rigidly connected buildings. When the MR damper worked in the semi-active controllers mode, both the peak and RMS displacement responses of the 12- syorey building are further reduced compared with those with the MR damper working in the passive off mode. The reductions in the peak acceleration, RMS acceleration, peak displacement, and RMS displacement responses of the 12-story building with the controller are about 65%, 78%, 67%, and 84% with regard to the rigid connection case.

Time histories of displacement responses at the top floor of the 12-story building under the excitation are depicted in Fig.4 for the logic controller compared with the rigid connection case.

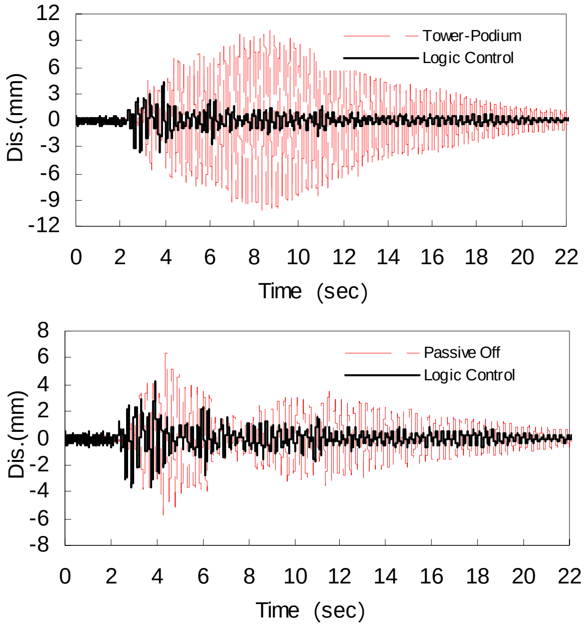


Fig. 4 Time histories of displacement responses at the top floor of the 12-story building under the excitation

It is clear that the MR damper with the semi-active controller can reduce the building response within the entire duration of the ground motion.

It is observed that the logic controller generally perform was better than the passive off control for both buildings and four types of ground excitations.

4. CONCLUSIONS

In the present paper, semi-active logic control algorithm using accelerations for seismic response control of complex civil structures using MR dampers have been proposed and verified through experiments conducted on a seismic simulator. The major advantages of the proposed semi-active logic control algorithm include the quick and simple control decision without requiring the accurate mathematical models for the control system and the MR damper, the full utilization of the nonlinear features of the MR damper, and the high reliability owing to the use of local information of the MR damper only. For the test 12-story building model with the 3-story building linked by MR dampers, one practical semi-active logic controller was designed of which the semi-active logic controller made use of acceleration response of MR damper. The experimental results showed that the effectiveness and efficiency of the proposed semi-active logic control algorithm was high.

5. ACKNOWLEDGEMENTS

The writers are grateful for financial support from the National Natural Science Foundation of China under Grant NNSF-50038010 and The Hong Kong Polytechnic University through its Area of Strategic Development Programme in Structural Vibration and Intelligent Buildings. The writers also very much appreciate the help from staffs of Structural Dynamics Laboratory of The Hong Kong Polytechnic University in conducting the comprehensive experiments.

6. REFERENCES

[1] Chen Jing, Qu W.L., Xu Y.L. Zhang N.L. Semi-active Logic Control of Tall Buildings with Ppdium Structure Based on Panboolean Algebras. Fourth International Conference On Industrial Automation, June 9-11, 2003.

[2] Carlson, J.D., Catanzarite, D.M., and St. Clair, K.A (1996), Commercial magnetorheological fluid devices, Int. J. Modern Phys. B, Vol. 10, No. 23/24, pp. 2857-2865.

[3] Dyke, S.J., Spencer B.F. Jr., Sain, M.K. and Carlson, J.D. (1996), Modeling and control of magnetor-heological dampers for seismic response reduction, Smart Materials and Structures, Vol. 5, pp. 565 – 575.

[4] Qu W.L. and Xu Y.L. (2001), Semi-active control of seismic response of tall buildings with podium structure using ER/MR dampers, the Structural Design of Tall Building, Vol. 10, No. 3, pp.179-192.

[5] Sasao, T. (1999), Switching Theory for Logic Synthesis, Kluwer academic Publishers, Norwell, Massachusetts, USA.

Chen Jing graduated from Wuhan University of Technology and received her MengSc and PhD degrees in 1997 and 2003, respectively. She is currently an Associate Professor in College of Automation at Wuhan University of Technology. Her research interest and practical experience are in the areas of control theory, computer control, and structural vibration control. She has published more than 20 technical papers in these areas.

Estimate Model of Delay in Autolever System and its Algorithm Design

Meigui Han, Jinglu Liu, Guangsheng Dongye
Information Science and Engineering College, Jinnan University
Jinan, Shandong Province 250022, China
Email: rose_xiaofeng@sina.com **Tel.:** 0531-2767500

ABSTRACT

In this paper we constructed a mathematical model by analyzing correlation of the input and the output yarns uneven signals. Using the model, we can realize on-line identification and timing adjustment the draft delay in autolever system. In order to improve the time efficiency of algorithm, we adopt the FFT method to realize this model, and we also discussed some detailed problems in algorithm realization.

Keywords: Auto-level, Delay, Signal Processing, Convolution, FFT, Algorithm

1. INTRODUCTION

In spinning technology, reducing the yarn's unevenness is required to improve the end product's quality and productivity. So it is necessary to equip with autolever at the proper technical place to improve the output yarn's unevenness. In traditional autolever system, the draft delay control is generally realized with machinery, such as memory steel roller and cylinder. It can't control the change of unevenness caused by some factors, such as random factors, the character of draft machinery and the yarn's self characters, after yarns passed through the measuring device. While the closed-loop autolever system can adjust drafting according to the feedback information of the output yarn's uneven, its uneven wavelength limits with long segment, it can't even the short segment unevenness. At present, there aren't some perfect autolever and control system. If we find some relations or rules between the draft and the yarn's unevenness on aspect of signal processing, we could construct some appropriate model to intelligently identify the delay's situation in drafting process and we can accordingly adjust the draft. So we can realize the intelligent control of delay and improve the capability and auto-control level of the autolever system.

2. CONSTRUCTING THE MODEL

This model is based on the mixed-loop control system, and we studied the delay problem in its open-loop part.

The effect on the output yarn's unevenness caused by the change of draft.

According to auto-level principle,

$$V_1 \cdot G_1 = V_2 \cdot G_2 \qquad \text{Eq(1)}$$

Thereinto V_1 is the speed of the back roller and V_2 is the front roller, G_1 is the weight of input yarn and G_2 is output yarn. Assume $G_1 = \Delta G + G_0$ (G_0 is the rated input weight, ΔG is the input deviation), then we can get

$$G_2 = (\Delta G + G_0)/E \qquad \text{Eq(2)}$$

Here **E** is the draft multiple, it is the ratio of the V_2 to V_1. If we assume the output is standard, that is G_2 keeps constant, we can get the basic auto-level formula:

$$E = E_0(\Delta G + G_0)/G_0 \qquad \text{Eq(3)}$$

E_0 is the rated draft, the ratio of the G_0 to G_2.

We assume the input yarns and the output yarns both have some uneven waveshapes, according to the spectrum analysis theory, any complicated wave can be analyzed as a superposition of many sine waves. Firstly, we study it in case of sine wave and then extend it to the complicated wave. Assume the change of ΔG is a sine wave,

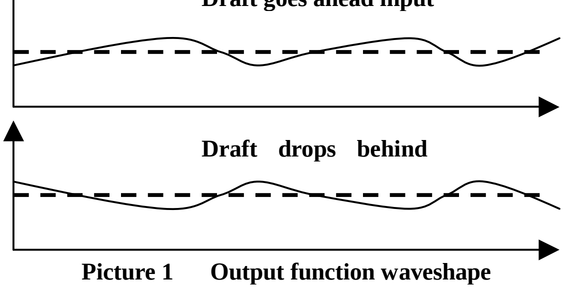
$$\Delta G = A \cdot Sin(2\pi x / L) \qquad \text{Eq(4)}$$

then input function $G_1 = \Delta G + G_0 = A \cdot Sin(2\pi x/L) + G_0$. According to Eq(3), $E(x) = E_0 \cdot (A \cdot Sin(2\pi x/L) + G_0)/G_0$, let $C = G_0 / E_0$, C is the rated output weight, then

$$E(x) = A/C \cdot Sin(2\pi x/L) + E_0 \qquad \text{Eq(5)}$$

According to Eq(2), $G_2(x) = G_1(x)/E(x)$. It shows the output function is decided by the change of the input function $G_1(x)$ and the draft function $E(x)$. The input function can be considered as the combination of two yarns, one is the ΔG and another is G_0. G_0 is an even yarn and no need leveling, but ΔG is an uneven yarn. The draft change, $A \cdot Sin(2\pi x/L)$, can be abstractly understood as the negative compensate to ΔG, that is, it will generate a yarn with equal amplitude but negative phase, these two yarns merge together to make the final yarn even. So it requires these two yarns have the equal amplitude and completely synchronous, or it can't get the proper compensate and will generate new output uneven.

The amplitude of draft function is decided by the sampled value of yarn's uneven signals, and it is easy controlled, while the phase change is difficult to be controlled because of the delay change. Here we mainly study how the phase change of draft function relative to input function affects on the output function, and we try to find some rules to solve problem.



Picture 1 Output function waveshape

When the draft function goes ahead with $2\pi x'/L$ phase change, that is, the leveling point goes ahead its position with x', the delay distance becomes longer. At this moment:

$$E(x) = A/C \cdot Sin(2\pi(x+x^{'})/L) + E_0,$$

$$G_2(x) = \frac{\mathrm{A}\cdot \mathrm{Sin}(2\pi \mathrm{x/L}) + \mathrm{G}_0}{\mathrm{A/C}\cdot \mathrm{Sin}(2\pi(\mathrm{x}+\mathrm{x}')/\mathrm{L}) + \mathrm{E}_0} \quad .$$

At this time the draft function is not synchronous with the input function, and the output function has some uneven waveshape, see picture 1.

When the draft function drops behind with $2\pi x'/\mathrm{L}$ phase change, that is, the leveling point drops behind its position with X', the delay distance becomes shorter. At this moment:

$$E(x) = A/C \cdot Sin(2\pi(x-x^{'})/L) + E_0,$$

$$G_2(x) = \frac{\mathrm{A}\cdot \mathrm{Sin}(2\pi \mathrm{x/L}) + \mathrm{G}_0}{\mathrm{A/C}\cdot \mathrm{Sin}(2\pi(\mathrm{x}-\mathrm{x}')/\mathrm{L}) + \mathrm{E}_0} .$$

The output uneven waveshape is shown in picture 1.

Analysis and building model:

According to above analysis, when the draft function is ahead or behind the input, the output shows some uneven waveshapes, and the two waveshapes have some similarity, that is, they are symmetrical relative to the line $G_2 = C$. So we consider that there may be certain correlation between these two waveshapes, they perhaps obey the same rule, which shows the different result when the draft function is ahead or behind the input.

The correlation and convolution in maths both reflect the relation of two functions in time domain. The convolution analyzes the delay status between two functions. For building the model, we have done both correlation and convolution analysis. The result of correlation analysis can't completely resolve the problem, while the result of convolution analysis is perfect. So here we adopt the convolution to build the model.

Firstly, we do the convolution of the input and the output, the result can change with the leveling point, it means it has some relation with the phase change of draft function. But it can't absolutely reflect the change of leveling point or the change of delay distance. So we multiply this convolution result with the input function, this result is perfect, it just shows the positive and negative value according to the leveling point, and when the leveling point is right, it equals to zero. Thus we build the model, with it we can intelligently estimate the delay distance in drafting process.

According to convolution of two functions $x(t)*y(t)=\int_{-\infty}^{\infty} x(\tau)y(t-\tau)d\tau$, the convolution of two discrete signals is $x(n)*y(n)=\sum_{m=-\infty}^{\infty} x(m)y(n-m)$.

The model can be expressed with

$$F(n\Delta) = [G_1(n\Delta) * G_2(n\Delta)] \bullet G_1(n\Delta) \qquad \text{Eq(6)}$$

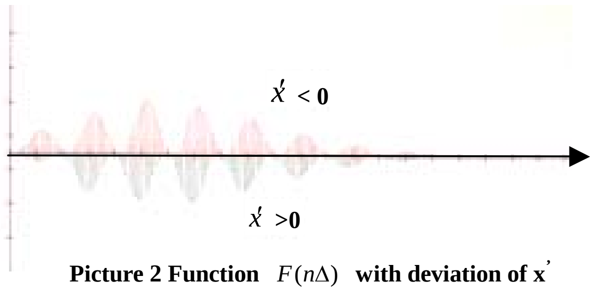
It can also be expressed with

$$F(x) = [G_1(x) * G_2(x)] \bullet G_1(x) \qquad \text{Eq(7)}$$

Using computer we adopt FIFO queue to record delay instead of the traditional memory steel roller and cylinder. When we sampled a point we store it into the FIFO queue. The length of queue is decided by delay distance, it is the amount of sample points. If the interval of sampling is △, the amount of sample points from detect roller to leveling point is AB/△. One sample point will pass by AB/△ sample points when it enters into the queue(being sampled) until it go out of the queue(being leveled). In draft process, we use the model to judge the delay change and then to justify the length of the FIFO queue, so we can control draft arrives synchronously with input.

Analyzing result:

When the leveling point moves to the front roller, that is,



Picture 2 Function $F(n\Delta)$ with deviation of x'

x' >0, at this moment the model value $F(n\Delta)$ <0. The discrete curve lies below coordinate axis, it is showed in picture 2 with the blue curve (This picture is a simulation by computer). It shows the delay distance becomes bigger at this moment, and the draft **E(x)** goes ahead input with $2\pi x'/L$ phase. So it is necessary to increase the length of FIFO queue and adjust the delay in order to make the draft go on synchronously with the input.

When the leveling point move to the back roller, x' <0, the model value $F(n\Delta)$ >0. The discrete curve lies above coordinate axis, it is showed in picture 3 with the red curve. It shows the delay distance becomes smaller at this moment, and the draft **E(x)** drops behind input with $2\pi x'/L$ phase. So it is necessary to decrease the length of FIFO queue to make the draft arrives synchronously with the input.

When the leveling point is perfect, x' =0, the model value $F(n\Delta)$ =0. It shows the delay distance is right, E(x) changes synchronously with G1(x), that is, the just level. The discrete curve is a line equals to zero.

3. THE MODEL'S REALIZATION AND ALGORITHM DESIGN.

If we directly calculate the convolution, the algorithm will be constructed with double circulations. Its time complexity is O(n2). In application, the value of n is very large, and calculation times are comparatively large, so the efficiency of algorithm is very low. According to convolution theorem, convolution of two signals in time domain corresponds with the multiple of signals in frequency domain. So we can use FFT to get the signals spectrum and multiply their corresponding spectrum, then use IFFT, the reverse transaction of FFT, to get the signals convolution in time domain. The algorithm steps are simply showed below:

First: Use FFT to calculate the limited discrete spectrum Xm and Ym .

Second: Multiply Xm and Ym, and get Zm= Xm ·Ym , (0≤m≤N-1)

Third: Use IFFT calculate Zm to get zn , zn this is the convolution of xn and yn .

Forth: Multiply zn with input signal xn to get the last result.

The basic problem of the algorithm showed above is how to use FFT to process the large quantity discrete signals. Now we firstly look how to get the discrete signals and its spectrum.

The discrete signal and its spectrum have one-to-one relation.

According to limited Discrete Fourier Transform DFT, if the sample interval is Δ, then the limited discrete signals are: x(n Δ)= x(0 Δ), x(1 Δ), x(2 Δ), …, x((N-1) Δ), its spectrum is $X(f)=\sum_{n-0}^{N-1}x(n\Delta)e^{-j2\pi n\Delta f}$. The period of $X(f)$ function is $\frac{1}{\Delta}$. If we divide its spectrum into N parts, the interval is $\frac{1}{N\Delta}$, then it limited discrete signal's spectrum is $X(f_m)=\sum_{n=0}^{N-1}x(n\Delta)e^{-jnm2\pi n/N}$, m=0, 1, 2, …, N-1. Whereas we can get the signal through its discrete spectrum Δ $X(f_m)$, $x(n\Delta)=\frac{1}{N}\sum_{m=0}^{N-1}X(f_m)e^{jnm2\pi n/N}$. The limited discrete signal $x(n\Delta)$ and the limited discrete spectrum $X(f_m)$ have one-to-one relation. Simplify the expression, we let $x_n=x(n\Delta)$, $X_m=X(f_m)$, $W_N=e^{-j2\pi/N}$, then the relation can be expressed with $X_m=\sum_{n=0}^{N-1}x_nW_N^{nm}$ and $x_n=\frac{1}{N}\sum_{m=0}^{N-1}X_m\bullet W_N^{-nm}$.

Fast Fourier Transform FFT --- division in time domain.

The calculation amount using DFT method to calculate the discrete spectrum X_m, through it signal x_n is very large. It needs to multiply and add complex number N*N times. We adopt the time domain division FFT method to complete the calculation of spectrum. It can greatly decrease the calculation times.

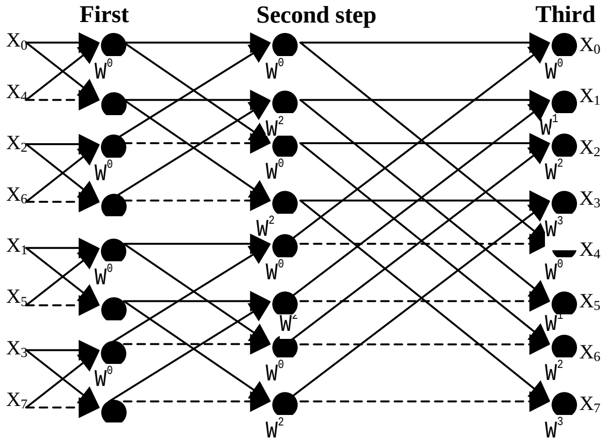
We divide signal into two parts, one is the even part $g_k=x_{2k}$ and another is the odd part $h_k=x_{2k+1}$, k=0, 1, …, $\frac{N}{2}-1$. Then X_m is composed of these two parts $\begin{cases}X_k=G_k+W_N^K\bullet H_K\\X_{k+N/2}=G_k-W_N^K\bullet H_K\end{cases}$, $0\le k\le\frac{N}{2}-1$. This formula suggests if we want to calculate the DFT spectrum X_m, we can get it through its two child part, the X_k of even part and the $X_{k+N/2}$ of odd part. Doing the same division, the calculation of $\frac{N}{2}$ term can be transferred into calculation of $\frac{N}{2^2}$ term. Go on doing this, N→$\frac{N}{2}\rightarrow\frac{N}{2^2}\rightarrow\cdots\rightarrow\frac{N}{2^p}$, until the one term $\frac{N}{2^p}=1$. According to the definition of limited discrete spectrum, the spectrum of one term signal is itself. In reverse, we can recursively calculate the other spectrums through one term spectrum using formula showed above. At last we get the final X_m.

Now we discuss how to get the one term signal arrange. We give an example of 8 points discrete signal. If the original arrange of these 8 points are (x_0 , x_1 , …, x_7), the first division is two parts (x_0 , x_2 , x_4 , x_6 | x_1 , x_3 , x_5 , x_7) , and we divide the 4 term signal into even and odd parts, it becomes (x_0 , x_4 | x_2 , x_6 | x_1 , x_5 | x_3 , x_7) , so we get it 2 term arrange. Because the 2 term signal arrange is itself whenever it is ordered by even or odd, we get its one term arrange (x_0 , x_4 | x_2 , x_6 | x_1 , x_5 | x_3 , x_7) . We see the binary of every signal's order; it is showed in picture 3. We can find the one term signals' order and the original signals' order have some relations. If we reverse the binary of original signals' order, we can justly get its one term order. So we can use the bit-reverse method to get the signal's one term arrange.

0	1	2	3	4	5	6	7
000	001	010	011	100	101	110	111
↓	↓	↓	↓	↓	↓	↓	↓
000	001	010	011	100	101	110	111
0	4	2	6	1	5	3	7

Picture 3 The order's binary comparison.

If one term spectrum arrange is (x_0 , x_1 , ⋯, x_{N-1}) , firstly we calculate 2^1 term spectrum. We group it by every two points ($\overline{x}_0$, $\overline{x}_1$ | $\overline{x}_2$, $\overline{x}_3$ | ⋯ | $\overline{x}_{N-2}$, $\overline{x}_{N-1}$). In each group, the front half part can be looked on as Gk and the back half part as Hk, so we can calcultate 2^1 term spectrum. Then we calculate 2^2 term spectrum through 2^1 term spectrum. Generally, we group each term and then calculate. Go on doing the same work, until we calculate $2^k=N$ term spectrum. The times we repeat doing so is $k=\log_2 N$. Now we give an example N=8 to show this deducing process. See picture 4. In picture the arrow's direction shows the calculation direction, W^k is the weight of term, the horizontal solid line means an add operation and the horizontal broken line means a subtract operation.



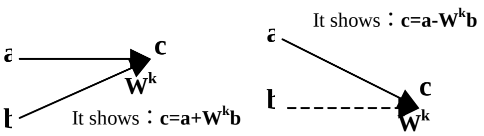
Picture 4 Division in time domain FFT

The basic FFT operation unit is called butterfly operation. It is showed in picture 5 below.

Other problems about algorithm realization.

From picture 5 we can see FFT calculation has a character of the only position operation, that is, the result is stored in original position of input data. This can save memory.

When we use IFFT method to calculate the convolution result to get the signals, we should change W^k into W^{-k} , and multiply the result by $\frac{1}{N}$. So we need only design one algorithm to realize FFT and IFFT calculation.



Picture 5 The basic FFT operation unit Butterfly operation

The weight of term $W_N{}^r = e^{-jr2\pi/N} = \cos 2\pi r/N - j \cdot \sin 2\pi r/N$, r=0，1，⋯，N/2, we give a recursive method to calculate it. Let $c_r = \cos 2\pi r/N$, $s_r = \sin 2\pi r/N$, $W_N{}^r = c_r - j \cdot s_r$, then $W_N{}^{r+1} = c_{r+1} - j \cdot s_{r+1}$ Because $W_N{}^{r+1}$ = $W_N{}^r$ · $W_N{}^1$ = ($c_r - j \cdot s_r$)($c_1 - j \cdot s_1$)=(c_1 c_r - s_1 s_r)－ j ·(c_1 s_r ＋ s_1 c_r), considering apart the real and the imaginary part, so $\begin{cases} c_{r+1} = c_1 c_r - s_1 s_r \\ s_{r+1} = c_1 s_r + s_1 c_r \end{cases}$ Meanwhile $c_0 = 1$, $s_0 = 0$, $c_1 = \cos 2\pi/N$, $s_1 = \sin 2\pi/N$. We can use above formula and the initial value to recursively calculate c_{r+1} and s_{r+1} , and then calculate the weight of $W_N{}^{r+1}$.

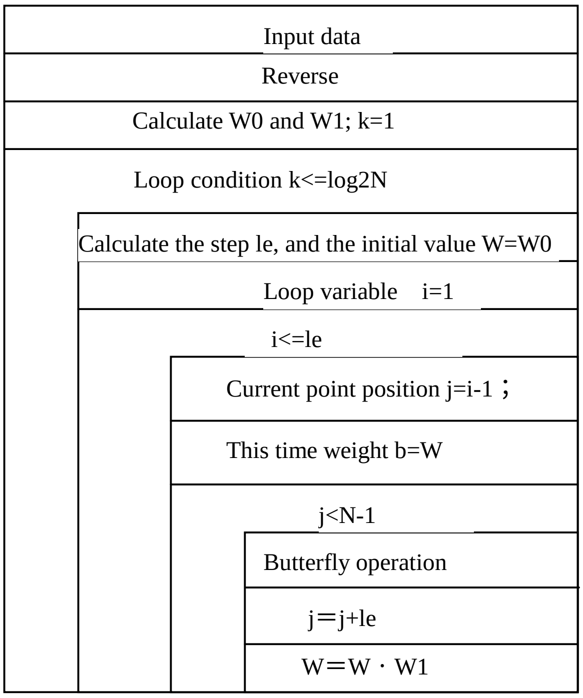
Algorithm of reverse rearranging.
According analysis above, if we want to get one term spectrum arrangement, we need to rearrange the signals, that is, reverse the binary of the points order. This rearranging method will greatly effect on the whole algorithm efficiency. Now we introduce a method to realize it. Look picture 4 above, in original data order's binary, the next data order always bigger 1 than the data in front of it. It also means the next data order can be looked on as its front order adding 1 to the low bit position. Whereas in last data order' binary, the next data order can be looked on as its front order adding 1 to the high bit position, of course this addition goes on from the high bit to the low bit, so it is called "reverse add".

Through analysis above, we design a FFT algorithm, its flow chart is showed in picture 6.

Analysis the algorithm efficiency.
With FFT method calculating the discrete signals spectrum, it needs N · $\log_2$N times addition and multiplication of complex number. In the model realization there are 3 times FFT operation and N times multiplication, so the whole time efficiency: addition operation is 3N · $\log_2$N, multiplication operation is 3N · $\log_2$N+N. While the direct method to calculate convolution needs N*N times multiplication and 1/2*N*N times addition. When N is very large, its efficiency is very low, while FFT method can greatly improve the time efficiency.

4. CONCLUSION

In this paper, we studied the auto-level technique in theory. By analyzing output uneven waveshape caused by the draft function's phase change, we got the elicitation to build a mathematical model based on convolution theory. Through this model we can realize the intelligence control of delay; it resolved the problem, which can't automatically adjust delay in any control traditional systems. The realization of this model is based on sampling and processing the yarn's uneven signals, it requires to sample the right yarn's signals， and try the best to sample the short middle segment uneven. So the detecting unit should have special spectrum characters. In addition, practicing this model, except for computer technique, it required cooperate with the advanced speed governing technique. Study and develop on alternating speed governing aspect, the automatically control capacity of autoleveler system will become better and better. So we still have many problems to discuss and study.



Picture 6 FFT flow chart

5. REFERENCE

[1] ZhuBorong, Autoleveler, Textile publishing company (in Chinese).
[2] W.D.Stanley, Digital Signal Processing, Science publishing company.
[3] ZhongZhili, Correlation analysis of yarn uneven, Journal of Tianjin Textile and Technology Institute(in Chinese).

Han Meigui was born in Heilongjiang Province in 1973. She studied the profession of textile product design in North-West Textile Technology Institute in 1990. When she graduated in 1994, she began her postgraduate study in the same school, which was the application of computer technology in textile process. After graduating on April, 1997, she engaged in a teaching career in Computer Engineering College, Jinan University. Now she mainly studies on computer technology and computer education.

The Design of Power Battery Management System Based on Distributing CAN Bus

Zeng ChunNian Chen Yu Qiao Guoyan
School of Automation, Wuhan University of Technology
Wuhan 430070, Hubei,China
Email: mars880716@163.com **Tel:** 013657299841

ABSTRACT

Power Battery Management System based on distributing CAN Bus is presented in this paper, which also expatiates the function and structure of the system and each of its form cell and discusses the system's work principle and some practical questions. This system is a new battery management system with practicality, which has many good characteristics, such as a logic design, advanced methods, legible structure, high reliability, easy servicing and etc.

Keywords: distributing, battery management, µC/OS-II, CAN Bus

1. INTRODUCTION

The big capacity storage cell that is commonly used as the power in vehicle is called as traction battery. The power battery is one of the important parts of EV or hybrid EV. In the further development of EV, the exigent problem which we must resolve is how to reasonably use the battery, how to take full advantage of the capacity of the battery sets and how to prolong the longevity of the battery. The traction battery generally is a battery module composed of many monomer batteries in series and a vehicle used battery package composed of a battery management system (BMS) with a micro control as its core.

Generally speaking, the main function of battery management system is to estimate the SOC (State of Charge) of the batteries, to measure the total voltage and current of the battery pack, to ensure the safety of the pack, to keep balance charge among each battery, to acquire and process data, and to implement communication. When developing prototype vehicle, a 25 KW fuel cell EV type Elysée, we chose 24 groups (each group 12V) of 12Ah monomer NiH battery to form a 288V battery package as vehicle used auxiliary energy. The following is the presentation of its distributing battery management system.

2. CONFIGURATION OF DISTRIBUTING BATTERY MANAGEMENT SYSTEM

The battery package is a close collectivity, which can only through analyses the information inside the package to grip what happened inside the package. In this aspect the more information is captured from the package, the easier control is achieved for the system. However, the BMS would be more numerous and complex. In the other aspect, to reduce cost and improve reliability of the system, the simplification of the system is needed. According to the performances of the NiH battery and the method of equalization charge, the reasonable number of the voltage and temperature measure points are needed to be chosen for the monomer batteries in the package. In each battery group two voltage and a temperature measure points is engaged, so there will be a great deal of measure points in the battery package. If centralized control system is adopted in acquisition and processing data in the system, too many joints will occur in the battery package, and will make the management system confused and bring extreme trouble to the mount, debugging and maintenance. Hence, distributed control system basing on CAN Bus [1] is used in the management battery system. The simplified configuration diagram of system is shown in Figure 1. In the system, the CAN Bus network is made up of seven bottom ECUs (Electronic Control Unit) and one top ECU, the internal CAN bus with bus topology. Transmission medium is twisted-pair, and transmission protocol is CAN2.0B protocol. The top ECU has double CAN controllers, one is connected with the bottom ECU to form the CAN network inside of the battery management system，and the other CAN controller composes with the other ECU the EV CAN Bus net, whose net topology structure is star form, whose transmission medium is plastic optical fiber and whose transmission protocol is also CAN2.0B protocol.

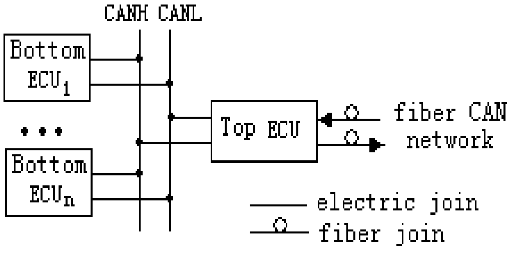


Figure 1 Battery Manage System block diagram

3. BOTTOM ECU (ELECTRONIC CONTROL UNIT) DESIGN

3.1 Configuration of the bottom ECU

In the battery management system, the main functions of bottom ECU are to measure and capture the data of voltage and temperature of the batteries and to send the values to the top ECU. The core of the bottom ECU is an 8-bit microcontroller—P87C591 [2], which is manufactured in an advanced CMOS technics, and which is designed for automotive and general industrial applications. The P8xC591 has on-chip 6-input 10-bit ADC and CAN-controller. As derived from the 80C51 microcontroller family, it uses the powerful 80C51 instruction set and includes the successful PeliCAN functionality of the SJA1000 CAN controller from Philips Semiconductors. The improved internal clock prescaler of 1:1 achieves a 500ns instruction cycle time at 12 MHz external clock rate. Due to the internal integration of an A/D converter and an independent CAN-controller，the system hardware circuit design is predigested.

In this paper, the battery package, which should be tested, is the 24-group (each group 12V) monomer NiH batteries and each group includes 10 monomer NiH batteries. If voltage and temperature of each battery was measured, the system would be numerous and jumbled. So it is feasible to choose a group of batteries as a point to measure voltage and temperature. In the group, even there is a battery in trouble, the management system can tell accurately the battery from others. In this way, there are 24 voltage measurement points and 24 temperature measurement points, which are managed by the 7 bottom ECU including 6 ECU to measure voltage and 1 ECU to capture values of temperature. Meanwhile, the bottom ECU communicates with top ECU by CAN Bus, thus formed a disturbing CAN net and facilitated the mount, debugging and maintenance. The block diagram of the bottom ECU to measure temperature is shown in Figure 2. The configuration diagram of the battery management system bottom temperature measurement ECU is shown in Figure 2, since the voltage measurement ECU configuration diagram is similar to it, only the voltage measurement circuit is need to be changed into temperature measurement circuit, we will not illustrate it in details.

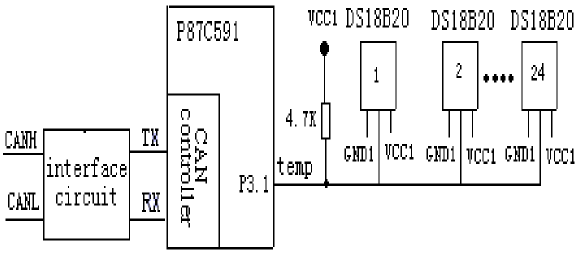


Figure 2 Bottom temperature measurement ECU of Battery manage System

3.2 Hardware design

The hardware circuit design is mainly comprised of the design of temperature measurement circuit, voltage measurement circuit and communication interface circuit of CAN bus.

A new type of one-Wire digital temperature sensor DS18B20 [3], which was produced by Dallas Semiconductor of American, was used to measure the temperature in the circuit. The interface between the DS18B20 and the single chip processor is very simple. As shown in Figure 2, information is sent to/from the DS18B20 over a 1-Wire interface, so only one wire is required to be connected from a central microprocessor to a DS18B20, using a 4.7k pull-up resistor tied to the I/O pin. There are two ways to supply the DS18B20. The first is parasitical power and another is external power. In the use of parasitical power, a MOSFET is required to supply the sufficient operating current and the number of the DS18B20 on the 1-Wire bus is limited. So the method of external power supply is adopted in the design.

In the voltage measurement circuit the value of the measured voltage is converted into 0~5v through the resistor net. And then this signal is sent to the filter circuit and the protective circuit. After that the signal is input to the A/D converter in the P87C591.Thereby the voltage measurement has been finished.

Communication interface circuit of CAN bus sends the data it has measured to the battery management system periodically. The structure of the circuit in which CAN bus is adopted is very simple and the method of communication is very flexible. Simultaneously the CAN bus can strongly suppress noise and disturbance. The communication between every ECU becomes convenient when CAN bus is applied to the EV. The transmission medium of CAN bus is twisted pair.

3.3 Software design

The software, which is programmed with the C language of 8051 series, is simple and easy to understand. Its emphasis is the digital temperature sensor DS18B20, which is a 1-Wire device and requires strict 1-Wire protocols to insure data integrity. First the bus master transmits a reset pulse to make all the DS18B20 on the 1-Wire bus reset, and then issues the ROM function command on the 1-Wire bus. The DS18B20 whose serial number is matched with the reference is active and prepares to receive the next memory function commands. The memory function commands control the work states of the DS18B20 who has been selected and allow the bus master to finish the temperature conversion and read the DS18B20's converted result. In all processes three important steps are as follows: initialize the DS18B20, write the DS18B20 and read the DS18B20, which must strictly obey the DS18B20's time slots, otherwise the correct conversion result will not be acquired.

4. TOP ECU DESIGN

4.1 Configuration of the top ECU

The top ECU of the battery management system is in charge of measuring the total voltage and current, estimating the SOC of the battery module, analyzing the temperature and the voltage of every battery. Based on these data the ECU can take corresponding action such as alarm, controlling the equalizer to start or stop and controlling the fan on the battery box. Furthermore this ECU also communicate with other ECU on the EV and the bottom ECU of the battery pack through CAN bus, so every system can share the information of EV. The top ECU system's block diagram of the battery management system is shown in Figure 3. The kernel of the top ECU is TMS320LF2407DSP [4], which is small and fast in processing speed, and the A/D converter and CAN bus controller are integrated in the slice. ADC0 and ADC1 are used to measure the voltage and the current of the battery module. D0 is used to control the start or stop of the equalizer. D1 is used to control the ECU of the battery package to startup or stop. CAN1 uses SJA1000 to spread to outside and CAN2 is the CAN controller in the slice of the LF2407. And the fiber using for communication is plastic fiber with 1mm diameter. The topological structure of the network is star form. The whole delay of the transmitting and the receiving is 120ns.

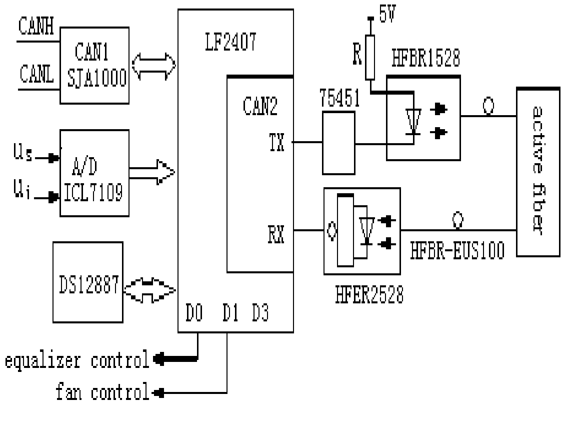


Figure 3 Top ECU block diagram of Battrery Manage System

4.2 Hardware design

The hardware is mainly made up of the following several parts: the DSP unit of TMS32LF2407, the interface circuit of SJA1000 [5](CAN controller), the interface circuit of A/D converter, the interface circuit of the real time clock and the interface of optic fiber CAN bus.

The unit of TMS32LF2407 consists of minimal system of the CPU and other peripheral circuits such as the CPU, bus transceiver with 3-state output, large capacity external RAM and decoder of GAL.

The external CAN bus controller of SJA1000 is in charge of communicating with the bottom ECU of the battery package in the battery management system, and receives the data about the temperature and the voltage of the battery transmitted from the bottom ECU via CAN bus. Because of the multiplexed address/data bus of SJA1000 and the self-existent address/data bus of DSP, the time slots of the SJA1000 must be simulated when the DSP reads/writes data from\to it. SJA1000 is looked upon as the I/O device of the DSP, so the 8 bits address/data bus of SJA1000 is connected with the low 8 bits data bus of DSP through level converter 74LVC16245. The SJA1000 can be read or written just when the time slots of SJA1000 are strictly referred, and the transmitting direction of the data is controlled by the DSP R/W signal.

The A/D converter module is mainly used to measure the total voltage and current of the battery package. Over 300 volts voltage is transformed into 5V voltage signal through the sensor and enters into the top ECU of the battery package. Before entering into the top ECU of the battery pack, the current signal is transformed into a ±5V signal through sensor. In order to calculate the value of SOC accurately, the sampling precision is expected relatively high.

The calendar/clock chip DS12887 includes high precision clock internally, and its inner cell can last more than 10 years. Besides, the address map of this chip consists of 114 bytes of nonvolatile static user RAM. The reason why this chip is adopted in the top ECU of the battery management system is that the chip can protect the SOC data in Power-down mode, and can contain the RTC time. The system contains the time in order to choose the SOC arithmetic, and to ensure the accuracy of the SOC result. If the interval between the last power-down (viz. the time when the batteries stop last working) and the current time is over 24 hours, viz. the state of the batteries has become more steady, the arithmetic of SOC in this system will obtain the initial SOC value again by opened circuit voltage arithmetic, otherwise, it will still adopt the last saved data as the initial SOC value.

The fiber-CAN interface is mainly used for the communication with the units of the other systems in the EV. The TMS320LF2407 DSP device integrates internally a CAN controller to get CANTX and CANRX signals. The CANTX signal needs to be connected to the HFBR2528 receiver block in order to transform the electrical signal into photic signal that is transferred in fiber. Equally, the CANRX signal needs to be connected to the HFBR1528 transmitter block to transform the photic signal into electrical signal that would be processed in the DSP system. Before CAN TX, DS75451 is used for enhancing driving capacity.

4.3 Software design

The software design in the top ECU adopts the Embedded Real-Time Operating System named µC/OS-II [6], which increases the capability of real time and the security of this battery management system. The µC/OS-II is a source opened and embedded real-time operating system kernel that is compiled by an American named Jean Labrosse. The µC/OS-II, which is a preemptive multitask OS, is ROMable, scalable and transplantable, is so powerful that it can sustain as many as 56 user tasks and supports the communication among multiform processes in common use, such as semaphores, message mail box and message queue, and currently has been successfully applied to numerous commercial embedded systems.

The main function of the software in the top ECU of the battery management system includes: receiving and processing the voltage and temperature data from the bottom ECU transmitting by the bottom CAN bus in interrupt mode; getting the gross voltage data and current data of the batteries by the time sampling (the higher the frequency of sampling is, the higher the accuracy of the SOC value measuring is); calculating the SOC value of the battery by the voltage data, current data and temperature data; transmitting SOC value and some other information to other CAN nodes in the control system by the top fiber-CAN at intervals of 50 ms; transmitting the voltage and temperature information of the pile by the top fiber-CAN at intervals of 1s; proportioning the voltage of the pile whose voltage is on the high or low side; inspecting abnormal state and sending warning signal at any moment if it is necessary.

The author divides the software function of the top ECU into four tasks to complete, and distributes different PRI to different task. In the µC/OS-II, multitask is executed at the system background, while the interrupt is to be the foreground application. Every task in the software is an infinite loop. The system illustrated in this paper has three interrupts: timer interrupt, CAN receiver interrupt of SJA1000, and A/D convert ending interrupt. The software block diagram of the top ECU is shown in Figure 4.

5. CONCLUSION

At present, the character of the power battery adopted in the R&D (Research and Development) of EV is always rather complex. In order to measure the real time and exact SOC value, and to prolong the use life of the batteries, high quality battery management system is required to develop.

In this battery management system, seven bottom ECUs communicate with the top ECU by CAN bus and form a distributed CAN network, while the top ECU communicates with the ECU of the other systems in the EV by the fiber-CAN and thus as a whole make up of a distributed battery management system. Consequently, the whole battery management system has powerful functions and very legible structure that improve the security and reparable capability of the system. Indeed, this system is still relatively complex, and with the development of the monomer cell technology, the management system can be further simplified.

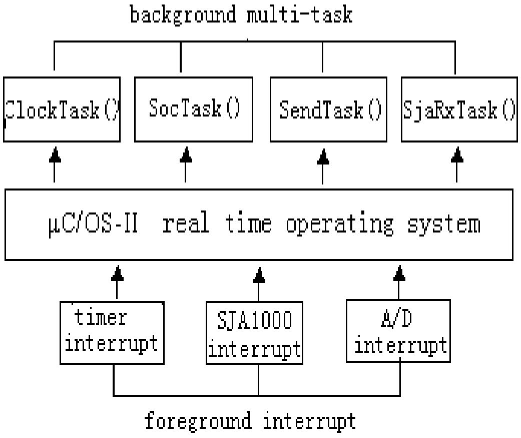


Figure 4 system software block diagram

6. REFERENCES

[1] Kuanming Wu, CAN Bus Principium and The Application System Design, Beihang University Press, Beijing, 1996
[2] PHILIPS Semiconductors Company, Single-chip 8-bit microcontroller with CAN controller, 2002
[3] 1-Wire Digital Thermometer Principium and Application, Electronic Technique Application Press, 2000
[4] Heping Liu, TMS320LF240xDSP Structure, Principium and Application, Beihang University Press, Beijing, 2002
[5] PHILIPS Semiconductors Company, SJA1000 Stand-alone CAN controller, 2002
[6] Labrosse Jean, μC/OS-II－Source Code Opened Real Time Embedded Operating System, China Electricity Power Press, 2001

Application of the Distributed Parallel Processing In the DNA Sequence Alignment[*]

Mao Liming[1], Wang Zhongjun[2], Guo Qingping[1]
[1]School of Computer Science and Technology, Wuhan University of Technology Wuhan 430063
[2]School of Science, Wuhan University of Technology Wuhan 430063

ABSTRACT

DNA sequence alignment is foundation of DNA sequence analysis. It is an important method of exploring the information of the DNA sequence. The Blast algorithm is an important algorithm in DNA sequence alignment. The key point of the method is to make DNA sequence to many sequence pairs for alignment and then join them together. This paper improves it and presents a local parallel algorithm in DNA sequence alignment based on the Blast algorithm. The paper analyses the advantage of the parallel algorithm to the other parallel algorithm by splitting the DNA sequence database. The theory analysis of the algorithm is described at the end of the paper.

Keywords: DNA sequence alignment Blast algorithm Distributed Parallel algorithm

1. INTRODUCTION

People have got many of DNA sequences with the developing of the biology. DNA sequence alignment is an important method to get the character, function of DNA sequence while it is the important domain of the study of the Bioinformatics. People can put the sequence which unknowing the function and character together with the sequence which knowing the function and character and get the comparability by alignment the two sequences. The people can get the function and character of sequence which unknowing the function and character by the comparability. The sequences in the sequence database are always very long and people always compare a sequence to the all sequences in a sequence database. So the time of the alignment is always very long. How to improve the speed of the alignment is a very important question. This paper use the distributed parallel technology to improve the speed of the alignment

2. THE DESCRIPTION OF THE ALIGNMENT OF DNA SEQUENCE

DNA is constituted of many Bases. There are 4 kinds of bases. These 4 kinds of bases can be denoted by 4 kinds of letters. These 4 kinds of letters are A,G,C,T. The DNA sequence is a string made by the 4 kinds of letters. For example, the AAGCT is a DNA sequence.

DNA sequence alignment is the way that compares a group of DNA sequences and gets the comparability of them by some algorithm [1]. There are some kinds of DNA sequence alignment. There are the multiple sequences alignment, double sequences alignment, local sequence alignment and global sequence alignment. This paper discusses a kind of them: double sequence local alignment. The follow is the definition of it:

There are two DNA sequences. One is the source sequence knowing the function and character. Other one is the comparing sequence unknowing the function and character. The main goal is to get some sequence pairs and analyses the function and character of the comparing sequence. The sequence pairs must satisfy the conditions

(1). Two sub sequences in the sequence pairs are both made of A, G, C, T. and gap.

(2). Two sub sequences in the sequence pairs have the equal length

(3). Sequence pair's score is higher than a giving score

(4). Then can get a subsequence of source sequence and a subsequence of comparing sequence by get rid of the gap of the two sub sequences

We must define an integral mechanism to measure the quality of the alignment of the DNA sequence. If the score is higher, the quality of alignment is better and the comparability of the DNA sequence is better. There are many kinds of the integral mechanisms. This paper adopts the follow integral mechanism: The initialize score of per DNA sequence pair is 0. If the corresponding bases of the DNA sequence are same, 1 is added on the score. Otherwise, 1 is subtracted on the score. If there is a gap in the two bases, 1 is subtracted on the score. Example in figure 1:

Before insert the gaps	
Sequence1 comparing sequence	AGGCTGGACG
Sequence2 Source sequence	AGGCTGAGG
The score of the DNA sequence pair is 6.	
After insert the gaps	
Sequence1 comparing sequence	AGGCTGGACG
Sequence2 Source sequence	AGGCTG –AGG
The score of the DNA sequence pair is 8.	

Figure1 Insert the gaps to get better alignment

In the DNA sequence, there is an important conception: sub string. Example: There are two sequences: A and B. If B is the part of the B and there isn't gap in the A, then A is called sub string of B. In the paper, the conception is always used.

3. BLAST ALGORITHM

The alignment of the DNA sequence is the foundation of the analysis of the DNA sequence. There are many kinds of methods to solve this problem. Such as BLAST algorithm and Smith-Waterman algorithm and dynamic layout algorithm. Blast algorithm is a classic method and is used very widely [2].

First, BLAST algorithm must code the two DNA sequences. Because there are 4 kinds of the Base, can use 4 kinds of the string whose the length is 2 to represent the 4 kinds of the abase. The representation is following.

[*] Research supported by the Natural Science Foundation of China (NSFC Grant No 70371063, 60173046).

A	00
G	01
T	10
C	11

The bases of the DNA sequence can be converted base on the representation method and can get a binary bit string. This binary bit string is the code of the sequence. Example:

DNA sequence: ACTTGCTACGTGCA
Binary code: 0011101001111000110110011100

After coding the DNA sequence, must define two point of per DNA sequence. The two points are called Left point and Right point and they point a base of the DNA sequence respectively. They can move belong the DNA sequence around left and right. But the moving must satisfy the following conditions:

1. The left point must point the base that its index in the DNA sequence must be odd.
2 The distance between the left point and right point can't be changed. It is 8 in the paper.
3. The moving step of the two points is even.

After having two points of the DNA sequence, can create a index table of Source DNA sequence by moving the two point step by step .The initialize place of the left point is 1. The initialize place of the right point is 8. The move step distance is 2. The binary bit string between of the two points is a sub string of the DNA sequence and it can be changed as the moving of the two points .The index table is made of the records. The structure of the record is:

A	B	C	D

A is the start place of the sub sting in the DNA sequence
B is the end place of the sub string in the DNA sequence
C is the sub string
D is number of the sub string

The start place of the left point of comparing sequence is 1 and the start place of the right point of comparing sequence is 8.Then compare sub string of the comparing sequence to the sub string with the record of the index table one by one. If the record is same with the sub string, move the left and right point in the two sequences simultaneously until there are two different bases. One is in comparing sequence and other is in source sequence. Then take down the all records that are same with the sub string. After that, move the two points backwards along the comparing sequence by step 2.The BLAST algorithm loop the process until the right point points the end of the comparing sequence.

By use this method, can get many high score sequence pairs. For getting the sequence pair with higher score, must join the sub string pair by gaps.

A is sequence and B is other sequence. M and N is a sub string in a high score sequence. The start place of M in A is P_{M1} and the end place of M in A is $P_{M2.}$ The start place of M in B is Q_{M1} and the end place of M in B is $Q_{M2.}$ The start place of N in A is p_{N1} and the end place of N in A is $P_{N2.}$The start place of N in B is Q_{N1} and the end place of N in B is $Q_{N2.}$ If the sequence pair satisfies the conditions, they can be joined

(1) $P_{N1} > Q_{M1}$ $P_{N2} > Q_{M2}$

(2) $P_{N1} - Q_{M1} = P_{N2} - Q_{M2}$

(3) $P_{N2} - Q_{M2} < 4$

If M and N can be joined by gaps, insert P_{N2}- Q_{M2} gaps into the M between N. The bases between M and N in the B can be changed by gaps. Then can get a sub string K whose length is $Q_{N2.} - P_{M2}$. The start place of the sub string K in the A is P_{M2} and the end place of it is $Q_{N2.}$ It and the sub string L of the B is a good sequence pair. The start place of the sub string L in the B is P_{M1} and the end place of it is $Q_{N1.}$

Base on that method, get the sequence pair M and N, which can be joined from high score sequence pairs. Then join the M to N and get sub string K and sub string L. K and L is the better sequence pair. Then M is replaced by K and N is replaced by L. Loop this method until there aren't sub string pairs like M and N in all sequence pairs.

BLAST algorithm isn't an algorithm of the sequence alignment only. It is a tool of sequence database research. It is always used in the sequence database research. There are many kinds of tools of the sequence database research. Such as FastA and PCI-BLAST[3].

4. PARALLEL ALGORITHM

For improving the speed of the DNA alignment, the distrusted parallel process can be used into the DNA alignment. In this part, the paper introduces two kinds of parallel methods to fast the speed of DNA alignment. One is by splitting the DNA database. This method is now widely used in the DNA alignment. Other is by splitting sequence. The advantage and disadvantage between the two kinds of methods are described at the end of the part.

4.1 Split the DNA Database

4.1.1 The Ideal of the Method [4]

In the alignment, one sequence is often compared to the every sequence in the DNA database. Because the number of the DNA sequence is very large, the speed of the alignment is very slowly. For fasting the speed, the database can be split some parts and put the parts on the some computer nodes. Then the alignment task can be split. The sequence can compare to the sequences in the specifically database on every computer node simultaneously.

4.1.2 The Description of the Parallel Algorithm

The parallel algorithm can be divided the master process and the slavery process:

Master process:
step1:
Send a part of the DNA sequence database to the every slavery process.
Step2:
Receive the alignment result from every the slavery process.
Step3:
Analyses the result and out the result.

Slavery process:
Step1:
Receive a part of the DNA sequence database from the master process.

Step2:

Compare the sequence to the every sequence in the local part of the DNA sequence database on the computer node using the BLAST algorithm.

Step3:

Send the alignment result to the master process.

4.2 Split the Sequence

4.2.1 The Ideal of this Method

For improve the speed of the alignment of the Blast algorithm, the algorithm can be improved and can get a local parallel algorithm. The most time spend on the getting many high score sequence pairs. All sub strings of one sequence must be alignment to all records of in the index table. If the length of the two sequences is L_1 and L_2, the algorithm complexity is $O(L_1L_2)$. If want fasting the algorithm, must reduce the time spending on the process.

For reducing the time of the step, can divide the index table in to some parts and put this parts of the index table on some computers. In per computer, the sequence need to alignment the local part of the index table. The processes of the alignment on the computers are at the same time. The serial algorithm is changed in to the parallel algorithm.

The index table can be divided in to some parts by dividing the sequence in to some segments. For example, the index table can be divided in to m parts and put the m parts on to the m computers. So the sequence can be divided in to m segments. The start place and end place of k-th segment in the sequence are $[(L_2\text{-}3)/m](k\text{-}1)+1$ and $[(L_2\text{-}3)/m]k+4$.The start place and end place of the last segment in the sequence are $[(L_2\text{-}3)/m](m\text{-}1)+1$ and L_2. The m segments are put on the m computes. The computer can create the local index table. The process is run at the same time.

4.2.2 The Description of the Parallel Algorithm

The ideal of the improving can be translated in to the asynchronous parallel algorithm base on master-slavery mode[5]. There are one master process and some slavery processes. The master process dispatches the mission to the slavery process and receives the compute result from the slavery processes. Slavery processes mainly complete the missions from the master process and give the compute result to the master process. The run of the all processes aren't synchronous. They can run, stop and end base on the different phases of the compute.

4.2.2.1 The Algorithm of the Master Process

Step1:

The master process divides the sequence in to the m segments.

Step2:

The master process dispatches the all m segments to the all slavery processes except the first segment. The master process dispatches two data to the every slavery process. One data is the start place of the sub string in the sequence and other one is the end place of the sub string in the sequence.

Step3:

The master process alignment the first sub string and local sequence use the serial Blast algorithm and get the result of the alignment.

Step4:

The master process receives the result of the alignment from the slavery processes one by one. For the load balance, the master process don't specify which slavery process when receive the result send by slavery process. Load balance is an important in the parallel technique. [6]

Step5:

The process of joining is run.

4.2.2.2 The Algorithm of the Slavery Process

Step1:

The slavery process receive the sub string of sequence which given by the master process.

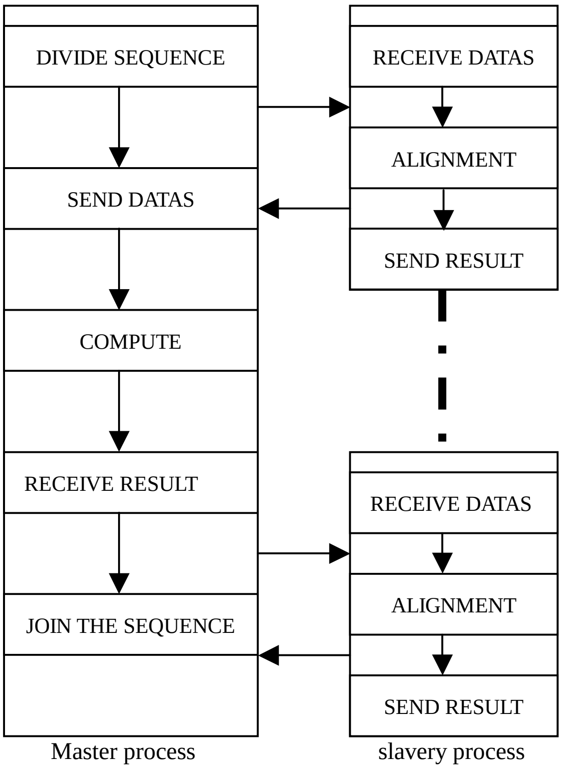
Step2:

The slavery process compares the sub string which receives from the master process to the sequence by using the serial Blast algorithm.

Step3:

The slavery process gives the alignments result to the master process

The figure2 can describe the parallel algorithm clearly.



Figur2 the parallel algorithm

4.3 The Advantage of the Method by Splitting the Sequence

In the process of the DNA sequence alignment, the sequence must be compared to every sequence in the DNA database. The lengths of the sequences in the DNA sequences database are very different. One sequence may be very long, but other may be very short. So the time of the sequence comparing to the all sequences in the part of the DNA sequence database on every computer node are very different. One may be is very long but other is very short. It must lead to the load of the parallel algorithm is in imbalance.

If the parallel algorithm split the sequence, the length of the subsequence on the every computer node is same. So the time of the alignment is same approximately. It makes the load of the parallel algorithm is in balance better. The load balance load is very important in the design of the parallel algorithm.

5 ANALYSIS OF THE PARALLEL ALGORITHM AND CONCLUSION IN THEORY

Base on the description of the parallel algorithm, we know the parallel algorithm is a parallel algorithm which there isn't communication between the slavery processes. Then the analysis of the parallel algorithm is follow:

The parallel algorithm has 3 phases: communicate phase, compute phase and communicate phase.

Phase 1: communicate phase

There are p slavery processes. The master process sends two data to the every slavery process. Then can get the time.

$$t_{comm1}=p(t_{startup}+2t_{data})$$

t_{comm1} is the communicate time of the first phase. $t_{startup}$ is the startup time of the every process. t_{data} is the time of sending the data to every process.

Phase 2: compute phase:

Every process completes the compute.

$$t_{comp} \le \frac{max_1 * m * n}{p+1} + t_{pj}$$

t_{comp} is the compute time. m and n are the length of the two sequences. Max1 is the max time of finding a same sequence pair. t_{pj} is time of joining sub string.

Phase 3: communicate phase

Last, every slavery process sends alignment result to the master process.

$$t_{comm2}<<4pmax_2t_{data}$$

t_{coom2} is the communicate time of the last phase. A is supposed the slavery process that can find most high score sequence pairs. Then max_2 is the number of which A can find high score sequence pairs.

Sum up, the total time of the parallel algorithm t is

$$t \le p(t_{startup}+2t_{data})+\frac{max_1 * m * n}{p+1} + 4pmax_2t_{data} \ (1)$$

The main body is $\frac{max_1 * m * n}{p+1}$

The time of serial algorithm is

$$T=mnmax_1 \ (2)$$

Because

$$t_{pj}<< mnmax_1$$

The algorithm complexity is O (mn) in theory.

Because m and n are very big and

$$max_1<<m \qquad max_2<<n$$

$$t_{pj}+4pmax_2t_{data} << \frac{max_1 * m * n}{p+1}$$

$$p(t_{startup}+2t_{data}) << \frac{max_1 * m * n}{p+1}$$

The algorithm complexity of the parallel algorithm is O(mn/(p+1)). The speedup of the parallel algorithm to the serial algorithm is

$$\mathbf{K}=\frac{T}{t} \ (3)$$

K is a bit smaller P+1 because the communication between the slavery processes.

From the formula (1) (2) and (3), we can find that parallel algorithm can fast the alignment of the DNA sequence .If we don't consider the communicate time; the speedup is p+1 perfectly. But if we consider the communicate time, we can also find that the differences between with

$$p(t_{startup}+2t_{data})+ 4pmax_2t_{data}$$

and $\frac{max_1 * m * n}{p+1}$ is smaller and smaller as the p is bigger and bigger. So the difference between with the speedup of the parallel algorithm to the serial algorithm and the p+1 is bigger and bigger. So the communicate time is a very important factor in the design of the parallel algorithm.

6、REFERENCES

[1] T.K Attwood & arry-Smith Introduction to bioinformatics LuoJingChu translate Peking University Press 2002.131~145

[2.] Stephen F. Altschul Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Research, 1997, 25(17): 3389–3402

[3] ATLSCHUL, S.F., GISH, W, MYES, E.W. and LIPMAN, D,J (1990) Basic local alignment search tool, Journal of Molecular Biology,215 403~410

[4] R.D.Bjornson A.H.Sherman S.B.Weston N.Willard J.Wing TurboBLAST:A Parallel Implementation of BLAST Built on the TurboHub TurboGenomics Inc

[5] Barry Wilkinson, Michael Allen Parallel Programming Techniques and Applications Using Networked Workstations and Parallel Computers LuXinDa translate China Machine Press 2002.67~87

[6] Rajkumar Buyya High Performances Cluster Computing Architectures and Systems Volume1 ZhengWeiMing ShiWei WangDongSheng translate Publishing House of Electronics Industry P243~P257 P443~P444

Mao Liming is a master candidate of Computer Technology Institute in School of Computer Science, Wuhan University of Technology. He graduated from JanHan University and got bachelor degree in 2001. His research interests are in distributed parallel computation, etc.

Wang Zhongjun is an Associate Professor in the Department of Statistics, School of Science, Wuhan University of Technology. She graduated from Beijing normal University, in 1987, and from Wuhan University of Technology in 1997 and acquire M.S. degree. Now she is Ph.D. candidate of department of computer science, Huazhong University of Science and technology, and majors on theory of computer.

Guo Qingping is a Full Professor and a head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He is one of the DCABES international conference founder, was the chairman of DCABES 2001, co-chair of DCABES 2002, and the chairman of DCABES 2004.

Communication between WDPF and Remote Terminal Unit in Power Plat

YiKui, ZhuTianqing
Wuhan Polytechnic University, Wuhan 430023,China
E-mail: ykll1903@126.com **Tel:** (027)62010300

ABSTRACT

In order to realize the direct communication between a computer and Distributed Control System(DCS), the structure of the WDPF type DCS made by the Westinghouse Company of America is analyzed in detail in this paper. The compatibility and a high speed communication between DCS and the existing computer systems such as Remote Terminal Unit(RTU) and Management Information System(MIS) are problems when employ a DCS. In order to realize a direct communication between computers of RTU and DCS, the structure of the WDPF type DCS made by the Westinghouse Company of American is analyzed in detail in this paper.By means of analyzing its inner agreement on communication, an interface between the WDPF type DCS and the computer is designed, which includes the software and hardware. This interface has been put into use successfully and efficiently in the RTU project of the WISCO Power Plant.

Keywords: DCS, RTU, communication port

1. INTRODUCTION

Along with our country industrial and quick development in electric power with technical exaltation in microprocessor, gather to spread to control the system the extensive applying in our country fire power station automatically in the control, realizes" scatter about the control, centralized management" of target, develop to emphasize to want the function in electric power produce. The dominating DCS products include Westinghouse’s WDPF, BAILY’s N-90 and INFI-90, Honeywell’s TDC-3000, and HITACHI’s HIACS3000.

Generally, DCS communicate with some other computer systems such as RTU and MIS in real-time control. On account of the independency of DCS, the communication between DCS and other computer systems should be solved when enhancing functions of DCS, and design an effective communication port is a key assignment.

In this paper, we use QLC (supplied by Westinghouse) as a bridge to design a communication port between WDPF and computers of RTU. The port applies the MODBUS protocol to realize point-to-point communication and can shield computers from attacks of outside network to ensure the reliability of generator units. The rest of paper is organized as follows. In Section 2,we describe the Westinghouse’s DCS—WDPF. In Section 3, we introduce the MODBUS communication protocol. Port design is described in Section 4 including hardware layout and software disposal. Section 5 offers the testing analysis from the communication system.

2. WDPF

WDPF adopts distributed structure, which consist of all kinds of workstations with different functions (Fig.1). These workstations [1] work on a high-speed bus---WESNET II --- designed by Westinghouse and through which the whole system is connected by coaxial cable. WDPF enables each station on the bus to write or read process parameters of any other stations without uniform communication commander, thus the distributed public database is stochastically shared, which enables a long-distance workstation to implement real-time processing to the process data of other workstations. In addition, WDPF offers many convenient services with friendly interface, such as printing diversiform production report forms at HISTORY/LOG workstation, retrieving for historical data, printing alert messages and etc. And all these functions can be configured according to the given instances of any power plant.

The distributed process control system of WDPF is deployed to control generators and boilers in power plants. The system collects relative data of generators, drums, forced draught, induced draught, fuel and feed water system, and controls these systems automatically. The application software of DCS on operator stations enables the operating personnel to monitor real-time change of all kinds of parameters in production process and ensure the reliability of system control.

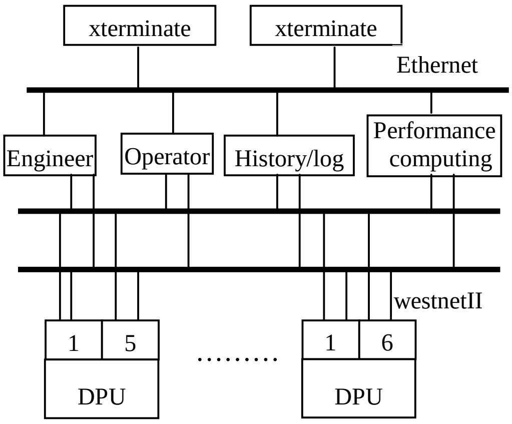


Fig.1 Distributed Structure of WDPF

3. MODBUS

3.1. Transactions on MODBUS Networks

The QLC card of WDPF applies the Modbus protocol to communicate with outside systems. On Modbus Networks, controllers communicate with a master-slave technique [2], in which only one device (the master) can initiate transactions (queries). The other devices (the slaves) respond by supplying the requested data to the master, or by taking the action requested in the query. Typical master devices include host processors and programming panels, and typical slaves include programmable controllers.

The Modbus protocol establishes the format for the master’s query by placing into it the device (or broadcast) address, a function code defining the requested action, any data to be sent, and an error-checking field. The slave's response message is also constructed using Modbus Protocol. It contains fields confirming the action taken, any data to be returned, and an error-checking field. If an error occurred in receipt of the message, or if the slave is unable to perform the requested action, the slave will construct an error message and send it as its response. Figure2 shows the Master-Slave Query-Response Cycle.

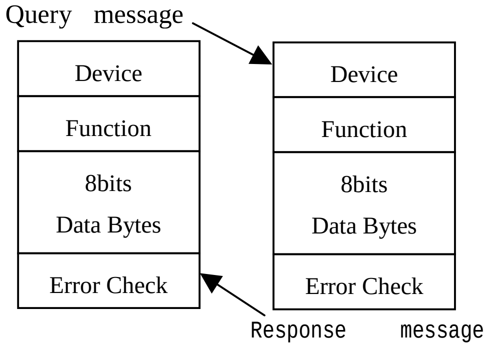


Fig.2 Master-Slave Cycle

The function code in the query tells the addressed slave device what kind of action to perform. The data bytes contain any additional information that the slave will need to perform the function. The error check field provides a method for the slave to validate the integrity of the message contents. If the slave makes a normal response, the function code in the response is an echo of the function code in the query. The data bytes contain the data collected by the slave. If an error occurs, the function code is modified to indicate that the response is an error response, and the data bytes contain a code that describes the error. The error check field allows the master to confirm that the message contents are valid.

3.2. RTU Mode

Protocol defines the communication formulae. Generally, MODBUS protocol is especially feasible to the communication type of networks consisted of PC in industry field. On standard Modbus Networks, controllers can be setup to communicate using either of the two transmission modes: ASCII or RTU. Actually, the bytes used to transmit messages in both modes of RTU and ASCII has the same definition except for the type of error check. But the characters used in ASCII mode are almost double more than which in RTU mode, and the slaves in WDPF communicate in RTU mode, thus we discuss the RTU mode emphatically.

3.2.1. RTU Framing

In RTU mode, Messages start with a silent interval of at least 3.5 character times and networked devices monitor the network bus continuously, including during the silent intervals. When the first field is received, each device decodes it to find out if it is the addressed device. If a silent interval of more than 1.5 character times occurs before completion of the frame, the receiving device flushes the incomplete message and assumes that the next byte will be the address field of a new message. RTU message frame is showed in Fig.3.

3.2.2. Address Field

The address field of a message frame contains eight bits in RTU mode. Valid slave device addresses are in the range of 0-247 decimal. The individual slave devices are assigned addresses in the range of 1-247. A master addresses a slave by placing the slave address in the address field of the message. When the slave sends its response, it places its own address in this address field of the response to let the master know which slave is responding.

START	ADDRES	FUNCTION
T1-T2-T3-T4	8 bits	8 bits
DATA	CRC CHECK	END
n*8 bits	16 bits	T1-T2-T3-T4

Fig.3 RTU Message Frame

3.2.3. Function Field

Function code tells slaves which response should be taken. All the function codes are showed in Fig.4 as well as their names and functions.

Code	Name	Function
01	read coil status	Reads the ON/OFF status of discrete outputs in the slave.
02	read input statuses	Reads the ON/OFF status of discrete inputs in the slave.
03	read holding	Reads the binary contents of registers holding registers in the slave.
04	read input registers	Read the binary contents of input registers in the slave.

Fig.4 MODBUS Function Code

3.2.4. Data Field

The data field of messages sent from a master to slave devices contains additional information, which the salve must use to take the action defined by the function code, and the data field of a response from a slave to a master contains the data requested. These messages may include items like discrete and register addresses, the quantity of items to be handled, and the count of actual data bytes in the field. For example, if the master requests a slave to read a group of holding registers, the data field specifies the starting register and how many registers are to be read.

4. Port Design

4.1. Hardware Layout

WDPF communicates with outside devices via the RS-422 port on the card QLC. To realize the communication of WDPF and RTU computers, we plant a PLC-745B card which offers a RS-422 port to the RTU computers.

PCL-745 series card provides two RS-422/485 serial ports. Each port utilizes a 16C550 UART with an on-chip 16-byte FIFO buffer for reliable, high-speed serial I/O. Main features of PLC-745B include:

Two independent RS-422/RS-485 serial ports
Supports a long-distance communi- cation of 1.5KM
Transmission speeds up to 921.6Kbps
Supports standard DOS COM1, COM2,COM3,and COM4
Supports TX, RX, RTS ,and CTS signals
Supports 2 wire or 4 wire operation

The pin assignments for the card's connectors in RS-422 are showed in Fig.5.

pin	signal	pin	signal
1	TX-（send data-）	6	CTS-（clear -）
2	TX+（send data+）	7	CTS+（clear +）
3	RX+(receive data+)	8	RTS+（require +）
4	RX-（receive data-）	9	RTS-（require -）
5	Ground		

Fig.5 Pin Assignments of PLC-745B

The RS-422 port of WDPF is offered by QLC, and the Fig.6 shows the pin assignments of QLC's connectors.

pin	signal	pin	signal
1	RXD-（receive data-）	6	RTS-（require-）
2	RXD+（receive data+）	7	RTS+（require+）
3	TXD+（send data+）	8	CTS+（clear+）
4	TXD-（send data -）	9	CTS-（clear-）
5	Ground		

Fig.6 Pin Assignments of QLC

As the pins 5,6,7,8 and 9 of QLC have not used by the MODBUS applications (as mentioned in 2.2.3), the communication port consists of 4 pairs of corresponding pins in QLC and PLC-745B(showed in Fig.7).

QLC		PLC-745B	
pin	signal	pin	signal
1	RXD－	1	TX－
2	RXD＋	2	TX＋
3	TXD＋	3	RX＋
4	TXD－	4	RX－

Fig.7 Pin Assignments of Port

4.2. Software Disposal

4.2.1. Modification of System Clock

At speak of the MODBUS agreement, a time for and a wanting staying 3.5 words sign partition that data flow, horary accuracy in system in machine in general PC is 1/18.2 a time for with rate is 9600, delivering each piece according to is 1/9600 second, and it takes 2.8ms to transmit 3.5 characters, thus the computers should have the ability to identify the interval of 2.8ms to realize accurate communication. But the system clock's precision of common computer is 1/18.2s (viz. 55ms), therefore the precision should be modified to 0.1ms.

To resolve this problem, we should modify the initial value of the Intel 8253 in system and choose the operation mode 3[3]. In this mode, the Intel 8253 works as a square-wave generator and the output will remain high until one half of the count has been completed and then low for the other half.

The 8253 chip has 3 channels (CNT0, CNT1 and CNT2), each of which is responsible for a different task on the PC. Channel 0 is responsible for updating the system clock. Because the initial value of CNT0 is set to 0000H, it is usually programmed to generate around 18.2 clock ticks a second and an interrupt 8 is generated for every clock tick. When we modify the initial value of CNT0 to 0080H, the precision of system clock will be improved by 512 times and system then could recognize the interval of 0.1ms.

4.2.2. Resident Program

In practice, the computers of RTU perform dispatcher program (YJC 1.00) in real time with DOS operation system, which occupies the computers completely. To perform the communication software and YJC respectively without disarrangement of each other, we apply the multitask program and store the communication software internally in inner memory (as a TSR procedure) of RTU computers.

Designing TSR procedure includes 3 segments[4]:

Initialization: put the entry address of TSR procedure into interrupt vector table;
Function segment: interrupt service routine to perform specific function;
Put the TSR procedure into inner memory with given length.

The segment program below shows the design of TSR procedure for RTU computers.

```
#include <dos.h>
static struct REGS rg;
unsigned int sizeofprogram;
rg.x.ax=0X3100;
rg.x.dx=sizeofprogram;
intdos(&rg, &rg)
```

As mentioned in Section 3.2, system clock has been modified to perform correct communication of WDPF and RTU computers, but all other performances including DOS operation system should use original system clock all the same. So, two functions, new_timer() and old_timer() are defined to harmonize the inconsistency. The function new_timer is generated by system function setvect() as an interrupt service routine[5] to keep the system clock's precision with 1/(18.2*512)ms, and the function old_timer will be called every 1/18.2ms to maintain the correct system clock precision with 1/18.2ms for all other intrinsic performances in computers of RTU.

5. SYSTEM TEST

In field control practice of DCS of power plant, there are over 50 analog variables and 66 digital variables should be sent in real time. The communication baud rate of the system is 9600bit/s, and 16 analog variables and 16 digital variables could be sent in one frame by QCL, thus 4 analog variable frames and 4 digital variable frames would be need to send total data.

According to the MODBUS protocol, to send a digital variable frame, 17 bytes(including a query frame consisted of 8 bytes and a response frame consisted of 9 bytes) will be sent between master and slave. Meanwhile, 36 bytes(including one ADDRESS byte, one FUNCTION byte, 32 DATA bytes and 2 CRC CHECK bytes) will be sent to transmit an analog variable frame. Each byte consists of 11 bits(including start bit, parity check bit and stop bit) and 0.1ms is taken to transmit each bit, thus the total time to transmit the whole data is :

(4*36+5*17)*11*0.1=251.9ms.

That will ensure the satisfaction of wants in real-time field control.

6. CONCLUSION

Before placing the communication port in service, we have performed experiments with several types data on the system and relative software. The test result shows that the communication system can perform transmission effectively with a perfect ability of anti-interference. Meanwhile, the other programs in RTU computers perform normally without effect of the communication. In practical control of power plant, the communication system plays a key role to realize the automatic process control of ash removal, water carburetion, coal handling and boiler analysis system.

7. REFERENCES

[1]. WDPFII WESTATION OPERATOR. Westinghouse Process Control, Inc.2001

[2]. MODBUS Protocol. http://www.modicon.com/techppubs/intr7.html

[3]. ZhouMingde. Principles and applications of Microcomputer system. TsingHua University Press. 2000

[4]. Limushun. TURBO C internally stored program and windows software program. Beijing KeHai Training Center.

[5]. Joe Campbell. C Programmer's Guide to Serial Communication(2nd Edition). TsingHua University Press. 1993

The Health Monitoring and Damage Identification Platform of the Civil Infrastructure Based on Internet

Xiao Chun[1] Qu Weilian[2] Zou Chengming[3]
[1]School of Automation, Wuhan University of Technology,Wuhan 430070, P.R. China
Email: xiaochun70@163.com **or** xiaochun@mail.whut.edu.cn **Tel.:** 8627-62680562
[2]School of Engineering and architecture, Wuhan University of Technology, Wuhan 430070, P.R. China
Email: qwlian@public.wh.hb.cn **Tel.:** 8627-62680906
[3]School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, P.R. China
Email: zoucm@mail.whut.edu.cn **Tel.:** 8627-62680511

ABSTRACT

The development of Internet technologies has facilitated real-time monitoring in more and more fields. This paper researches the Internet-based platform of remote health monitoring (RHM) and damage identification (DI) for civil infrastructure, and introduces its main task, researches platform framework, network architecture, software platform and hardware frame, and implements the network monitoring system. Data acquisition is the first step of a remote monitoring system. In this paper, data is acquired by the distributed sensor network, processed and displayed by real-time monitoring system, which integrate on-line real-time heterogeneous sensor data, database of complex structural system, data analysis and rational decision-making.

Keywords: health monitoring, civil infrastructure, damage identification, distributed sensor network, remote transmission, real-time monitoring

1. INTRODUCTION

All man-made structures and machines have finite lifespan and begin to degrade as soon as they are put into service. Processes such as corrosion, fatigue, erosion, wear and overloads degrade them until they are no longer fit for their intended use. It would be very dangerous to people’s life and possession if we could not estimate the magnitude of the damage of the civil infrastructure timely. Everyday we can hear of by news all kinds of accidents due to the deterioration and damage of civil structures. In recent times, infrastructure health monitoring has made great strides in North America, Europe and Japan. Immense progress has been made in this regard[1-3]. However, the application is still not enough. Meanwhile, the development of Internet technologies has facilitated real-time monitoring in more and more fields, and the structure health monitoring is not exceptional. The platform is a test platform about the application of the health monitoring and damage identification of the civil infrastructure based on Internet. The research advances the application of remote health monitoring in the field of bridge and civil infrastructure.

2. MAIN TASK

The aim of the civil infrastructure health monitoring is following the tracks of micro changes in a structure. These micro changes are not so obvious unless they reach the stage of distress cracks or displacements. At the proper locations are installed heterogeneous sensors to record on-line real-time micro changes. These sensors give us valuable information about the health condition of an infrastructure. If this is monitored regularly, helpful information is obtained. When the helpful information is compared against a reference model of the undamaged structure, structural changes can be detected, serving as damage detection. The major tasks in structural damage identification are identifying the existence of damage, and identifying the location of the damage, and estimating the magnitude of the damage. Once the damage is completely identified, it can be decided whether to remedy the structure, replace it or throw it away. The research of this platform will focus on the framework of remote health monitoring and damage identification platform, on how to get a helpful data from a wide variety of sources by distributed sensor network, and on implementing the remote transmission.

3. RESEARCH FRAMEWORK

This platform mainly integrates three subsystems. They are network supervision subsystem, signal detection subsystem and damage diagnosis subsystem. Fig. 1 is built to achieve the platform frame and Fig. 2 to show the function flow chart of the platform. Network supervision subsystem implements data (including environmental data and output result by damage identification algorithm) transmission in intranet and Internet, system user administration and network security etc. Signal detection subsystem consists of distributed sensor network, sensors data processing and outputting helpful data for the diagnosis center. Damage diagnosis subsystem makes structural undamaged modal, fusions data from signal detection subsystem, and implements diagnosis..

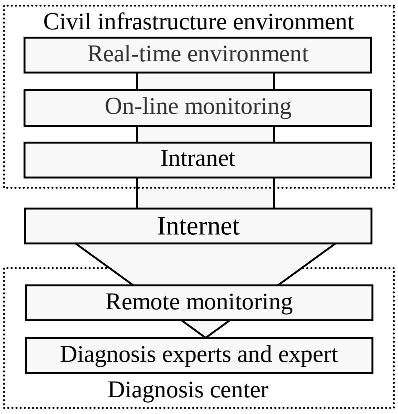


Fig. 1 The frame of the platform

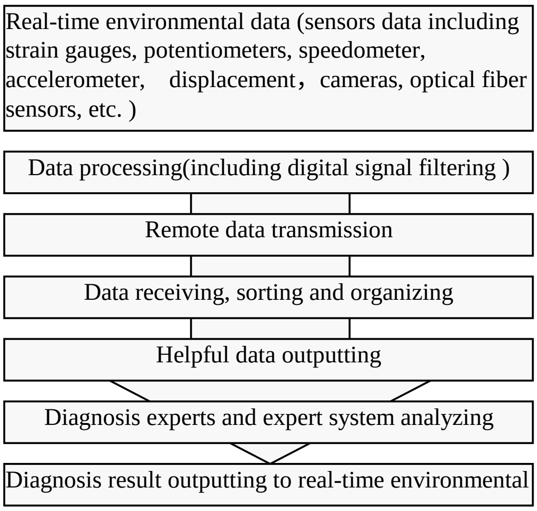

Real-time environmental data (sensors data including strain gauges, potentiometers, speedometer, accelerometer, displacement，cameras, optical fiber sensors, etc.)
Data processing(including digital signal filtering)
Remote data transmission
Data receiving, sorting and organizing
Helpful data outputting
Diagnosis experts and expert system analyzing
Diagnosis result outputting to real-time environmental

Fig.2 The function flow chart

4. CONSTRUCTION AND IMPLEMENTATION

Construction

Figure 3 is built to achieve construction of platform in detail.

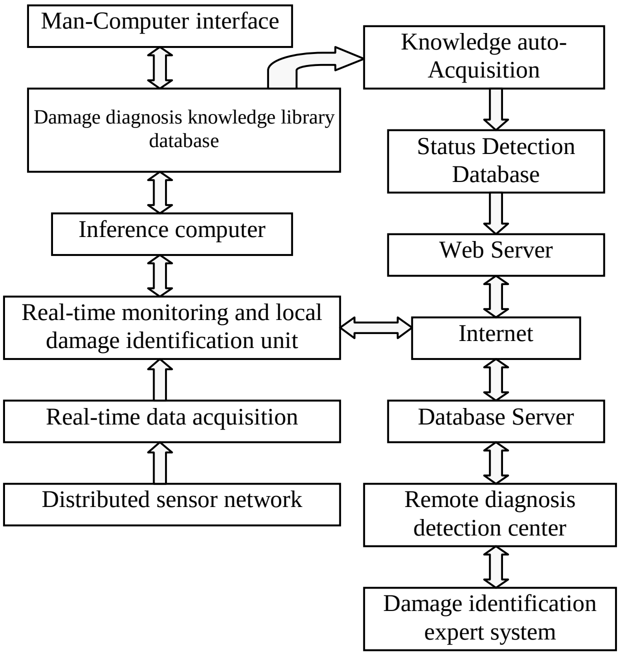

Man-Computer interface
Knowledge auto-Acquisition
Damage diagnosis knowledge library database
Status Detection Database
Inference computer
Web Server
Real-time monitoring and local damage identification unit
Internet
Real-time data acquisition
Database Server
Distributed sensor network
Remote diagnosis detection center
Damage identification expert system

Fig. 3 The detailed construction of the platform

Distributed Sensor Network

The data acquisition of the structural health monitoring and damage identification is from the sensor network that is the first step to design a monitoring system. The sensor network consist of selecting the types of sensors to be used including required bandwidth and resolution, the locations where the sensors should be placed, the number of sensors to be used, the relation to sensors how to be made. In this paper, a distributed sensor network is shown in figure 4.

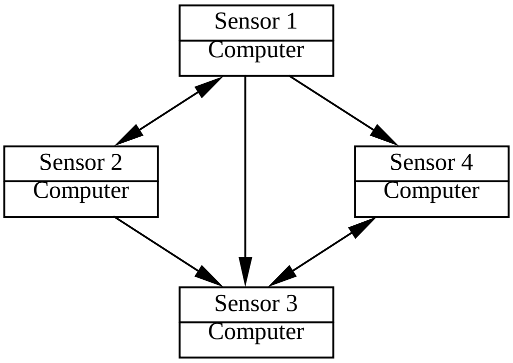


Fig. 4 Distributed sensor network

The sensor network allows some different type sensors (e. g. strain gauges, potentiometers, speedometer, accelerometer, displacement, cameras, optical fiber sensors, thermometer, etc.), and each having local processing and low-bandwidth communication capabilities with other sensor. Therefore, each sensor unit has the ability to process data and to cut off the false data because of interfering before the data is transmitted to real-time monitoring unit. From the forgoing analysis， we know this sensor construction can improve system performance. The system can compensate some module through others when it makes fault [4].

Network Architecture

There are three usual remote communication technology which are wireless technology, GPS(Global Positioning System) technology, and Internet-based technology. In this paper, Real-time monitoring and local damage identification unit communicates with remote diagnosis center by Intranet and Internet. Figure 5 is built to achieve the network architecture of the platform which is a three-level architecture based on WWW (World Wide Web) mode.

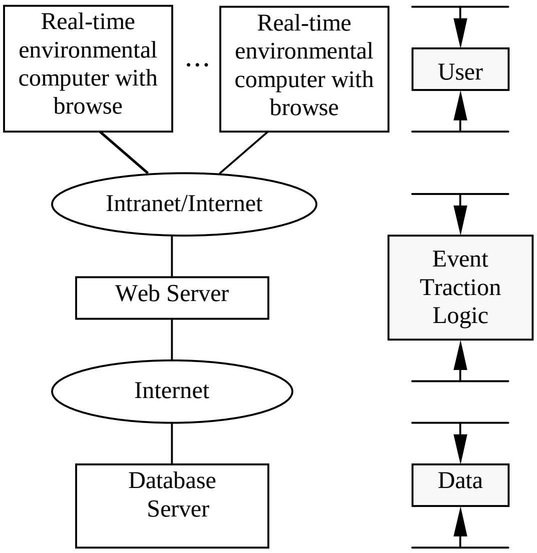

Real-time environmental computer with browse
…
Real-time environmental computer with browse
User
Intranet/Internet
Web Server
Event Traction Logic
Internet
Database Server
Data

Fig. 5 The three-level architecture based on WWW (World Wide Web) mode

Software Platform

The software that is applied in this platform includes Linux (a free operating system), J2sdk1.4.1(Java 2 Development Kit

JDK, a free JavaSoft by Sun company, you can get it from http://java.sun.com/j2se/1.4.1/download.html), Mysql4.0.13(a free database management system software which can be downloaded from http://www.mysql.com), Tomcat4.1.24(a free server developing software which can be download from http://jakarta.apache.org/, mm.mysql-2.0.4-bin.jar(a free port software between Mysql and Java to be download from http://mysql.ihostunit.com/Downloads/Contrib/mm.mysql-2.0.4-bin.jar), and Matlab 6.5(a simulation software which is not free). From the foregoing analysis, the cost of system software is very low. For system security based on networked application, firewall is used and every user access the Web should pass the authentication.

5. DAMAGE IDENTICATION STRATEGY

A two-level diagnosis mode (including local diagnosis and remote diagnosis) is applied in this damage identification. Real-time monitoring and real-time diagnosis is processed in the local diagnosis. Some problems unable to be identified in the local expert system will be sent to the remote diagnosis center by email. Remote expert system does damage identification as soon as it receives message. The strategy is showed in Fig. 6.

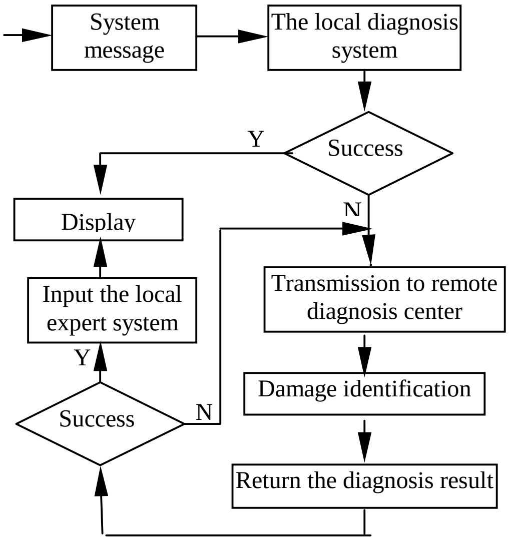

System message
The local diagnosis system
Success
Y
N
Display
Input the local expert system
Transmission to remote diagnosis center
Y
Damage identification
Success
N
Return the diagnosis result

Fig. 6 Remote diagnosis flow

6. CONCLUSIONS

An integrated remote monitoring platform for civil infrastructure is researched, which will promote the development of structural health monitoring methodologies. The paper applies a high-performance data acquisition method and some free software. It is worth of time and effort invested here.

7. REFERENCES

[1]. Stalling, J. M., et al. 2000. Field performance of FRP bridge repair. ASCE, Journal of Bridge Engineering, 5/2, 107-113.

[2]. Aktan, E., Chase, S., Inman, D., and Pines, D. 2001. Monitoring and Managing the Health of Infrastructure Systems. Proceedings of the 2001 SPIE Conference on Health Monitoring of Highway Transportation Infrastructure, SPIE, March 6-8, 2001.

[3]. Catbas, N., Ciloglu, K., Celebioglu, A., Popovics, J., and Aktan. E. 2001. Fleet Health Monitoring of Large Populations: Aged Concrete T-Beam Bridges in Pennsylvania. 6th Annual Int'l Symposium on NDE for Health Monitoring and Diagnostics, SPIE, Newport Beach, CA USA, March 4-8, 2001.

[4]. J. M. Manyika and H. F. Durrant-whyte. Data Fusion and Sensor Management: A Decentrialized Information-Theortic-Approach. Eills Horwood, NY, 1994:112-117

Xiao Chun is a lector of Control Theory and Control Engineering Institute in the College of Automation, Wuhan University of Technology, and is currently a PhD student in the College of Civil Engineering and Architecture at Wuhan University of Technology too. She received her MengSc from Wuhan University of Technology in 2000. Her research interest and practical experience are in the areas of control theory, computer control and structural health monitoring.

The Research on Intelligentized Distributed Cooperative Virtual Environment

Gao Shu[1] **Cheng Ding-Fang**[2]
[1]**Computer College ,Wuhan University of Technology , Hubei 430063,China**
[2]**Department of Logistics Engineering , Wuhan University of Technology , Hubei 430063,China**
Email:gshu418@163.com **Tel.**: 027-62707682

ABSTRACT

The paper firstly introduces the developing state of the distributed cooperative virtual environment and the problems in the research about it. Secondly, it analyses the necessity and feasibility of integrating the Agent technology into the DCVE system .In the end, the paper studies the key aspects about the Intelligentized Distributed Cooperative Virtual Environment ,including the system's structure , the structure of the cooperative relation-oriented Agent, the communication frame of the Agents, cooperation and negotiation mechanism, the cooperative awareness mechanism and etc.

Key words: the distributed cooperative virtual environment; the distributed virtual reality ; CSCW; Agent

1. INTRODUCTION

As a new kind of research branch, the distributed cooperative virtual environment (DCVE for short) integrates the technology of the distributed virtual reality (DVR for short) and the technology of CSCW. It provides the people ,who are separated in the space time, and depend on one another in the work, with an environment of natural communication and cooperation . Therefore ,it pays more attention than the DVR to the perceptibility and the cooperation among users in the shared virtual space and to the coordination control among cooperators in order to support the cooperation and enhance efficiency .On the other hand, as compared with the CSCW, as technology of DVR is introduced, DCVE can create for the cooperators a kind of 3-dimension video and audio virtual environment ,which is really united, harmonious ,open and cooperative ,because the user feels as if he did interact with people ,not machine ,when he communicates with others in a working group.

Since the 1990s, because the DCVE system can efficiently solve the cooperation among the users in the distributed environment ,the research on DCVE has been emphasized highly in the developed countries, and some prototypes using the idea have been developed, representatives of which are as follows:

- Simulation Based Design (SBD) and Intelligent Collaboration and Visualization (IC&V) of DARPA in America.
- Partnership for Advanced Computational Infrastructure (PACI) and Knowledge and Distributed Intelligence (KDI) of NSF in America.
- Intelligent Synthesis Environment(ISE) of NASA in America.

In spite of the emphasis on the research on the DCVE and appearance of prototypes and application system , a lot of difficulties in study of DCVE still remain. They have impeded the DCVE system to enter the commercial and applied stage. The problems can be classified into the several aspects:

- Low supporting technology

The establishment of DCVE is related to lots of technologies, including DVR, CSCW, Network communication and its security, Integration of System, Artificial Intelligence and etc. It is the first problem to be solved that how to integrate these technologies together in order that every part in the DCVE system can perceive, adapt and adjust itself to the change of outer environment dynamically and in real-time.

- Architecture

Architecture is defined as orderly arrangement and structure of parts in a system, and a rule and method which control the design and construction of a system. Due to its complexity, the architecture of DCVE play an important role in the deciding the its success or not. Nowadays, generally, the kind of system adopts the layered idea, i.e. some layers support the implement of DVE, and the others CSCW. The disadvantage of the idea is that the cooperative activity in the system ,in which the machine and users are all "inhabited", are not regarded as a whole .So it causes a serial of problems, such as the scalability , interoperability, extensibility and reusability of the system.

Apart from these, there are problems including the research on the cooperative mechanism, how to effectively support the perceptibility, virtual cooperative space and etc. Therefore, it is necessary to use the new method as low supporting technology to integrate the DVR with CSCW in order to build up the large-scale, distributed, dynamic and cooperative network environment.

2. THE RESEARCH ON THE INTELLIGENTIZED DCVE

2.1 The application of Agent theory

The newly arisen theory and technology Agent provides a new idea and method for solving above difficulties:

- It is possible to divide a large-scale and complex problem into a serial of smaller and simpler sub-problem by means of the Agent technology.
- The entities and their relations in the real world can be mapped into the agents in the virtual reality, which hold relevant resources and have the solving problem ,interactive and cooperative ability .
- The relating agents can form a collaborative group to complete a complex task. They can regulate themselves, negotiate with each other about conflict and collaborate in order to achieve their common goal.
- The Agent-based developing method can describe the activity in a complicated and concurrent system more accurately than the traditional one.

Consequently, as far as DCVE system is concerned ,in which the environment change constantly and dynamically ,the relating information can not be obtained completely and the total goal must be divided into a lot of collaborative sub-goal achieved by many users together ,it is suitable to use multi-agent system to establish. The reasons are as follows:

Idea of the Agent theory is that the software based on Agent can imitate the social activity and recognition of human, which include the organizing mode, the cooperative relation, evolution mechanism, and pattern of recognizing, thinking and solving problem of people in the society. Therefore the people themselves can regard as Agents. If so, the users and computers can be united in the multi-agent system, the communication among users and interaction between the user and computer can be implemented by the interaction among the Agents: driven by the goal ,they are able to adopt a variety of behaviors actively ,including social and studying activities and etc, in order to percept ,adapt and adjust themselves to change in the dynamic environment; their intelligence and collaborative abilities conform to the opening of the DCVE system and enable them to achieve the system's total goal by means of loose federal mechanism ; their initiatives and sociality provide the natural human-human interaction and human-machine interaction for the DCVE; their mobility ability enable it possible to accomplish efficient location of resource and service matching; their ability of reactivity and study also enable them to be able to adjust their activity in time in order to adapt to the environment, which assure the synchronization and consistency of information in the DCVE system; their high autonomy ability make the members in the Agent-based environment go in or out so freely that the structure of system can be regulated dynamically, which assure the system has more flexibility , expansibility and robustness; at the same time ,the Agent technology is specialty-oriented ,and can have rich knowledge in the specialty and intelligence ,so it is quite suitable for dealing with the tasks in the DCVE ,which are complex and fuzzy.

In brief, it is necessary and feasible to develop the DCVE using the Agent technology. By providing the low supporting technology –Agent-for the DCVE ,It can integrate DVR with CSCW to build up a virtual reality having the characters of the group, interaction ,distribution and cooperation.

2.2 The structure of the agent-based DCVE

Through analyzing the function of the DCVE, we will describe the DCVE using a four-tuple as follows:

DCVE=（Agents, Tasks, KB ,CVE）。

where Agents is a set of Agents, in which every Agent has the certain ability of solving problem. Tasks is a set of task, which explains the tasks to accomplish. KB is a knowledge base, which describes the a variety of the knowledge used by the Agent. CVS is a collaborative virtual space, which offers the working environment for the Agent.

2.3 The research on the structure of the cooperative relation-oriented Agent

The study on the structure of Agent is the base of the research on the multi-agent system. Recently, various structures of the Agent, such as the structure of spirit- state-based Agent, the structure of

Knowledge-based Agent, the structure of goal-oriented Agent and the structure of time-centered Agent ,are proposed .According to the character of DCVE, we put forward to a structure cooperative-relation-oriented Agent .See figure 1.

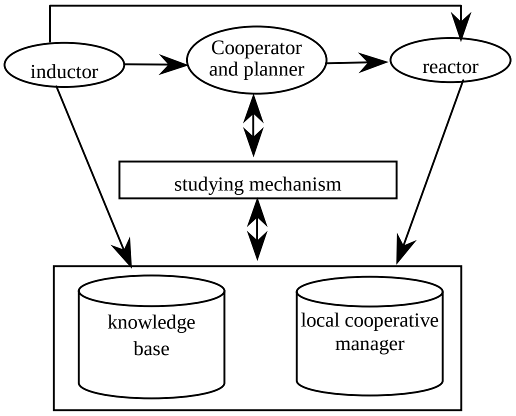


Figure 1. The structure of the cooperative relation-oriented Agent

Where the inductor is responsible for receiving the input information and forming outputs at knowledge level using relating information in the KB. In the meantime ,it classifies the outputs into two, some of which are sent to reactor, and the others the cooperator and planner. It consists of receiver of information and processing mechanism.

Cooperator and planner is the center of Agent .It makes full use of a variety of information to divide, plan, infer the complex activities of Agent in order to form corresponding result, which will be sent to reactor. On the other hand, it can use and modify data about present running state in the local cooperative manager, and coordinate the activities inside the Agent.

As the part of interacting with outside, reactor transforms information, which is from the inductor or cooperator and planner, to corresponding action to accomplish the relevant task.

Studying mechanism is a on-line study system through the collaborative work and can update the knowledge base using the new information.

Knowledge base stores the a variety of information about control, specific field and relating data. The fuzzy theory is used in the knowledge representation and exchange of Agent so that it can adapt the dynamic and uncertain characters of Agent itself or relation between Agents.

Local cooperative manager records the Agent's data related to the present working state, the present running state, the present cooperative state of the environment.

The structure of Agent emphasizes that some collaborate functions should be implemented inside the Agent. Meanwhile, it is also a hybrid Agent. On the one hand, its two parts —inductor, reactor—compose a reactive Agent, which directly transforms the information from inductor into action so as to imitate the reactive action of human efficiently. On the other hand, the inductor, cooperator and planner, reactor together compose the deliberative Agent, which is used to plan, infer the complicated activities. So it has higher priority than the reactive Agent .Therefore ,the hybrid Agent whose structure has hierarchy can implement the long plan ,and has good reactive capability as well.

2.4 The communication frame of the Agents

The communication ability of Agent is the representation of its sociality ,the base of its intercourse, corporation ,competition and etc. It is very important in the distributed environment. In order to enable the Agent technology to be effectively used in the WWW ,it is necessary to solve two problems:(1)how to decide the standard message format which is structural and semantic ;(2)how to decide the mechanism used by Agent to explain the structural message and exchange the knowledge about specialty .A XML-based communication frame of the Agents is designed in our DCVE system. The Agent still uses the ARPA KQML as communication language, uses the ARPA KIF as knowledge representation language .At the sending end, the data of the KQML message is encapsulated into XML text in the DTD format, and at receiving end ,the XML text is parsed into KQML message in order to obtain the data. So the XML can use the standard WWW technology to solve such problems as interoperability ,and the small difference of operational semantics in the implementation and usage between different Agent communication languages ,which will benefit for the standardization of operational semantics ,make it possible interoperability in the heterogeneous system .Therefore ,it also is good for the coordination and corporation among Agents ,and enhance the communication efficiency among the different kinds of Agents. At the same time, the communication methods of the different modes and different levels are adopted in our DCVE system to support the synchronous and asynchronous message pass.

2.5 The research on the cooperation and negotiation mechanism in the DCVE

The cooperation work in the DCVE is a complicated and dynamic process which changes constantly in the space and time and have the structural and semi-structural characters. So the Agent technology should be introduced into the cooperation mechanism in our DCVE system ,i.e. the entities in the cooperation work are regarded as Agents ,and it is regarded that the process of cooperation consists of a serial of Agents' activity. Therefore, the corporation activity among the users can be described by the Agents' action , behavior and their relation.

The negotiation is the key to realizing the cooperation in the multi-agent system, which is regard as a promising solution to collision. Generally ,the policy about negotiation can be classified into five kinds: unilateral concession policy, competition policy, negotiation policy, destructive policy and delay policy. A self-adapting scheduling algorithm is designed to enable the Agents to dynamically and intelligently select the suitable policy so that they can cooperate and compete with one another. At the same time ,the contract-net protocol is still used as the basic method to assign the tasks ,but the spirit state of Agent (including the degree of trust, stability ,and enthusiasm) is introduced into the bidding process of the protocol so as to improve the quality of corporation.

2.6 The research on the cooperative awareness mechanism

In order to support the human-human interaction, the DCVE system make users aware of not only the reaction of machine, but also the existence ,action ,state and feedback information of the other users. The cooperation work is based on the cooperation awareness. We integrate the intelligence of Agent with the visualization of the Virtual Reality so as to offer more ways to obtain information for the users. So the cooperation awareness in such Agent-based system is greatly improved.

2.7 The research on the real-time character of DCVE

The verisimilitude of the DCVE depends on the real-time reaction of output ,such as image and audio ,to the user's input. In order to meet the real-time interaction, the delay time from the user's input to the output of image and audio must be reduce to the minimum. Generally speaking ,the factors affecting the real-time in the DCVE are mainly from two aspects: the network delay and computation delay. On the limited bandwidth network, for example, Internet, the problem becomes quite obvious. The DR algorithm ,AOI technology ,information filter and compression are usually adopted to reduce the amount of communication, but the methods ,including the grading management about interest, the DR algorithm with multi-threshold ,data package compression with the entity state ,are used in our system. Meanwhile, in order to reduce the amount of the computation, the levels of detail algorithm is used to simplify and compress the geometry model.

2.8 The research on the consistency in the DCVE

Some effective measure must be taken to maintain the consistency of time, space, state of entities and the shared information. As for the consistency of time and space, the common method is used in our system. We emphasize the study on the later two. As to the consistency of state of entities, we will study the methods, including pre-obtaining and pre-estimation for state of information, synchronization with the system clock. As for the consistency of shared information, it is achieved by setting the token and the cooperative awareness Agent, i.e. only if he holds the token can the user access the information .And cooperative awareness Agent monitors the information momently . Once the change happens, it picks up the changed result and distribute to the users on the network.

3. CONCLUSION

A prototype based on the above research is being developed. The practice shows that it is necessary and feasible to introduce the Agent into the DCVE, which make it possible that the users separated by time and space make full use of the rich shared resource to do the variety of cooperation work.

4. REFERENCES

[1] Mills K L. Introduction to the Electronic Symposium on Computer-Supported Cooperative Work. ACM Computing Surveys,1999-06

[2] Zheng Qinghua, Li Renhou. Study on the Model of Multi-level and Multi-group Cooperative Work Based on Agent. Proceedings of Fourth International Workshop on CSCW in Design, Compiegne France,1999-09，

[3] Ju Chunhua, Wang Guangming. Research of decision support system on commerce [J]. Proceedings of 9 Intl Conf on Management Science & Engineering [C], 1999. 156-161

[4] R J Rabelo, L M Camarinha-M atos, H A fsarmanesh. Multi-agent-based agile scheduling, Robotics and Autonomous Systems,1999,27:15～28

[5] HUANG C Y, SHIMON Y N. Formation of autonomous agent networks for manufacturing system[J]. Int. J. Prod. Res.,2000, 38(3): 607-624.

Study on Application of Data Fusion in Erosion Detection of Furnace Lining

Liu Quan, Zhang Xiaomei
School of Information Engineering, Wuhan University of Technology, Wuhan 430070
Email: qliu@public.wh.hb.cn

ABSTRACT

In this paper, three different methods which are capacitance method, flame image processing and laser thickness measure are used to detect the erosion condition of steel-making converter's lining, and then their result are processed by using the data fusion technology based on BP neural network. The simulation results show that this kind of detection method can get relatively ideal detection results.

Keywords: BP neural network; data fusion; steel-making converter; erosion condition

It is very important to detect the erosion condition of steel-making converter's lining and to adopt corresponding measures in time for steel-making furnace to maintain its high yield and longevity. The thermocouple method, heat flux meter method and infrared ray imaging method are adopted in early years. Recently/In recent years, thermocouple trigger pulse response method was invented overseas. With the development of laser technologies, most countries adopt the laser thickness measure to detect the erosion condition of steel-making converter's lining. However, the information acquired by one sensor is incomplete usually. While multi-sensor data fusion technology acquires information from multi-source signal, it greatly decreases the information uncertainty, provides better illustration of signal source and improves the precision of the detection. More in-depth research has been done into multi-sensor data fusion technology, and the primary theory involves data base theory, reasoning theory, blackboard structure, artificial neural network, Bayes rule, D-S evidence reasoning theory, fuzzy set theory, Statistic theory, clustering technology, Entropy theory, and so on. Each of these methods and theories has its own advantages and disadvantages, and they are applied to different fusion layer. Generally speaking, in order to meet with special application background requirement we adopt fusion tools and methods, which aim at different application backgrounds. In this paper, we put forward a multi-sensor data fusion model based on artificial neural network and study on using this model to detect the erosion condition of steel-making converter's lining.

1. DATA FUSION TECHNOLOGY

Data fusion can detect, associate, correlate, estimate and synthesize the data from multi-source to acquire more accurate and more credible conclusion. What makes the difference between data fusion and conventional information processing technology is data fusion technology can process more complex information and fuse the data at different layer. We can divide the data fusion system into 3 layers according to the information layer processed by data fusion system:

1) Data Fusion Layer In this layer, original data acquired from sensors is correlated directly, sent to fusion center. Then the object to be measured is estimated synthetically.(Then the synthetical estimation of the object needing measured is completed.) Data fusion layer fuses the data based on sensors. Its advantage is maintaining as much original information as possible, while its disadvantage is needing to process too much information, low processing speed and having poor real-time performance. So it usually is used to process images with high data match accuracy.
2) Character Layer Fusion In this layer, characters abstracted from original data will be correlated, normalized and sent to fusion center. Then the object to be measured is estimated synthetically.(Then the synthetical estimation of the object needing measured is completed.) Character layer fuses data in mid-layer. Its advantage is maintaining as much original information as possible and compressing the data to extent as well. Popular as it is, character layer fusion's stability, compatibility and reliability in complex environment are still needed improving.
3) Decision Layer Fusion In this layer, before signals acquired from the sensors are fused, they undergo local processing such as characters abstracting, making decision. After that, the processed signals are correlated and sent to fusion center where these signals are processed ulteriorly. As an advanced data fusion, decision layer fusion excels in real-time performance, small data transmission, asynchronous information processing ability, system compatibility and reliability. Hence, decision data fusion is a focus of research on data fusion.

In practice, which fusion method is selected depends on concrete problems. In the detection of erosion condition of steel-making converter's lining, due to the different characters belong to each detection method, the bug information model varies. As a result it is difficult to correlate them directly and to fuse the data at data layer. Moreover, bug characters detected by different methods aren't identical. And that hobbles data fusion at character layer. Therefore, decision layer fusion is relatively feasible in the detection of erosion condition of steel-making converter's lining.

2. STUDEY AND SIMULATION ON DETECTION OF EROSION CONDITION OF STEEL-MAKING CONVERTER'S LINING

2.1 Some Methods in Common Use to Detect Erosion Condition of Steel-Making Converter's Lining

2.1.1 Capacitance Method

Make the inner surfaces of capacitance sensor's metal plate cling to converter's lining surface. To spray insulating material onto the outer surface of capacitance sensor's mental plate to ensure that plate is insulating. If the converter's lining is eroded, the dielectric capacitance of the corresponding area will changes which leads to changes of capacitance sensor array's output. As for the capacitance sensor array consist of N plates, there are N(N-1)/2 pairs plates of which N pairs plates border upon. In these N(N-1)/2 pairs plates, those pairs border upon can reflect the erosion condition of lining most effectively. Because limited elements analysis shows that the sensitive area of plate pairs, which border upon centralizes in lining other than hearth[3].

Compared with conventional inserted measurement like thermocouple method, thermocouple trigger pulse response method, capacitance method has high accuracy and is more endurable. Because inserted measurement needs to embed sensor into lining, the sensor and lining both are eroded as converter produces.

2.1.2 Laser Thickness Measurement

Laser thickness measurement is non-contacting measurement technology. Frequency modulator modulate laser to make the laser transmit modulated light with stable frequency. Then the modulated light radiates to a certain spot on lining surface and received by measure device after being reflected. The received signal is processed by system processing unit which gets the distance needing measured by figuring out the span of time when signal light making trips between measured spot and the center of measure device. Sent these distance data to host computer and compare with the distance predefined. Then the erosion condition of the steel-making converter's lining is acquired.

Because laser has good monochromaticity and isn't easily affected by environment, laser thickness measurement finder always has high precision under various outside conditions. Therefore, laser thickness measurement has the highest precision of the single sensor methods to detect erosion condition of steel-making converter's lining at present.

2.1.3 Flame Image Processing Method

When steel-making converter's lining is eroded to some extent, the color, shape and brightness of the flame in the hearth will change accordingly. In early stage/years, it mainly depends on experienced workers to judge the erosion condition qualitatively by watching changes of flame according to their experiences. With the development of image processing technology, the method which detects erosion condition by flame image becomes more quantitative. But this method is still need probing into with a little production and low measure precision.

2.2 Multi-Sensor Fusion Detection Method

2.2.1 The Design of Multi-Sensor Fusion Detection System

The detection system erosion condition of steel-making converter's lining based on multi-sensor data fusion is shown as fig1.

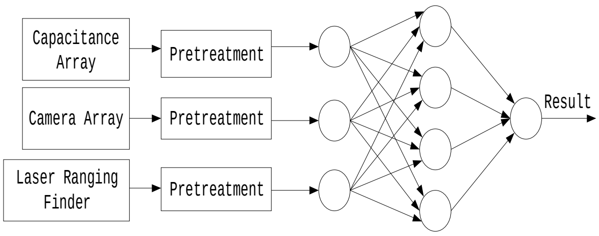


Figure 1 Detection System of Erosion Condition of Steel-Making Converter's Lining based on multi-sensor data fusion

This system has capacitance sensor array, camera array and laser ranging finder. It detects erosion condition of steel-making converter's lining by 3 different methods and fuses data acquired by those 3 method using BP neural network to get more precise result.

Firstly, the data detected by those 3 methods undergo pretreatments which implements character abstracting and judging respectively. Then the data is sent to 3-layer BP neural network made up of input layer, output layer and connotative layer where the data is correlated and fused.

2.2.2 Fusion Algorithm

Because each method measures different physical variable and processes by different method, it is very difficult to correlate directly. Since neural network can learn by itself, BP neural network is used to fuse data.

The algorithm of BP neural network is trained by tutor and it is optimized by carrying out grads declining algorithm[5,6]. Training samples are input-output pair $\{X_i^k, T_j^k\}$. The input layer has m nodes; the output layer has n nodes; connotative node's and output node's transmission function are network f_h and net work f_o respectively. As the k_{th} sample is inputted, the input and output of each layer in the network are:

The input weighted sum of the h_{th} node at connotative layer is: $S_h^k = \sum_{i=1}^{m} W_{ih} \bullet X_i^k - \theta_h$ (1)

According output is:

$$Y_h^k = f_h(S_h^k) = f_h(\sum_{i=1}^{m} W_{ih} \bullet X^k{}_i - \theta_h) \tag{2}$$

The input weighted sum of the j_{th} node at output layer is:

$$S_j^k = \sum_{h=1}^{l} W_{hj} \bullet Y_h^k - \theta_j \tag{3}$$

According output is

$$Y_j^k = f_o(S_j^k) = f_o(\sum_{h=1}^{l} W_{hj} \bullet Y_h - \theta_j) \tag{4}$$

Given output error of network is:

$$E = \frac{1}{2}\sum_{j,k}(T_j^k - Y_j^k)^2 \tag{5}$$

According to grads declining algorithm, adjusted weight

W_{hj} connecting connotative layer and output layer is:

$$\Delta W_{hj} = -\eta \frac{\partial E}{\partial W_{hj}} = \eta \sum_{k} (T_j{}^k - Y_j{}^k) \bullet f_o{}'(S_j{}^k) \bullet Y_h{}^k \tag{6}$$

adjusted weight W_{ih} connecting input layer and connotative layer is:

$$\Delta W_{ih} = -\eta \frac{\partial E}{\partial W_{ih}} = \eta \sum_{k,j} (T_j{}^k - Y_j{}^k) \bullet f_o{}'(S_j{}^k) \bullet W_{hj} \bullet f_h{}'(S_h{}^k) \bullet X_i{}^k \tag{7}$$

here: η is learning rate, $f_o{}'(S_j{}^k)$ and $f_h{}'(S_h{}^k)$ are output node's and connotative node's transmission functions' derivatives respectively.

When sum error of all learning samples $E < \varepsilon$, stop doing iterative computations and the training for network weights and threshold value is over.

2.3 Simulation and Result

In this paper, simulating experiments on multi-sensor fusion detection are carried out based on 330 groups of data detected by a steel company during 1995-2000. 300 groups of data are used as learning samples and the other 30 groups of data are used as checkout samples. Table 2 lists the network parameters selected from simulating experiments; table 3 presents the compared results of the 4 methods' detection of 30 checkout samples.

Table 2 Network Parameter Config

Parameter	Value
M	3
L	4
N	1
f_h	*S Function*
f_o	*S Function*
Sum error（ε）	0.05
Initial weight (ω)	0.3
Initial threshold-value (θ)	0.5
Learning rate (η)	0.7
Number of samples	300
Training time	*35 minutes*

Table 3 Comparative Results of the 4 Detection Methods

Detection method	Capacitance method	Flame image method	Laser thickness Measurement	*Data fusion method*
Detection rate	82.5%	74.5%	85.4%	*92.3％*
Detection precision	*80.3％*	*70.8％*	*83.6％*	*88.7％*

Table 3 shows that in steel-making converter's hearth the abominable environment and a great deal of interference make it difficult to abstract signal and as a result, the 3 methods mentioned above are not able to get ideal detection result alone. However multi-sensor data fusion technology can solve the problem to some extent.

3. CONCLUSION

To detect steel-making converter's erosion condition by using single sensor or method will have poor results. In order to improve detection precision, this paper puts forward the idea that applies multi-sensor data fusion technology to detection of erosion condition of steel-making converter's lining. Because each detection method has different characters , the bug information model varies. As a result it is difficult to correlate them directly. Whereas neural network has such advantages as self teaching ability, parallel processing ability, excellent real-time performance, we adopt neural network to fuse data. And simulating results shows that applies multi-sensor data fusion technology based on neural network to detection of erosion condition of steel-making converter's lining will achieve relatively ideal effect.

4. REFERENCES

[1] Waltz E., Llinas J. Multisensor Data Fusion [M]. New York: Artech House INC., 1988.

[2] Fabrizio Russo, Giovanni Ramponi. Fuzzy Methods for Multi-sensor Data Fusion [J]. IEEE Transaction on Instrumentation and Measurement, 2000, 49(2): 840-843.

[3]Hua Yan,Fuqun Shao,Shi Wang. Simulation Of Lining Erosion Detected By Capacitance Method For Blast Furnace [J].Journal Of Iron And Steel Research,1999,11(4):61-64.

[4] Quan Liu，Yiwen Zhu，Yali Yang. The Design and Research of Thickness Measure System of Furnace Lining of Converter by Use of Laser [J]. Journal Of Wuhan University Of Technology，2001,23(12): 64-67.

[5]Shouren Hu，Shaobo Yu，Kui Dai.An Introcuction to Neural Network[M].Changsha：National University of Defence Technology Press，1999.

[6] Xia Youshen, Leung Henry, Bosse Eloi. Neural Data Fusion Algorithms Based on a Linearly Constrained Least Square Method [J]. IEEE Transactions on Neural Networks [IEEE TRANS NEURAL NETWORKS], 2002, 13(2):320-329.

[7] Pad Y C，Krishnaprasad P S. Analysis and Synthesis of Feed-forward Neural Network Using Discrete Affine Wavelet [J]. IEEE Trans on NN, 2000, 11(1): 73-75.

Fiberglass Molding Technology and Control Method On Tank Kiln

Chen Jing, Yuan Youxin
College of Automation, Wuhan University of Technology, Wuhan 430070, P.R. China
Email: jingchen680@163.com Tel.**: +86 (0) 27-62356851**

ABSTACT

This paper advances a fiberglass molding technology and control method on Tank Kiln. Of the main parameters that affect the fiberglass molding process, modern control method is proposed to overcome the blight and try to achieve an effective produce approach and high quality.

Keywords： Fiberglass Molding, Photo-fiber Sensor, Hybrid Control.

1. INTRODUCTION

Since 1950's, when the first Tank Kiln product line was gone into production, it has become a powerful impact to the traditional production technics because of its low power requirements, high efficiency, high quality, and adaptability to varieties of products. At present, the fiberglass produced by Tank Kiln is up to above 90% overseas.

A fiberglass molding is a process from glass melt to steady solid, it is also a movement of transmitting mass and heat and repressing all kinds of force balance. Compared with crucible drawbench, Tank Kiln drawbench has many advantages for fiberglass molding. For example, melting area is large, glass liquid surface is steady, clarifying time is long, and temperature undulation is very low. Besides raw material confecting and solution making technics, molding technics is also a main factor that affects on fiberglass moulding [1].

At present, technics and equipments used in fiberglass moulding of Tank Kiln drawbench production include: big Nozzle Plate of many rows and holes, forcible cooler with filling-in bits, long technis line of zolaesque double deck, airflow control, automatic drawbench with both tubes of big coil.

The main parameters which affect fiberglass molding include: temperature of Nozzle Plate, drawbench speed and drawbench technics inclination, etc. The affection and the control strategy of the parameters will be described in detail in the following sections.

2. RELATION PARAMETERS OF FIBERGLASS MOLDING

In the process of fiberglass Tank Kiln drawbench, the most important quality index is to keep the number of the precursor fiber invariable. It has been documented from theoretic analysis to manufacture practice, the relation is governed by

$$N = \frac{\mu L K^{2} V}{30.7 H D^{4}} \times 10^{-8} \quad \text{(mm)} \qquad (1)$$

Assume $K1 = \frac{LK^{2}V}{30.7HD^{4}} \times 10^{-8}$, Equation (1) can be re-written as

$$N = K_{1}\mu V = K^{2}VT \qquad (2)$$

where u is the viscosity of glass fluid, L is the length of the discharge spout, D is the diameter of discharge spout, K is the process modulus, H is the height of fluid surface, T is the temperature of glass fluid.

The fluid surface of Tank Kiln drawbench is stable and the other techniques parameters are invariable. So equation (2) describes the relationships of the number N of fiberglass and temperature T and drawbench speed V [2].

3. TEMPERATURE OF NOZZLE PLATE

From equation (2), the parameter N which indicates fiberglass molding is governed mainly by temperature. Nozzle Plate compensates the heat dissipation of glass fluid, and is also a key facility to maintain the molding temperature and viscosity and to insure the fiberglass molding. The function of Nozzle Plate is to adjust fluid temperature by yielding high temperature through low voltage and high current. As a result, Nozzle Plate controls glass fluid in a range of stable molding temperature.

The big Nozzle Plate of many rows and holes (1000-4000 holes) is usually chosen by Tank Kiln drawbench. When local temperature of Nozzle Plate exceeds the temperature range of fiberglass molding, the molding process will pause. In a consequence, the base condition of fiberglass molding is to keep on the temperature of Nozzle Plate uniform and invariable.

The making quality of Nozzle Plate

Nozzle Plate is a glow body like a resistance. Therefore it is not only a mechanical machine but also a electrical equipment. The components of Nozzle Plate can be regarded as a series-parallel connection resistance. When the material, thickness of each component is not uniform, the resistance value will be unequal and inevitably, which will arise areas of high or low temperature.

Installation and application of Nozzle Plate

In the process of Nozzle Plate installation, whether theinterface area between the electrode clamp and the electrode of Nozzle Plate should be rational or not directly affects the temperatures on both sides of Nozzle Plate. When reduce the interface area, the temperature around Nozzle Plate will rise. On the contrary, the temperature will drop.

In the process of application, high resistance alloy is formed by argyria that arises from moderate flow. With the time prolongation, high or low temperature areas will be formed when metal crystals elongate and change the resistance.

The cooler location

The cooler is the facility to translate natural heat exchange of fiberglass molding area to forcible heat exchange. The role of cooler is to keep the fiberglass shape and take away the radiant heat between discharge spout and fiberglass. As a result, glass fluid will retain within the appropriate molding viscosity range before it flows out of discharge spout.

The cooler with filling-in silver bits is usually selected because its performance of heat conduction is very good. Whether the distance between Nozzle Plate and cooler is far or near, it will make the temperature of molding area change remarkably. If the distance between each location is unequal, the difference in temperature on Nozzle Plate arises. The sound location is the place where the top of cooling pieces is at the same height with the bottom of discharge spout, the error is within ±1mm. There are four kinds of ordinary cooler location.

Airflow influence

Airflow divided into forcible and natural airflow is the main cause of the temperature undulation of Nozzle Plate. Forcible airflow is the one which Air-condition wind is blown by bellows into molding area, and natural airflow is the one which environmental gas is carried into molding area by fiberglass fast motion. Their function is to carry away the heat of molding area and to make its environment steady.

The main factors that forcible airflow affects the temperature of Nozzle Plate are as: inclination of wind blown, wind speed, temperature, humidity etc. In the process of fiberglass molding, air-condition wind isn't allowed to blow directly on Nozzle Plate. Only when drawbench is stopped, it is used to cool fiberglass root for drawbench. When cooling wind is blown onto the motherboard of Nozzle Plate or thermocouple, or the wind speed is instability, the temperature of Nozzle Plate will oscillate acutely. The solution to this question is to decide reasonable inclination of wind blown, wind speed and temperature, to try to choose automatic drawbench to reduce the frequency of the forcible airflow.

Wind speed of natural airflow is governed by drawbench speed, and its temperature is governed by environmental temperature. The environmental temperature around Nozzle Plate forms a temperature field, which decides the optimal temperature range of Nozzle Plate when the process of drawbench is steady. If this temperature balance is broken, the temperature of Nozzle Plate will oscillate acutely. A Fiberglass molding area is disposed to be close area controlled by steady airflow, which provides a favorable condition for fiberglass molding.

Temperature control of Nozzle Plate

In addition to the factors above, in order to keep the temperature of Nozzle Plate stabilize in the process of production, Nozzle Plate must be heated by tuning its temperature through low voltage and high current control to stabilize the temperature. Its control Principle is depicted in Fig.1.

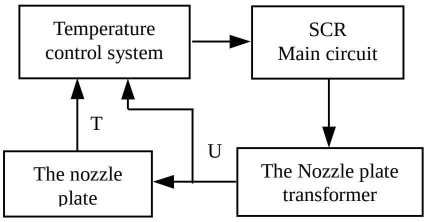


Fig. 1 Block diagram of the control

4. SPEED OF DRAWBENCH

The speed of drawbench is given by $V = n\pi D_s$ then equation (2) is rewritten as：

$$N = K_2 T n \pi D_s \tag{3}$$

where n is drawbench speed, D_s is the chees diameter.

When the temperature of Nozzle Plate is invariable, given the diameter of drawbench is D_{s1} at the beginning to enlace fiberglass, corresponding to it, the speed of drawbench is n_1, then the number N is computed as

$$N = K_2 T n_1 \pi D_{s1} \tag{4}$$

In a similar way, given the diameter of the drawbench equal to D_{s2} at falling off fiberglass, which corresponds the speed of drawbench equals to n_2, then the number N are

$$N = K_2 T n_2 \pi D_{s2} \tag{5}$$

To keep N (V) immovable, set equal (4) equals to equal (5), get:

$$K_2 T n_1 D_{s1} = K_2 T n_2 \pi D_{s2} \tag{6}$$

or

$$n_1 D_{s1} = n_2 \pi D_{s2} = n_2 (D_{s1} + \Delta D_s) \tag{7}$$

$$\Delta D_s = \frac{n_1 - n_2}{n_2} D_{s1} = \frac{\Delta n}{n_2} D_{s1} \tag{8}$$

From equation (8), to maintain N to be invariable, the speed of drawbench must be diminished when the diameter D_s goes up. Many methods can be used to change the speed of drawbench, the best one in which is to tune the speed of drawbench by a transducer. Figure 2 shows block diagram of control in the drawbench by a transducer.

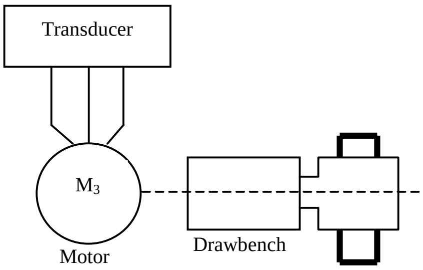


Fig.2 Block diagram of control in the drawbench by a transducer

The speed of drawbench mainly affects on the strain and cooling performance while fiberglass is formed. The strain is proportional to the speed of drawbench, but the variation rate of the strain is smaller than that of the speed of drawbench.

The coiling or drawing speed increases, which improves the cooling performance. In the process of production, the temperature of Nozzle Plate is very low for a long period of time when thick fiberglass is changed to thin fiberglass, this state is abnormal, which is caused by natural airflow speed increase too fast to improve cooling performance of Nozzle Plate after the speed of drawbench quicken. When cooling performance reinforces, the temperature of Nozzle Plate must be raised to reduce the amplitude of vibration of the original fiberglass root, which is advantage for steadying fiberglass molding. However, it is not always good with a higher speed and a higher temperature. The appropriate speed of drawbench and the temperature of Nozzle Plate are decided according to all kinds of original fiberglass.

5. TECHNICS INCLINATION OF DRAWBENCH

The techniques inclination of drawbench mainly affects on the fiberglass strain. Tank Kiln drawbench applies long technis line of zolaesque double deck, it can reduce inclination σ_1 and σ_2. Therefore the phenomena of breaking fiberglass arisen from the excessive strain may be decreased as shown figure 3.

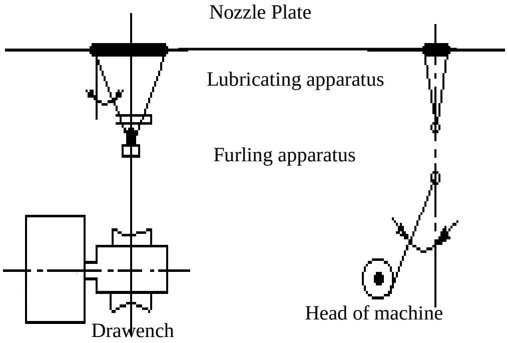


Fig. 3 Drawbench technics

Through observing and statistics, the locations where the fiberglass is breaking are on the upside of the lubricating apparatus. Therefore it is favourable for production to properly decrease the inclination of Nozzle Plate centerline. But the improper inclination σ_2 between the axes of Nozzle Plate and the axes of bolthead rarely cause the strain of the fiberglass increase and fiberglass break , however, it impacts on the characteristic furling and unfurling of certain product. In general, synthetically consider σ_1 and σ_2 when change technics line, then take as σ_1=9 ° and σ_1=17 ° -20 ° .

6. HYBRID CONTROL METHOD

According to above analysis, besides raw material confecting, founding technics, there are many factors impacting on fiberglass molding such as the temperature of Nozzle Plate, the speed and inclination of drawbench, etc. As far as the inclination of drawbench is concerned, it is easy to be fixed. In the case of controlling the temperature of Nozzle Plate and the speed of drawbench, the traditional way is that single point constant voltage-current control is used to control temperature, yet constant velocity control is used to control the speed of the drawbench. Thought the control performance of the traditional way is bad and because of the small coiling number, and low production efficiency, it should be replaced by other approaches.

Along with development and application of computer control technology and fiber optic network communication technology, it is possible that Tank Kiln drawbench is controlled automatically. DCS (Distributed Control System) and fiber optic sensing technique is applied to control the parameters, such as temperature of Nozzle Plate, the speed of the drawbench, glass liquid level and so on, in which there is good control performance, large the coiling number, and high production efficiency. Furthermore, it can achieve computer remote supervisory and control, and computer remote management shown as figure 4.

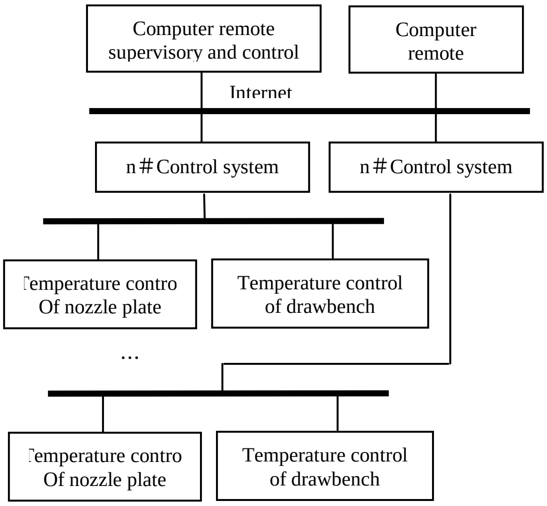

Computer remote supervisory and control
Computer remote
Internet
n＃Control system
n＃Control system
Temperature contro Of nozzle plate
Temperature control of drawbench
…
Temperature contro Of nozzle plate
Temperature control of drawbench

Fig. 4 Block diagram of DCS

7. CONCLUSION

It is a complicated process how to make favourably fiberglass molding. However, it is avoidable to break fiberglass in the process of production as long as we solve the main factor in molding technics and apply appropriate control methods to control fiberglass molding.

8. REFFENCES

[1] Qiyin Wu, Weiqun Zhang, "dicuss Simply the technology of fiberglass molding of the Tank Kiln with no natrite ", Fiberglass, 2000.5.

[2] Liangcai Wei, "Development trend of production technology in the fiberglass industry in China", Fiberglass,2000.3.

Chen Jing graduated from Wuhan University of Technology and received her MengSc and PhD degrees in 1997 and 2003, respectively. She is currently an Associate Professor in College of Automation at Wuhan University of Technology. Her research interest and practical experience are in the areas of control theory, computer control, and structural vibration control. She has published more than 20 technical papers in these areas.

Design of a Distributed Monitoring and Control System Based on DSP

Qin Juanying Wu Guoping Zhu Rongbo
College of Automation, Wuhan University of Technology Wuhan,Hubei, 430070, China
Email: qinjuany@mail.whut.edu.cn, wgp62000@163.com**, Tel:** +86 (0)27 87850114

ABSTRACT

The TMS320F240 device is the first member of a new family of DSP (Digital Signal Processor) controllers of Texas Instruments Incorporated. The system adopts it as basic hardware, and all communication programs were written in Visual C++ or C. A PC (Personal Computer) communicates with all front-end units (DSPS) through the public telephone lines manipulating central control functions. The system is cheap, transplantable and easy to realize. And it has been put to use.

Keywords: distributed monitoring and control system remote communication front-end units MSComm(Microsoft communication control) DSP

1. INTRUDUCTION

In an actual monitoring and control system, workstations and information of working states are often distributed, which makes the front-end units a multi-unit system. With the development of computer science and communication techniques, monitoring and control technique approaches unattended-remote communication, remote diagnosis and remote maintenance. When a PC works as a central controller and controlling all front-end units, they form a local area network (in other words, a distributed monitoring and control system) in which data is transmitted. A remote communication system based on modem and public telephone network is chose when considering the reliability of data exchanging, the complexity of the lines and the long distance for its cheapness and easy to realize. All these are conspicuous characteristics. The system has been put into use in a signal station satisfactorily. In this paper, the authors introduce the system on both hardware and software in detail.

2. DESIGNING OF HARDWARE

2.1 DSP TMS320F240 and its SCI(1)(3)(serial communications interface)

The TMS320F240 device is the first member of a new family of DSP controllers based on the TMS320C2xx generation of 16-bit fixed-point digital signal processors(DSPs). It combines the enhanced TMS320 architectural design of the `C2xLP core CPU for low-cost, high-performance processing capabilities. It uses an advanced Harvard-type architecture that maximizes processing power by maintaining two separate memory bus structures---program and data---for full-speed execution. This multiple bus structure allows data and instructions to be read simultaneously. Coupled with a four-deep pipelining, it allows the F240 device to execute most instructions in a single cycle.

Transmitting and receiving operations of 8-bit data are accomplished through TXD (transmitting pin) and RXD (receiving pin) connected with the full-duplex SCI. All SCI registers-SCITXBUF (transmitter-buffer), SCIRXBUF (receiver-buffer), SCIRXST and SCITXST (status registers), SCIHBAUD and SCILBAUD (baud rata register), TXSHF and RXSHF (shift register) are accessed by setting SCICTL (SCI control register). Data signals and handshaking signals, two different signals, are used in SCI communication operations. Character pattern, protocols and communication mode are determined by the settings of SCICTL. Features of data word format include: one data start bit, optional even/odd/no parity bit, one or two stop bit, length of data word, 8 bits. The SCI receiver and transmitter are double-buffered, and they both can be operated in full-duplex mode. The baud rate is determined by SCILBAUD.

2.2. Hayes modem

Before transmitting data, communication channels should first be set up through Modem. Hayes modem is easy to use and fixed conveniently, it can be connected to a mainframe or a slave by RS-232 interface, and it supports a set of general AT commands (a set of string command to control modems) and result codebook. Modem can accomplish initialization, dialing-up and hanging-up with certain fundamental commands (which can also be done by public telephone lines based TAPI(Telephone Application Programming Interface).

2.3 Structure frame of the system

The structure frame of the distributed monitoring and control system is shown in next page(fig.1). The PC and all front-end units are connected to the telephone network through a modem. Communication lines between both sides should first be set up through Modem Before transmitting data. Modem hang up and rescind the communication lines when the transmitting operation is completed.

DSP sends the data to be transmitted to the modem of PC through its SCI and modem, PC receives it as digital signals and implements central control and monitoring functions. The interface between DSP and modem can be designed according to RS-232C. Voltage is converted by making use of MAX232. The PC receives data transmitted from all front-end units and analyzes and makes decisions and sends commands to control the whole system continuously to ensure it working well together.

3. DESIGNING OF SOFTWARE(2)

Because the PC works continuously controlling the whole system, the reliability depends mostly on the design of the communication programs. Meanwhile, fluency of the communication channels is a premise. The software design task includes two parts: one is to ensure controlling the modem, and the other is to finish communication programs.

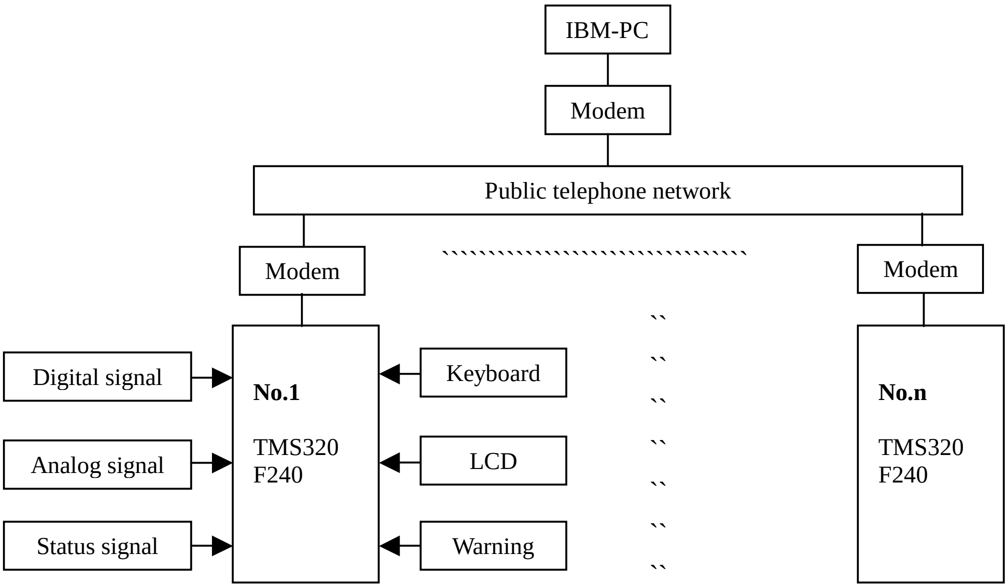


Fig.1 Structure frame of the system

3.1 Control the modem

The modem is operated by AT commands sent by MSCOMM control. Here are the main contents of the dialing up program written in Visual C++:

```
   //dialing up program
if(! m_Comport.GetPortOpen( ))
         //m_Comport is a variable of the MSCOMM control
m_Comport.SetPortOpen(TRUE);    // open the port
m_Comport.SetInputMode(comInputMideBinary);
         //   set input mode
m_Comport.Settings("9600,n,8,1");   //set baud rate etc.
m_Comport.SetRTreshold(1)     //generates an event each
                   time receives a character
m_Comport.SetInputLen(0);
m_Comport.GetInput( );       //   clear the data left
m_Comport.SetOutPut("     ATDT+Number");
                        //   dialing up in audio mode
//   waiting
```

Data transmitting completed, an AT command is sent to order modem hang up and rescind the communication lines. And the port also shut at the same time.

3.2 Communication programs for PC

Flowing chart of data communication is shown blow as Fig 2. PC sends different dialing numbers to distinguish different workstations. Transmitting instructions are also sent out, ordering each DSP transmit data and assure the quality of all information. Properties of CommEvent in MSCOMM control should be configured properly (for example, generates an event each time receives a character) to affirm the operation completed and to assure the reliability and real time of communication,.

Reading from SCI and writing to SCI are often viewed as two different tasks in serial communication. The program first do some initial work and set up monitoring program in proper time when necessary. Communication program listens on serial port, MSCOMM control generates a receiving event whenever received dada, and then data processing program verifies the data received. Events in the visual user interface call transmitting subroutine to accomplish data processing when a transmitting operation is needed. Processing data received should be top-priority because of the randomicity and real time of the communication events.

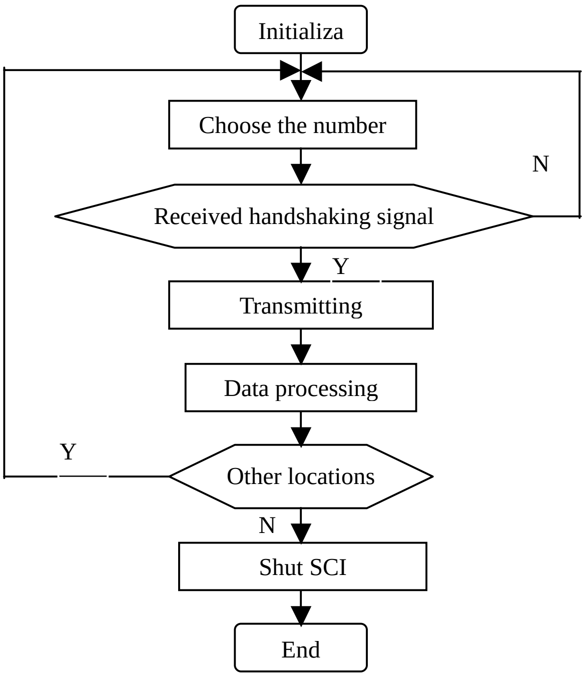


Fig.2 Flowing chart of data communication

Software designing of communication protocols is also important because the priority among the instructions is a problem. Here are the source codes for transmitting data frames:

```
Void CcomDlg: Senddata( )
{// TODO: Add your control notification handler code here
   int i;
   updateData(TRUE); // acquire data input by user
   SData[0]=0x1E;        // flag of start bit
   SData[1]=0x03;        // object address
   SData[2]=0x01;    // identity code of command frame
   SData[3]=0x08;            // length of command frame
   SData[4]=COMMAND0; // info of command frame
   SData[5]=0x02;        // number of data package
   SData[6]=0x08;             // check sum code
   SData[7]=0x0F;             // frame ending code
   unsigned char sum=0, count=0;
   CBytarray array;
   char sOutput[10];
   count=SData[3];               // length of the frame
   for ( i=0;i<count;i++)
      sum+=SData[i];            // compute the check sum
   SData[count-2]=sum;
   Auuay.RemoveAll( );      // empty the array
   Array.SetSize(count);      // setting size of the array to be
                              the length of the frame
   for ( i=0;i<count;i++)
      array.SetAt(i,SData[i]);    // store data into the array
   if(m_Comport.GetPortOpen( ))
      m_Comport.SetPortOpen(TRUE);
   m_Comport.SetOutput(ColeVariant(array));
                              // transmitting data
}
```

Confirming frame and repetition mechanism should be set to make sure that data transmitted successfully. If the same data frame has not been transmitted within certain times, it means that something abnormal occurred and the transmitting operation should be terminated, avoid plunging into a infinite loop.

3.3 Program design for serial interface communication
Figure 3 shows the flowing chart of the communication process.

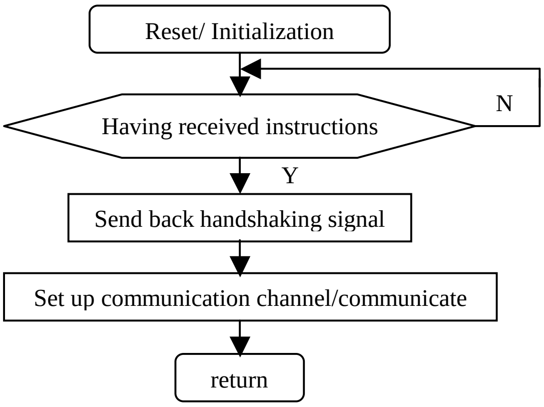

Reset/ Initialization
Having received instructions
N
Y
Send back handshaking signal
Set up communication channel/communicate
return

Fig.3 Flowing chart of the communication process

Source codes(written in Turbo C language):

```
volatile int *SCI_CTL=( vatile int *)0x7050;
                            // set address for SCI register
volatile int *SCI_STDATA=( vlatile int *)0x7059;
          // set address for transmitting data register
void SCI_Initialization(void)
                       // subroutine for Initialization
{
   SCI_CTL[1]=0x00;        //    reset
   SCI_CTL[0]=0x77;     // Character pattern, protocols and
                           communication mode
   SCI_CTL[1]=0x53;     // enable transmitter and receiver
   SCI_CTL[2]=0x25;
   SCI_CTL[3]=0x80;     // set the baud rate
}

void Send_Data()       // subroutine for transmitting
{
   int data;
   while(!(SCI_CTL[1]&0x20))       // ready to transmit
        SCI_STDATA[0]=data;          // transmit data
}
```

Dialing-up program should also be included in the main program of DSP, so that it can set up the communication channels and transmit data at anytime to assure the fluency.

4. CONCLUSIONS

The remote distributed monitoring and control system introduced above can be adapted for different practical systems according to different conditions. For example, if it is a large system of many workstations with a large amount of information to handle, SCI then can be extended (e.g. adopt CAN mode). As for the software, no obvious changes are necessary and just make sure the system coordinate timely.

The scheme above has been put into use in a signal station successfully, it operates with high effectiveness stably.

5. REFERENCES

[1] GUIDE BOOK for TMS320C240X DSP, WUHAN CHINA, P&S Information Technology Service Co., Ltd.

[2] David J Kruglinski, Visual c++ 6.0, Beijing: Chinese Machine Press, 1999

[3] Zhang Weixiong, Principles of DSP Controller, Beijing: Publishing House of Electronics Industry, 2001

Qin Juanying is a Full Professor in college of automation, Wuhan University of Technology (WHUT). She graduated from University of Electronic Science and Technology of China in 1969. Her research interests are in Control Theoty and Engineering, Computer Science.

Wu Guoping and **Zhu Rongbo** are both MS candidates of Control Theory and Engineering，college of automation, Wuhan University of Technology (WHUT).

Visual Federation Control Mechanism

Feng Zhe, Xu Dongping
Science and Technology of Computer, Wuhan University of Technology
Wuhan 430063, China
Email: DPXU@public.wh.hb.cn **Tel.:** +86(0) 27-86551167

ABSTRACT

This paper introduces the main idea and architecture of Visual Federation, visual in network distribution environment. This paper introduces the main composition of visual Federation Object Model. This model reduces data redundancy and transmission, so it can solve the conflict between scene creation speed, data transmission and the limited bandwidth in a certain extent. The writer defines a king of recursion enumeration language and realize the flexible describing of visual Federation Object Model. The result of this research is very important to directives about how to solve the problem of mutually operation in distributing interaction scene simulation control system.

Keywords: Visual Federation, Visual Federation Object Model, Mutually operation

1. THE MAIN IDEA AND KEY QUESTION OF VISUAL FEDERATION

In our human life, information is the same primary resource as energy and raw materials. Through the five sense (seeing, hearing, touch, olfaction, gestation), people can obtain information in which the 75% percent obtained through seeing. Therefore, Chinese people always say “the thing that saw by eyes is true” to stand out the important of visual information. But in the real life many phenomena is not visual, and sometime it’s too expensive if you want to know all sides of something. For example, in the transportation system, the training of control the ship avoiding collision in complex water area, particularly in the condition where several ships running as same time, it is a big challenge that drivers have to face. Several decades ago, the training of ship drivers must in the true environment, but the educate expense is expensive doubtless. Several decades ago, if you want to train a batch of tankman the ability of battle, can you send them to the true battlefield like the ship drivers? The answer is affirmative negation, so these tankman lack the true war experiences in the learning stage is an unavoidable pity. People pursue the feeling of “the thing that saw by eyes is true” in process of learning and studying. As the development of communication and technology, many things that can’t do or cost too expensive in the cast can be satisfied with paying more little with the support of “Virtual Reality” technology today[1].

From the above examples we can abstract a description as follows: To express a complex environment of three-dimensions scene, in this scene there have one or many complex running active objects O_i(i=1…n), each object has different observational point V_i(i=1…n), The motion of these objects will affect mutually in a certain way, or mutually constitute a certain kind of connection R_{ij}(i,j=1…n). Among them: S is a digitized model constituted with three-dimensions geometry model, additional texture, environmental illumination and etc. O_i（i=1…n）is a digitized model constituted with three-dimensions geometry model, additional texture, motion control model and etc. V_i(i=1…n) is the viewpoint coordinate binding of the position and direction of O_i (i=1…n) . R_{ij} is the relation of O_i(i=1…n) and O_j(j=1…n), such as transfer relation, possible bump relation, relative position restriction relation, triggering condition relation varying from time or position. We can use R_i to denote the relation of O_i(I=1…n) and S, such as resemble relation, floating relation, immersion relation and etc. as the base of the description, we can make a depiction that the realtime perspective of the binding viewpoint of motion multi-objects in the same scene. Actually, V_i(i=1…n) is observation simulation point of every controller in Simulation network. It not only relates to the motion of $O_{i,}$ but also relate to the motion of O_j. Because of its has complexity of computing, the mutual operation O_i(i=1…n) must be realized in the distributed environment. That is to say, parallel computing of the motion of O_i(i=1…n) must be realized in network environment. Each O_i(i=1…n)’s computing has high degree cohesion, as the same time it must ensure the relation with O_j （j≠i,j=1,2…n）and maximum reduce data transfer on the network to ensure visual continuity of each V_i(i=1…n). That is the idea of “Visual Federation” we introduced in this paper.

2. VISUAL FEDERATION ARCHITECTURE

Visual Federation Architecture is constituted with LAN network(LN), realtime service(RTS) and visual database(DB) [2]. DB is constituted with federation object model(FOM), common scene(S) and federation member F_i(i=1…n). Federation member F_i(i=1…n) is constituted with O_i,V_j(i,j=1…n). O_i is constituted with mainpart and subpart. The interaction between RTS and Fi(i=1…n) can be carry through manipulable-table object publicized message po(unordered O_i i=1…n), object order message co ⊂ po, running object synchronized similitude message sy. FOM adopt federal describes language to express the relation of R_{ij}. The visual federation architecture is shown as figure 1.

Realtime Running Service(RTS) look assignable right in federation and control property as “manipulable-table and publicized object” po to broadcast to the whole federation environment. When federation member join in the federation environment, it can request “object order” co according to its own control requirement. After federation member receives control right, it calculate the orientation of M_{ik}(k=1…m) according to dynamics motion control model and send “synchronization similitude” signal to RTS, and then, RTS send broadcast to sy. After each member receives sy messages it calculate the M_{ik}(k=1…m) according to “similitude dynamics control model”. Therefore, the realization of cooperational publicized computing of each motion object’s

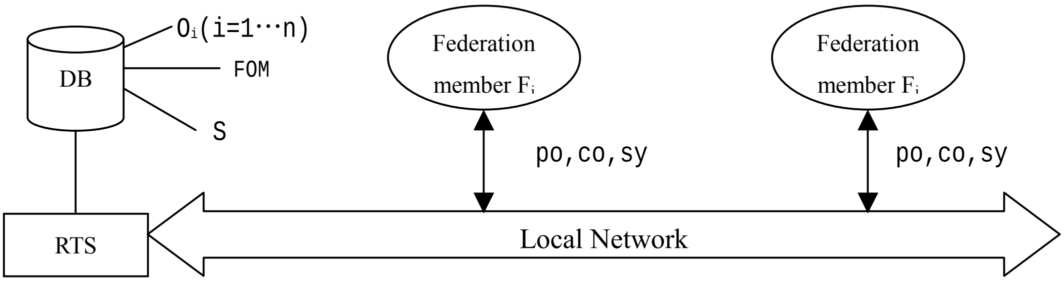


Fig1 The vision federation architecture

position in network environment is not simplification computing of the orientation of non-main control federation member, but to debase motion control computing complexity of each federation member. On the other side, in federation system only transfers similitude orientation information of each federation member O_i's mainpart Mik(k=1…m) to minish network transmission burthen.

3. VISUAL REALITIME MUTUALLY OPERATION IN NETWORK DISTRIBUTED ENVIRONMENT

In actual application, it needs lots of calculation to create three-dimensional animation scene with real feeling. And at the same time, scene simulation always does not exist alone, but is a part of network distributed environment and needs interaction with control simulation subsystem. So the system have to seek a kind of balance among the quality, precision, realtime of the image and the rapid response of control simulation subsystem. As a result, realtime and interaction of the three-dimensional animation simulation and the vivid display of dynamic scene are the key technology of the scene simulation all through. The scene simulation based on PC is limited by the bus, the capability of CPU and memory, so we need ensure the realtime and interaction of three-dimensional animation simulation.

The scene interaction in network distributed environment mostly indicate the capability of cooperation between scene simulation and control simulation, especially interaction capability of the data between entity. Scene simulation needs control simulation subsystem provide information to describe reality system. So, scene realtime interaction mostly include realtime transport of control simulation order parameter and realtime creation of scene.

Thus, we must think over the conflict between scene creation speed, data transmission and the limited bandwidth. In a simulation system, simulation model run in control simulation system and animation model run in scene simulation subsystem. Animation model needs each parameter from simulation model drive three-dimensional object moving in virtual space. These parameters are transmitted in the network. In order to solve the conflict between scene creation speed, data transmission and the limited bandwidth, reducing data redundancy and transmission is the essential way. So rational distributing describe of scene parameters is the key to solve this problem. In the follow content, we'll introduce Visual Federation Object Model (VFOM).

In the base of rational distributing describe of scene parameters, integrating network communication, scene control and other technologies, to realize data alternation between simulation entities.

4. VISUAL FEDERATION OBJECT MODEL VFOM

The establish of Visual Federation Object Model VFOM is the importance measure to resolve the question of visual interaction. VFOM include motion command table, announced motion object table , and motion object parts table. The interaction of visual simulation subsystem and control simulation subsystem is decided by the proclamation order relation which defines in VFOM.

Motion Command Table is a connection window between control simulation federation members and visual simulation members. The data of command window will refresh each △tc（as 300ms）, control simulation federation members to send motion command to visual simulation members through network. Motion Command Table mainly record data message such as the ID of motion object, Command Speed, Command similitude Orientation, and Location Flag.

Announced Motion Object is an object which could be controlled by another federation member. It means a object of the whole motion. Announced Motion Object Table contains object motion attributes which Control Simulation System maybe public motion command for. Announced Motion Object Table mainly record data message such as the ID of Announced Motion Object, Current Speed, Speed Granularity, Command Speed, and Command similitude Orientation.

Motion Object Part Table descripts the parts'motion link and the motion attributes. Motion Object Part adopts part coordinate orientation and this is related to one object or another part. Motion Object Part include announced mainpart which can directly get the motion command from other federation member, and unannounced subpart which can move by the traction from the motion of mainpart. The relation of traction motion is defined by the traction orientation equation and subpart can tow lower subpart.

Motion Object Mainpart Table mainly records information of ID of mainparts, mainpart motion upper limit of freedom degree M, mainpart motion lower limit of motion freedom degree L, mainpart current speed V, mainpart moving speed C and mainpart similitude location data Pw which sending from another federation member, in each frame refresh time $\triangle t_R$ (30ms-50ms) mainpart speed increment or we can say it reflecting the mainpart itself motion inertia feature a, the indicator(pointer) B_L which directing motion subpart, etc. Motion Object Subpart Table mainly records information of ID of subparts, subpart motion upper limit of freedom degree M, subpart motion lower limit of freedom degree L, the indicator (pointer) S_P which directing motion subpart, motion traction equation.

All the tables we mentioned can be divided into two kinds: one is static state data which can be dealt with beforehand. It not only include inherence property of object such as speed granularity of Announced Motion Object, motion object mainpart's motion upper limit of freedom degree, but also include the motion object part, the Initial value of announced motion object motion parameter. Traction location equation can be give as static data beforehand. The second is dynamic data, viz. the data in motion command table, which is provided by control simulation member when the system running.

Since each frame of simulation visual object needs six freedom degree statement parameter information, if making motion orientation synchronized in each animation refresh cycle, the transmission quantity between simulation subsystem and visual subsystem will be too big to achieve realtime data transmission. Therefore, we adopt method as follows: enlarge the synchronous cycle; when the process of system running, control simulation member transfer the refreshed command speed and synchronous similitude orientation command of announced object and announced parts only when operation affair happened or synchronous cycle started, video frequency of announced object and announced parts, unannounced parts video frequency orientation dealt with by animation model. Above method ensure motion visual object statement parameter error in visual simulation within the permitted range. As the same time It can maximum debase the data redundancy and the data transmission quantity.

From the Visual Federation Object Model we can see that motion object part divides into announced mainpart and unannounced subpart can maximum debase the information transferred in the network. Or we can say that in the limited network bandwidth, according to the traction equation of mainpart and subpart, calculating the motion of subpart by mainpart table's data announced on the network, It can increase realtime of simulation and verisimilitude degree of dynamic visual.

5. VISUAL FEDERATION OBJECT DESCRIPTION

Visual Federation Object Model VFOM decides visual control announcement characteristic，motion and explanation of its transfer relation. It adopts grammar G= (N,∑,P,S) to define the O-model language and realize the text description of VFOM. N is non-end symbol collection, ∑ is end symbol collection, P is rule collection, S is start symbol collection.

N={O,M,A,V,DK,H,L,F,T,X,Y,W,S, ⊕, ⊙}
Σ={#OBJECT_TABLE, #OBJECT_TABELEEND, {, }, [,], <, >, step,now,
#MAINPART_TABLE, #SUBPART,
#SUBPART_END,BEGIN,END
#MAINPART_TABLEEND,
mdof, ldof, ox, oy, oz, oh, op, or, =, :, +, -, *, /, constant, x, y, z, h, p, r, sin,cos, log,tan, acos, asin, atan, sqrt, pow, exp, ceil, floor,fabs, fmod,round, sqr }
P:
S::=OM
O::=#OBJECT_TABLE{[obj_ID]{AV}}$^{+}$#OBJECT_TABLEEND
M::=#MAINPART_TABLE{<mainpart_ID> BEGIN{[obj_ID]DAV}{#SUBPART<subpart_ID> H #SUBPART_END}$^{+}$ END}$^{+}$#MAINPART_TABLEEND
H::={[obj_ID]DW}|BEGIN{[obj_ID]DW}{#SUBPART <subpart_ID>{[obj_ID]DW} H
#SUBPART_END}$^{+}$ END|NULL
D::=mdof:L ldof:L
A::=step:L
V::=now:L
L::= constant，constant，constant，constant，constant，constant
W::=ox=F oy=F oz=F oh=F op=F or=F
F::=F ⊕T|T
T::=T ⊙K|K
K::=(F)|Y(F{,F}*)|X| constant | variable
X::=x|y|z|h|p|r
Y::=sin|cos|log|tan|acos|asin|atan|sqrt|pow|exp|ceil|floor|fabs|fmod|round|sqr
⊕::=+|-
⊙::=*|/
constant::= integer | real number |symbol constant
variable::= ASCII code cluster started with letter

F if traction motion equation. maiopart_ID is mainpart identification. Obj_ID is object identification in SOM. Subpart_ID is subpart identification in VOM. {}$^{+}$ Denote once or more repetition. {}* denote none or more repetition.

The VFOM text method belongs to the recursion enumeration language. The purpose which adopting recursion enumeration language to realize the description of VFOM is to ensure its flexibility and VFOM's comparative independence.

6. CONTINUE RESEARCH

This idea of scene simulation was used and achieved in the research of the scene simulation of 1750 ship dredger work, that is one of 95 key projects in china. But the description of the change of relationship between flexible object description and visual object needs more research.

7. REFERENCES

[1]. Alexander B. Aranson . Computation and applications of the Newton polyhedrons. Mathematics and Computers in Simulation .-2001,57(3-5).-155-160

[2]. Xu Dongping. Real Time Dynamic Interactive Visual Simulation: [Ph. D. thesis]. Wuhan University of Technology，2001

The Study and Application of OGSA Grid Based on Digital Manufacturing

Zhang Fan, Zhou Zude, Liu Quan
School of Electron-mechanical Engineering, Wuhan University of Technology
School of Informational Engineering, Wuhan University of Technology, Wuhan, Hubei, China
Email: bluedak@sohu.com zudezhou@mail.whut.edu.cn qliu@public.wh.hb.cn
Tel: 13071221331 027-87651445 13908659571

ABSTRACT

In the paper, the mode that based on the research of digital manufacturing was brought forward. It takes advantage of the characteristic distance resources co-share of high capability of the Open Grid Services Architecture (OGSA). The component structure and achievement of function of the Digital manufacturing grid are emphatic introduced. The mode, inheriting the characteristics from OGSA and digital manufacturing, achieve the co-share of design, manufacture, information, technology resources to great degree, and overcome the obstacle from the distance of space gap among different corporations.

Keywords: Digital Manufacturing; Grid technology; OGSA; Manufacturing Grid; Virtual enterprise

1. INTRODUCTION

The development of network & information technology and economics globalization make the communication of people more convenience and speedy, and inject new energy for some traditional trade. Manufacturing is one of the trade which achieve those benefits. Grid is one of the great achievement in information times, it is a tendency of future information technology development. The essential characteristic of grid is high capability distance resources co-share. In the international environment of information and economics globalization, there are requirement that manufacturing take advantage of network especially Internet span the space gap among different corporations to provide technology support environment and means, these are based on integration in information and business process, resources co-share among corporations, carry out different place cooperate design and manufacture, network sales, provision management for the corporation. The support that grid technology provided agree with the requirement of digital manufacturing, so grid became the best candidate that develop and meet digital manufacturing. The Open Grid Services Architecture based on digit manufacturing can carry out the co-share of distance resources of design, producing information, technology and overcome cooperating obstacle from the distance of space gap among different corporations.

2. THE CONCEPTION AND HARACTERISTICS OF GRID

2.1 The Conception of Grid

Grid is a new technology which base on internet. Generally, grid assemble whole internet to a huge super computer. It will achieve co-share of calculate resources, store resources, data resources, message resources, knowledge resources and specialist resources. At present, the most influencing defining is the one defined by Ian Foster in 2001: resources sharing and coordinated problem solving in dynamic, multi-institutional virtual organizations[1]. Grid connect with different structures resources distributed wide-ranging through speedy co-share network, solve single problem that generally need much CPU or memories disposal and visit. Compare with traditional internet and web, grid would achieve the connection of computer hardware and web page and try to achieve all-round connection of all resources in network.

2.2 The Characteristics of Grid

The most important characteristic of grid is not its scale but co-share of resources. therefore, the grid must have single image space that can shielding hardware border so that achieve diaphanous distant resources visit and clear off resources isolated island; support multi-manage realm and station autonomy so that guarantee the relative independence of the resources; support efficient security and compatible fault so that guarantee the security and secrecy of all resources providers in the whole system; dynamic flexible so that guarantee the dynamic enlarge and repeal of the grid resources. We can get several important differences hereinafter if we compare Internet with grid[2]:

(1) Grid technology base on Internet. Grid is not the alternative of Internet but the organic combination and development of Internet, high capability computer and data resources.

(2) Grid has higher capability than Internet. There are more higher capability computer on the grid, the system architecture of grid can take advantage of resources effectively, so the calculated speed of grid and the disposal speed of data can be improve greatly.

(3) Grid has knowledge produce feature that Internet has not. Internet isn’t create or produce knowledge, after produce knowledge by other ways people “put it in Internet” so that user can research. Grid can automatic produce knowledge according to the requirement of user, it can get original data from data resources by high capability computer, so it can operate particular program to process knowledge.

(4) Grid is more integration than Internet. Convenient for use, grid shall not like Internet that provide several million web station and even user search difficulty according to his requirement, grid shall like a machine in logic.

***Project support:** This paper is supported by The National Planning Project on Technology: Research and Industrialization on the Platform of Cooperating and Sharing Resources Based on Network Manufacture(Project NO. 2003BA103C).

3. MODE OF GRID BASED ON DIGITAL MANUFACTURING

3.1 Grid and Digital Manufacturing

Grid technology is named the third information technology tide after internet and web, but from the development history of grid we can see that today application mainly focus on high power physics experiment, biology gene test suchlike advanced science, apply to the feature of grid gathering calculate capability. The essence of grid is the co-share advantage of high capability distance resources still having not been enough developed.

Since middle period of last century, the traditional manufacturing have been combined with computer technology, information technology, network technology, controlling technology, new material technology and advanced managed technology. Manufacturing coming into digital manufacturing times gradually which virtual enterprise gleaned resources information rapidly according to the demand of user and analysis, scheme and reorganize for products information, craft information and resources information, thereby achieved the design of the products and the simulation of the function and the prototyping manufacturing then produce the products which accomplish with user's demand[3]. Since our country went into WTO, manufacturing had been into the international market of economics globalization and faced to rigorous challenge. We adapt to the dynamic and competitive times none but accelerate the informational progression of manufacturing. In the international environment of information and economics globalization, there are requirement that manufacturing take advantage of network especially Internet span the space gap among different corporations to provide technology support environment and means, these are based on integration in information and business process, resources co-share among corporations, carry out different place cooperate design and manufacture, network sales, provision management for the corporation. However, unnecessary enterprise information are serious, gross data increased rapidly. We can make whole design speed increase multi-times if we take advantage of some new technology to resolve the difficulty of the calculation, thereby the development of manufacturing would go beyond. Fortunately, the high calculation capacity of grid can achieve operating rapidly by hundred-fold even thousand-fold operating speed of common high capability computer and promote the advantage of resources and reduce redundancy through the resources of co-share scope. To the grid stage, it create a virtual cooperated working space for people, so we can see the whole design process of the equipment from our desk workstation real time and real spot and virtual operate natively.

3.2 The Basic System Architecture OGSA Based on the Grid Mode of Digital Manufacturing

The production was not the produce process of single corporation but the product whole life period cooperated of virtual enterprises league for nowadays manufacturing came into digital manufacturing times[4]. The digital manufacturing enterprises (virtual enterprises) based on grid can make resources achieve flexible co-share in wide range. The grid system architecture of digital manufacturing enterprises we mention here adopt OGSA (Open Grid Services Architecture) [5], OGSA is the most new grid system architecture at present that is "services architecture", OGSA is more beneficial to achieve flexible, unanimous and dynamics co-share mechanism compare with the traditional architecture of five layers of sand leakage.

The application of grid mode that based on digital manufacturing is looking for advanced co-share based on existing technology for former distributing differ structure resources instead of displacing it.

OGSA based on digital manufacturing has three layers structure(Fig1):

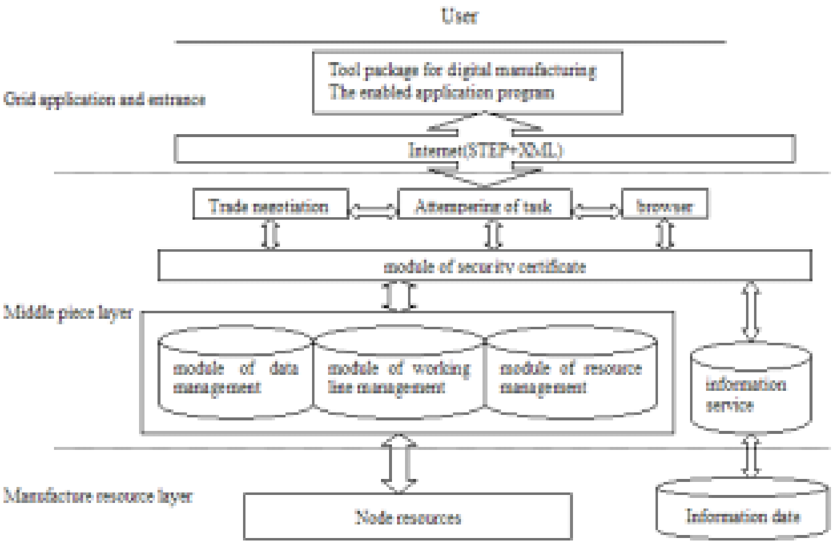
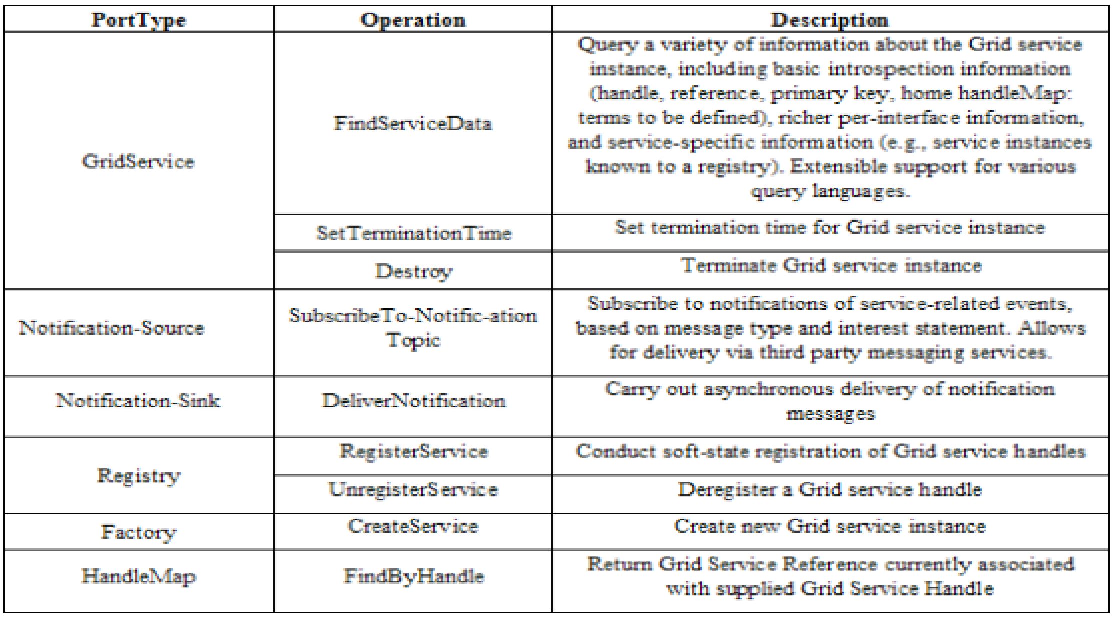

User
Grid application and entrance
Tool package for digital manufacturing
The enabled application program
Internet(STEP+XML)
Trade negotiation
Attempering of task
browser
module of security certificate
Middle piece layer
module of data management
module of working line management
module of resource management
information service
Manufacture resource layer
Node resources
Information date

Fig.1 Basic system architecture of OGSA Grid Based on Digital Manufacturing

① The bottom layer is called manufacture resources layer, namely service contact layer: upwards offer co-share resources of grid such as manufacture equipment resources (numerical control machines, working line), software resources (design software: AutoCAD, UG etc.), calculated resources, store resources, human resources, data resources etc. These resources maybe physical or logical thing such us distributing document system, computer clan or distributing computer pool etc. The manufacture resources layer shall achieve the fundamental function including the inquire mechanism (discover the structure of resources and condition) and the resources management mechanism of control service quality. The former allow inquire its structure, state and function, the latter offer the control of distributing service quality, The corresponding component in Toolkit detect the information of the feature of useful software and hardware resources in the module of information date, current load, state etc. and pack it for upper levels treaty to calling, the first we will describe and assemble various resources to service (described by XML), thereby achieve the agreement of resources, information and data, and make distributing system management achieve standard joints and actions.

② The second layer is middle ware layer: This layer uses the Globus Tookit3 which based on OGSA for the grid flat, it includes all the general functions of grid, such as security certificate, purview and authorization, dynamic land, information service, resources management, working line



PortType	Operation	Description
GridService	FindServiceData	Query a variety of information about the Grid service instance, including basic introspection information (handle, reference, primary key, home handleMap: terms to be defined), richer per-interface information, and service-specific information (e.g., service instances known to a registry). Extensible support for various query languages.
	SetTerminationTime	Set termination time for Grid service instance
	Destroy	Terminate Grid service instance
Notification-Source	SubscribeTo-Notific-ation Topic	Subscribe to notifications of service-related events, based on message type and interest statement. Allows for delivery via third party messaging services.
Notification-Sink	DeliverNotification	Carry out asynchronous delivery of notification messages
Registry	RegisterService	Conduct soft-state registration of Grid service handles
	UnregisterService	Deregister a Grid service handle
Factory	CreateService	Create new Grid service instance
HandleMap	FindByHandle	Return Grid Service Reference currently associated with supplied Grid Service Handle

Table 1 OGSA Grid service interfaces

③ The third layer is grid application and entrance: the application layer of OGSA based on digital manufacturing need exploit tool package for digital manufacturing, namely ,it take advantage of grid enabled language and the enabled application program of grid tool exploited. The joint and bind of middle ware layer can be described and discovered by XML and can directly support the intersect of Internet agreement with other software application based on XML massage. At the same time, utilizing the Standard for the Exchange of Product Model Data that resolve problem of the product data co-share, we create resources co-share and cooperation work environment for different manufacture service nodes[7].

3.3 Example of Application Base on the Digital Manufacturing of OGSA

The resources of grid base on the digital manufacturing in the virtual enterprises mainly include the promulgation, discovery and distribution of computer resources, storage resources and numerical control machines of production equipment etc. Due to the widely distribution of these resources in the geography area, and the large differ structure existed between the resources, the main task of the grid is to organize these resources and supply them to the users. GRAM (Resources Allocation Manager) in the Globus mainly conduct the request of resources, execute long-distance application, distribute resources and management activity etc[8]. Herein concretely explained according to the simple execution flow of the task request(Fig.2)

Firstly, users must send out request to the grid for agent submission task (e.g. some spare parts of motors), here, users can send request according to the RSL (Resources Specification Language) and submit the task to GRAM for transaction. In short, when users submit a task (e.g. spare parts of automobile), send a request for task transaction (e.g. quantity, specification, deadline of delivery of the spare parts,) to the Gatekeeper of long-distance computer. The Gatekeeper’s main task is to transact the distribution of request, identify the safety each other and establish a job manager for this task. The job manager analyzes the task parameter according to the RSL description in this request, then starts and monitors the execution of the task, and finally sends information of the task condition to the users. In this course, the job manager gets information of nodes via information service of the grid middle ware, and users negotiate about price via resources nodes in the trade service module. Both sides feed back the respective results to the job manager. The job manager adjusts granularly according to the task parameter, character, load and communication condition of every node, and distribute the task to the proper resources nodes for execution as well. During the execution, job manager must feed back the task state and results in the resources operation nodes to the users at any moment and adjust according to the task condition.

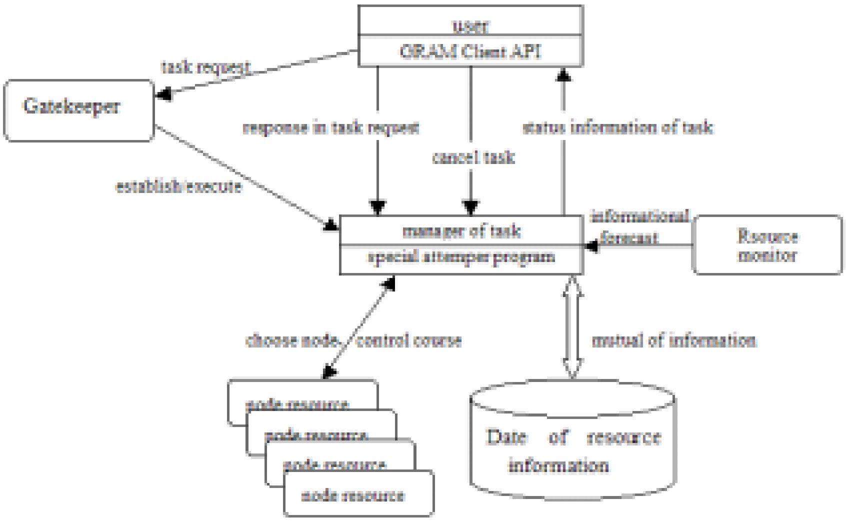


Fig.2 The simple execution flow of the task request

4. CONCLUSIONS

The Open Grid Services Architecture(OGSA) based on digital manufacturing can achieve the co-share of resources of design, manufacture, information, technology and overcome the obstacle from the distance of space gap among different enterprises. It offer support for achieve agile manufacturing and the operation of virtual enterprises, thereby, make it impossible to set up complementary cooperate enterprises with the feature of digital, flexible and agile. For the integration in information, process and resources among enterprises in network environment, it achieve the optimal operation of object line, massage line and value line in cooperate manufacture process and make the manufacturing of the whole manufacture grid system produce the high quality products meet market demand by cheaper cost and shorter exploit period.

5. REFERENCES

[1]. I.Foster, KesselmanC, NickJ,etal. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration[EB/OL]. http://www.Globus.orgresearchpapersogsa.pdf, 2003-02.

[2]. Zhangliqing, FanYushun. Grid Technology and its Application in Manufacturing. Aviation Manufacturing Technology.32-37. 2003-2.

[3]. Zhouzude, Ligangyan. Actuality and development of digital manufacturing. China Mechanical Engineering, vol. 13, no.6, pp. 531-534, March. 2002

[4]. Zhou zude, Liu Quan. Design and Reaserch on the Platform of Network Manufacture Product Electronic Trading. The Third International Symposium on Multi-spectral Image Processing and Pattern Recognition 2003. 2003,10.

[5]. Duzhihui, Chenyu, Liupeng. The Grid calculation. Tsinghua university publishing house.2002.

[6]. http://www.gridforum.org/ogsi-wg/drafts/ogsa_draft2.9_ 2002-06-22.pdf.

[7]. Liulanli, Yutao, Shizhanbei, Fangminglun. Rapid Manufacturing Grid and Construct of its Service node. Computer Applications. Vol.23,No.9,Sept.,2003.

[8]. I.Foster, A.Roy, V.Sander. A Quality of Service Architecture that Combines Resource Reservation and Application Adaptation.8th.International Workshop on Quality of Service,2001

Fan Zhang, male, was born in July 1977, is a postgraduate in School of Electron-mechanical Engineering, Wuhan University of Technology, his research interests is Digital Manufacturing.

Zu-De Zhou, male, professor, tutor of doctor, is the President of Wuhan University of Technology, his research interests are CNC Theory and technology, Intelligent Control, Digital Manufacturing, Reliability and Fault Diagnosis of the Modern Manufacturing Systems and etc.

Research on NN and RKB Based Expert System of Resistance to Corrosion of Sulfate on Concrete*

Luo Zhong, Qiong Jiang, Fei Huang, Jingling Yuan, Hao Pan, Qiwei Tong
Department of Computer Science, Wuhan University of Technology
Ma Fang Shan Campus, Wuhan, HuBei (Province), China
Post Code: 430070 **Tel:** +8613071226453 (Cell Phone)
E-mail: jiangqiong2651@163.com jiangqiong@mail.whut.edu.cn

ABSTRACT*

The problem of sulfate corrosion on concrete has been focused by the people all over the world. However, the concrete field knowledge is so complicated that it is hard to conclude some certain rules and express them. Therefore, a kind of ES based on relationship knowledge base (RKB) and neural network (NN) was designed in the paper by analyzing on limitations of traditional ways about knowledge representation, knowledge acquisition and knowledge discovery. In the paper, knowledge was stored in the RKB and data table. One data table can relate well to another by a special index field. Additionally, because neural network (NN) has abilities in self-learning, associative memory and parallel distributed computing, a kind of distributed computing algorithm was proposed in the paper, which was used to solve many problems, such as knowledge acquisition, knowledge discovery, parallel reasoning and so on. Due to using RKB and NN, the performance of the ES has been improved greatly.

Keywords: Expert System; Neural Network; Distributed Computing; Relationship Knowledge Base; Resistance to corrosion of sulfate

1. INTRODUCTION

Concrete, as a structure material, has been widely used in all facets of life. Nevertheless, due to the corrosion of sulfate, many concrete projects have been destroyed in the period of validity, which severely affects the lifetime of projects and makes a great economic loss. Therefore, it is important for concrete safety to study on the breakage mechanism and protection of resistance to corrosion of sulfate on concrete.

Although much knowledge and experience about concrete have been accumulated through a lot of practices, they can't worked well because of all kinds of limitations on time and space. As a result, ES is considered to solve the problem. Expert system is an intelligent system. It can analyze and conclude existed knowledge without considering environment effects, and then form a new system by self-learning, which can handle not only original knowledge but also new things.

2. BASIC STRUCTURE OF EXPERT SYSTEM OF RESISTANCE TO CORROSION OF SULFATE ON CONCRETE

The program design method based on knowledge was adopted in the ES and professional experience was used to make intelligent conclusion, just as an expert. The structure of the ES is in the following:

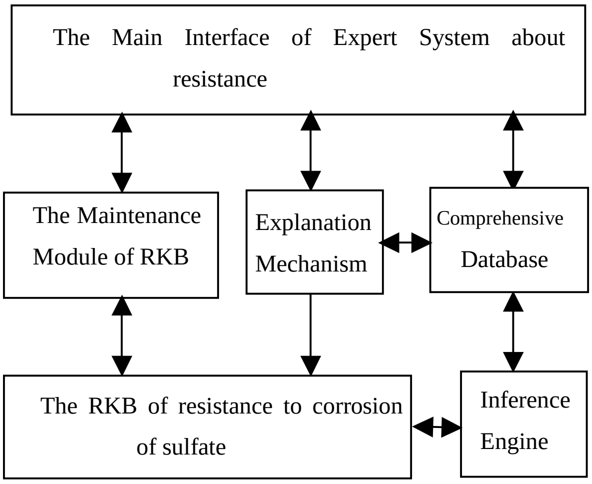

The Main Interface of Expert System about resistance
The Maintenance Module of RKB
Explanation Mechanism
Comprehensive Database
The RKB of resistance to corrosion of sulfate
Inference Engine

Figure 1 The structure of the expert system

It comprises six modules:

1) Relationship Knowledge Base: professional knowledge about resistance to corrosion of sulfate on concrete, acquired knowledge from the outputs of NN and mined knowledge from historical database and original knowledge base are stored in the RKB.
2) Comprehensive Database: it is made up of initial data about special problems and all inner calculated information, including descriptions of problems, inner results and process recorders.
3) The Maintenance Module of RKB: the RKB is updated and modified by database managers in the module.
4) Inference Engine: it chooses appropriate knowledge in the knowledge base, and then employs NN and data mining technologies to infer the best results and mine new knowledge. Then, updating and modifying the RKB at real time.
5) Explanation Mechanism: it answers users' questions and provides explanation about the process of calculating solution.
6) The Main Interface: it translates experts and uses' inputs into systemic inner format; on the contrary, transforms the output messages into the information that people can understand easily.

* This paper is supported by the "Ten Fifth" Key Project

3. NEURAL NETWORK EXPERT SYSTEM

The development of NN provides a new way for study on artificial intelligence and expert system. NN regards non-linear parallel distributed computing process as kernel, so people can make use of its abilities in self-learning, associated memory and parallel process to solve many problems in ES, including knowledge acquisition, knowledge discovery and parallel reasoning.

Of course, NN also has some obvious shortcomings. For example, the reasoning process of NN is a black box and people can only see inputs and outputs, but the inner steps of reasoning process hardly can be explained. Moreover, ES owns independent knowledge base easy to maintain, but the knowledge in NN is stored with a fixed structure and not easy to modify and supplement. So, NN needs to be combined with traditional symbol process in order to establish a good performance neural network expert system (NNES), which is characterized by massive parallel distributed process, knowledge acquisition automatization, self-organization reasoning, and parallel association.

In addition, NNES can cope with problems at real time and it also has good robustness, heuristic and transparence. Specially, if there are many different arguments among experts, traditional ES can not deal with them well but NNES can do it.

4. KNOWLEDGE STRUCTURE REPRESENTATION ABOUT THE ES OF RESISTANCE TO CORROSION OF SULFATE ON CONCRETE

4.1 principles of selecting systemic knowledge representable technology

It is important for the performance of intelligent system to choose proper knowledge representation technologies [1]. The current expert system selects knowledge representable technology based on the following principles:

a) Analyze the field knowledge and then choose perfect knowledge representable strategies, which can reflect those unknown information exactly.

b) sually, an expression is fit for a kind of knowledge type. But in fact, many fields are so complicated that it is hard to express with a single format. Therefore, some basic expressions need to be combined [2].

c) Propose a set of knowledge representable strategy, which can express field knowledge effectively and it is different from traditional one.

d) Combine the expert system with other technologies, such as artificial neural network, genetic algorithm and so on, which makes the system become more intelligent [3].

4.2 The systemic strategy of knowledge representation

The systemic strategy of knowledge representation is “rule frame + rule body”, which can be described by BCNF:

Rule groups：=<rule frame><rule body>

Rule frame：='IF' <the set of reason factors> 'THEN' <the set of result factors>

“：=” means “be defined” .

For example, according to intensity grades and types of concrete, the ES chooses appropriate admixture and its quantity. The rule groups are as follows:

RSn: IF the intensity grades and types of concrete
THEN types and quantity of admixture
RB: IF (C7.5<= intensity grades of concrete<C30)
and (types of cement=Portland cement)
THEN types of admixture=second grade fly ash and
quantity of admixture=30
IF (C35<= intensity grades of concrete<C45) and
(types of cement=slag cement)
THEN types of admixture=second grade fly ash +
super fine slag and quantity of second grade fly ash=30
and quantity of super fine slag=10

The example is a rule group, which confirms types and quantities of admixture. RSn means the identifier of rule frame in the group, n is both the number of the rule frame and serial number of the rule group. RB is the identifier of rule body. Between RB and the identifier of next rule frame is its contents. The advantages of this representation are described in the following:

a) Hierarchical structure of representation is clear. The relationship between cause and effect among problems is reflected in the frame of rules, and then the ES can use the relationship to infer unknown problems.

b) Expressional ability is good. It is convenient to express logical, processing and calculating knowledge.

c) Because the same kind of knowledge is collected in the same rule group, redundancy of rules is avoided and it also makes RKB compressed and inferential speed more rapid.

d) Uncertain factors can be calculated in the rule body.

e) Good expressions are provided for neural network and model-based knowledge. [4]

4.3 The expert systemic memory structure of knowledge

The expert system designs a set of suitable memory structure of knowledge on the basis of correlative field knowledge and the experiences of experts. Because admixture has many choices changed with time and places, the conclusion is uncertain. As a result, RKB and data tables are adopted in the system. Its format is following:

Rule (conditions) structure:

Rule Name	Rule	Index

Field “Index” points to a table (conclusion structure table), which contains all results aiming at the rule.

Conclusion (facts) structure:

Fact Name	Fact	Certainty factor of fact

Field “Certainty factor of fact” means the possibility of choosing a fact. Its maximal value is 1, that is, 100%. Under a certain condition, selected fact is regarded as the maximal reliable fact.

4.4 Acquire and discovery knowledge with distributed computing algorithm of NN

Because the knowledge is complicated and uncertain in the ES of resistance to corrosion of sulfate on concrete, NN and its distributed computing algorithm are employed to acquire existed knowledge and infer new knowledge. At the same time, the RKB is updated with those discovered new knowledge to make preparations for next reasoning. Its specific steps are showed in the following:

a) Translate rules in the knowledge base into the structure of NN: At first, initial the network structure is formed with symbol rules. Its steps includes rules re-written, which makes every disjunction only expresses one reason rule and then transforms those rules, which include logical relationships (AND, OR, NOT), into neural cells of the NN. New nodes need to be added for the sake of making the NN keep integrity. At last, fully connected multilayer NN is established by adding the weights of combination, which values are zero. In the application, the NN need to be compressed in order to improve its learning efficiency.

b) Train initial NN with BP learning algorithm and then correct the rules in the NN.

c) Remove redundant nodes and weights of combination with generic algorithm, and then train NN again with BP algorithm to form a more accurate NN.

d) Select discovered knowledge from trained NN. Learning-based method is adopted in the ES of resistance to corrosion of sulfate on concrete. The method considers selecting process as a learning process and its goal is the outputs of NN. Existed knowledge is inputted into NN and the outputs of NN are acquired knowledge or the new knowledge. Using the method, the certainty factor is high and inner structure of NN needn’t be considered, so it is applied widely.

The ES of resistance to corrosion of sulfate on concrete was developed with above the structure of RKB and NN. The production rules were used to express in the RKB and new production rules were generated with NN and its parallel distributed computing algorithm. Finally, new discovered knowledge was put into the RKB.

5. CONCLUSIONS

The paper put forward a new ES establishing method combining RKB with the distributed computing of NN and an ES of resistance to corrosion of sulfate on concrete was developed based on the method. On the one hand, because a large number of data have been stored in the RKB, if rules or application data change later, the ES modifies only items in the database, not program[5]. On the other hand, the structure of RKB benefits for establishing NN model and it is convenient to make distributed computing and mine new rules. In addition, the distributed computing algorithm of NN can not only heighten the ability of knowledge acquisition but also refine RKB, which makes the performance of system improved. Therefore, the study on RKB and NN based ES has important reference value for integration of intelligent system.

6. REFERENCES

[1] Guan jiwen, Liu dayou. Theory of Knowledge Engineering [M]. Changchun: JiLin University Press, 1988

[2] Giarratano J. Theory and Programming of Expert System. Beijing: Machine Industry Press, 2000

[3] Tian shengfeng, Huanghoukuan. Artificial Intelligence and Knowledge Engineering [M]. Beijing: Chinese Railway Press, 1999

[4] Zhong luo, Wei zhihua, Lei ting. Knowledge Acquisition Based on Neural Network [J]. Microcomputer Development, 1997, 7(1): 1-3.

[5] Wu xiqi, Zhong luo, Zhang hongmei. The Structure Design of the Object-Oriented Project Expert System [J]. Microelectronics and Computer, 1994, (5): 39.

Zhong Luo is a Full Professor and a tutor of Doctor, a head of the School of Computer Science and Technology, Wuhan University of Technology, the principal of Graphic & Image Lab and Research Center of Intelligent Technology & Intelligent System, a judge of Nature & Science Fund. He graduated from Wuhan University in 1982 and achieved Doctor’s degree from Wuhan University of Technology in 1995 with specialty of structure. He is one of leaders of “Ten Fifth” Key Project; visited Japan and France as a scholar; was awarded many prizes by government and has published a lot of papers in kernel journals, which can be searched partly by EI/SCI/ISTP. His research interests are in intelligent technology, expert system, neural network, software engineering, artificial intelligent, distributed computing, image & graphics and parallel processing.

Jiang Qiong is a Master and graduated from Wuhan University of Technology in 2002. She majors in computer application technology and participated in the “Ten Fifth” Key Project and developed an ES of resistance to corrosion of sulfate on concrete with others together. Her research interests are in intelligent algorithms, expert system, neural network, parallel distributed computing and image & graphics.

A New Neural Network Models Based for Rule-Based Reasoning *

Xiong Shanqing[1] Pan Hao[2] Zhang Yingjiang[1]
[1] School of Information Engineering Hubei University of Technology
[2] School of Computer Science and Technology Wuhan University of Technology

ABSTRACT

Neural Network models of rule-based reasoning are investigated and it is shown that while such models usually carry out reasoning in exactly the same way as symbolic systems, they have more to offer in terms of commonsense reasoning. CONSYDEER, is proposed to account for common sense reasoning patterns and to remedy the brittleness problem in traditional rule-based system. This work shows reasoning of connection models are not only implementation of their symbolic counterparts, but also computational models of commonsense reasoning.

Keywords: Rule Reasoning; Commonsense Reasoning; Neural Network

1. INTRODUCTION

Rule-Based Reasoning is the most typical symbol of symbolized Artificial Intelligence, whether Neural Network models can effectively deal with symbolized Artificial Intelligence or not depends on how they express relative rule behavior and rule-based reasoning ability. In order to present the rules in connecting model, Human beings have done research on a lot of methods to realize them in different systems, but most of them can be realized directly, and there are no different bet between reasoning ability and symbol system.[1]

Connecting model is very effectively in expressing active、complex and all kinds of commonsense reasoning, especially in rule-based reasoning. A rule-based system is very hard to express data but it supports connecting model of rule-based reasoning to explain data. This paper tries to explain these problems by a shared part between a connecting reasoning system and commonsense reasoning. Firstly, Let us check out the existing of the rule-based reasoning connecting model. Then a new network structure which can solve the problem mentioned above will be built up.

2. RESEARCH BACKGROUND

Early rule-based reasoning has very long history in the application of AI and identification fields. Bocharam(1984) used model unit(called rule) to analyze problem. These rules which consisted of some conditions and conclusions lead to the learning of some complex identification theories. The solving problem approach is built up on this basis.

As the appearance of Neural Network, human beings began doing research on how to realize rule reasoning in network environment. Touretzy(1985) put forward a system structure beyond based rule. Which divides memory、rules and facts into basis unit. In order to store relative regular data and decide which rule will be executed, it can get a satisfied result by competition.

Although Neural Network has the ability to realize rule-based reasoning, Can it explain commonsense reasoning better through its models? What is the needed condition which a rule-based connecting model can make use of by imitating commonsense reasoning effectively. We will solve the problem by building up a new Neural Network model through analyzing agreement data and then computing a structure which is similar as commonsense reasoning.

3. BRITTLENESS AND ITS BASIC REASONING MODEL

Allan[2] has collected some protocol of common sense reasoning. He has pointed that it is irrelevant to explain those typical reasoning by traditional logic. For example, somebody asked if South Africa is an ox-production country? Answer was that it is a country just like England in West Europe. To some extend, a guess that it is an ox-production country can be gotten. An uncertain conclusion is based on a kind of similar known knowledge because that there is no known knowledge. Examples like this have harassed research on symbolized artificial intelligence which is called brittleness for a long time. In general, it contains the things as follows: (1) partial information;(2) uncertain and dim information;(3) unmatched rules;(4) lacking coherence and integrality among rules;(5) universality;(6) inheriting from bottom up and from top down; (7) importing new rules.

The analyzed results show that all traits of the above problem can be gotten by reasoning when the relevant rules that sustain the reasoning are endowed even though the above contents looks like a group of independent questions. A similar concept is defined as follows:

$$(A\sim B) = \left|\frac{F_A \cap F_B}{F_A}\right| \in [-1,1]$$

If $ACT_A=a$, $ACT_B=a*(A\sim B)$. Here F_i is a characterful expression of dot i, ACT_i is the activation of dot i. Knowledge linking value is given by:

$$(A\rightarrow B) = \gamma \in [-1,\ 1]$$

That is $ACT_B=a*(A\rightarrow B)$ if $ACT_A=a$. And γ is the knowledge link strength between A and B.

What’s more, unmatched case can be described as follows:

$$A\sim B \qquad A\rightarrow C$$

A is activated $(ACT_A\neq 0)$,thus we can get

$$ACT_C=(B\sim C)*ACT_B=(B\sim C)*ACT_A*(A\sim B)$$

Other cases can be described analogously. The problem about brittleness can be solved entirely in CONSYDERR structure.

* Funded by national “Fifteenth –year plan” key project （2001BA307B01-02-01）

4. COMPENDIA OF MODEL

The CONSYDERR structure is composed by two layers: CL and CD.

CL is conjunctive network that conduct cursory reasoning in concept level. The rules are embodied by the linkages among nodes. CL can contain traditional reasoning based on rules totally and catch knowledge of causality between commonsense. It also can conduct reasoning about accumulated partial information and uncertain information. The basic operation of this scheme is a simple add-weight calculation. And the accumulating operation of node add-weight can be realized easily in conjunctional networks.

CD is a distributed denotation that equals to conjunctional network of reasoning in sub-concept level. Concepts and rules are described dispersedly by some groups of units which are overlapped each other. It has characteristic perceptive unit and inherent target and external state in this unit. Among these original descriptions, concepts are defined by comparability between themselves and other concepts. For the similarity description of concept in a higher lever, we use these geometric primitives as sub systems.

Now we can use these distributed networks to link every “close ” of every concept to all the dots describe the same concept in CD, which means they have been linked to local network. The linkage of nodes in net is conducted by a crossbar-like structure. The rules that linked to CD has been duplicated approximately. The interaction of two net components is a fixed recycled process.

First, the crossbar-like switch is opened to let the activated unit of “close” in CL flow into the dot in CD. Then a temporary stable phase of two parts is gotten .At last, the crossbar-like switch let the activated unit in CD flow back to CL to combine the same dots in CL.

This system is a unity based on rules and similar components whose functions are all defined on similarity of knowledge and concept. This relationship is helpful to conquer the above problem of brittleness.

Let us look at the specific example about the reasoning of ox-production country. This example can be described as follows:

South Africa ~ England, England~ ox- production country

By setting ACT $_{\text{South Africa}}$=1, the conclusion of ox-production country can be gotten by calculation as follows:

ACT $_{\text{ox- production country}}$= （England → ox- production country） * ACT$_{\text{South Africa}}$ * （South Africa~England）

South Africa~ England =P * $\left| \dfrac{F_{SouthAfrica} \text{ I } F_{\text{England}}}{F_{SouthAfrica}} \right|$

P∈[0,1], is a parameter used to adjust the action of system. The above formula can be transformed to CONSYDERR structure directly.

5. SYSTEM WORKING PROCESS

After system received input data, CONSYDERR architecture is operated as fixed cycle:

(1) Top-down step;
(2) Processing step;
(3) Bottom-up step.

During top-down step, calculation is given by:

$$x_i(t+1) = \max ACT_a(t)$$

a is a arbitrary node in CL, $x_i \in CD_a$.During processing step, calculation is given by :

$$\Delta ACT_a = a\sum W_i I_i(t) - \beta ACT_a(t) \text{ ;}$$

$$\Delta x_i = \mu \sum W_i I_i(t) - U x_i(t)$$

where W_i is sustain link weight, I_i is an activated eigenvalue.

During bottom-up step, calculation is given by:

$$ACT_b(t+1) = \max(ACT_b(t), \sum_{x \in CD_b} \frac{x_i(t)}{|CD_b|})$$

This system can be used to the above example. During top-down phase, nodes represent South Africa and England in CD will be activated by system based on similarity. During processing phase, the law of linkage will work. It conducts the reasoning as follows: England is an ox- production country, so the nodes represent ox-production country in CD will be activated. At last, during bottom-up phase, nodes of ox-production country which have been activated in CD will activate nodes of ox-production country in CL. The conclusion can be gotten from CL directly which is given in Fig 1.

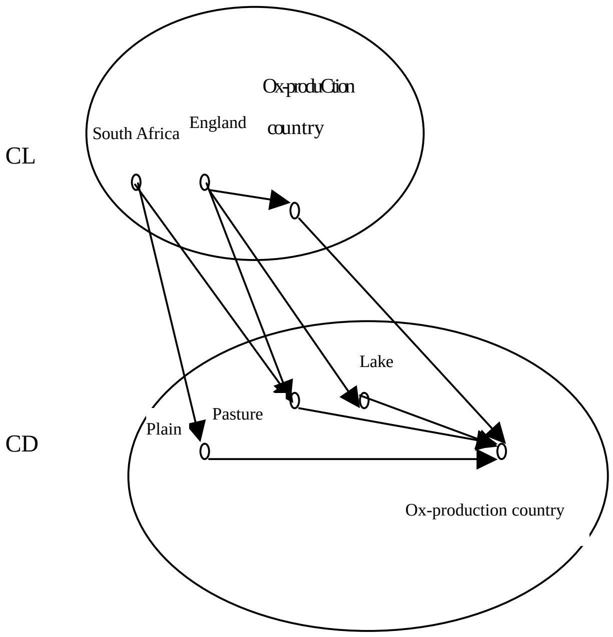


Fig. 1 System work process example

6. CONCLUSION

Compared with other systems, the neural networks model based on rule reasoning can utilize the property of data in a parallel way. The CL part is similar with other link architecture based on rules. The specialty of CD part is that it provided another effective way to satisfy the reasoning based on rules though similarity and it conquers the brittleness in typical signalized system .Thus it is helpful to process the problems such as partial information ,uncertain match、interaction of reasonable inherited rules.

7. REFERENCES

[1] JiaoLicheng. Theory on neural networks system.Xi’an:XiDian University Press ,1997.
[2] HuShouren. Neural networks [M].Beijing: Press of National University of Defense Technoloty,1997.
[3] Simpson P.K., Artificial Neural System, New York Pergamon Process,Inc,1998
[4] Kosko B. , Neural Network and Fuzzy Systems. New Jersey: Prentice Hall. Inc.1996

Study the Application of Neural Network in the Prediction of Regional Integrated Transport Structure

Huiyuan Jiang, Jiang Li
Transportation College, Wuhan University of Science and Technology
Wuhan, Hubei Province, 430063, China
Email: p930212h@public.wh.hb.cn **Tel.:** +86 (27)86581330

ABSTRACT

Each transport method in the transport system has its own technical economic characteristics and application range. However, in the past, each transport method was usually investigated separately in the quantitative analysis of transportation development trend. It obviously has some shortages. And it is always a technological problem to analyze the development trend of the comprehensive transportation structure with mutual effect of each transport method. The presence of the artificial neural network provides us a new way to resolve this problem. Although the BP network is suitable for the short-term estimate, after its self-adapting in algorithm and error correction it can also be suitable for the long-term estimate. This article establishes a neural network model according to the principle of artificial neural network, for the structure prediction of the comprehensive transport freight transportation volume and the freight transportation turnover. We have displayed the running result of the model which has been proved to conform to the natural law of transport economic development with an example of the comprehensive freight transport system in a Province. Therefore, the BP network is more suitable to the multi-factor complicated transport system, which is unable to be expressed or analyzed by overt formulas.

Keywords: Neural Network, Integrated Traffic, Cargo Structure, Forecast.

1. INTRODUCTION

Carrying trade is the vein of economic society. Even since the production and the exchange of products came into being in the society, carrying trade, which severs for the circulation, has been born. Only by the transportation can the material products be exchanged. Without the transportation, the exploitations and sale activity of resources in the range of the whole world and even in each country's interior can't be realized, and the social productivity can't be developed thoroughly. The forming and development of carrying trade push the development of social productivity powerfully. While the social productivity develops greatly and the degree of social product becomes higher and higher, carrying trade certainly comes to face the higher request.

The transport supply system includes railway, highway, shipping, airline and pipe. Every kind of transport method in the transport system has its own economic characteristics in technique and orientation extent. Since in different period the industrial structure is different, the development of every kind of transport method and the structure of comprehensive transport has different characteristics. Right full with confidence of the trend of every kind of transport method and the structure of comprehensive transport system is the antecedent of establishing the right planning of transportation development.

Each kind of transport method may be replaced by other kind, so every kind of transport method suffers the influence from the other greatly. In the past, every kind of transport method was usually investigated separately in the quantitative analysis of transportation development trend. This obviously has some shortages. It is always a hard nut to crack in technique how to quantitatively analyze the mutual effect of every kind of transport method to analyze the development trend of comprehensive transportation structure. But the presence of the artificial neural network provides us a good way to resolve this problem.

2. BASIC PRINCIPLE IN THE BP NEURAL NETWORK [1]

Artificial neural network (ANN) is simplification and simulation of the biological nervous system, it has been one of the methods of using artificial intelligence in technique, economy, management and some other aspects. The model is just a network formed by a large amount of interconnected neurons. According to the connection mode, the neural network usually is divided into two main types: the neural network of forward direction of no feedback and joining together neural network. ANN theory is still currently in the development step. More than tens of kinds of models have been studied, but the three most popular kinds of models are forward direction neural network, feedback neural network and self-organizing neural network.

The BP neural network is one of forward direction neural network models. It is a multi-layer forward direction network that unidirectional spreads, and it's structure is shown in Fig1.

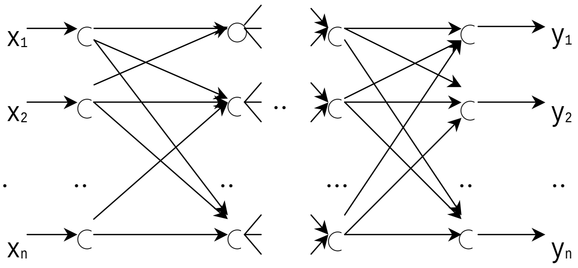


Fig1. BP neural network

In addition to the input and output panel points, the network has implicit panel points of a singer layer or several layers, and there is no link among the same layer panel points. Input data x^1, x^2… x^n, from the input layer panel points through each implicit layer panel points one by one to the output panel points. At last we get the output data y^1, y^2… y^m. Because the panel points at the same layer have none coupling, the output of points only affects the output of the panel points at the next layer. Each point indicates a single neuron, and its homologous transferring function is the function of Sigmoid. Sometimes the transferring function of the panel points in the output layer also follows the linear function.

The BP neural network substantially reflects the highly nonlinearly mapping between the input and output, namely f: $R^n \rightarrow R^m$, f(x)=Y. And there has been the theorem which proves that an arbitrary continuous function can be realized by a treble BP neural network. Thus, for the assembly whose amount of sample is T: inputting $x^k \in R^n$, and outputting $Y^k \in R^m$, t=1,2,…,T. The mapping relationship between the input and the output can be reflected in a certain precision by creating a treble BP neural network model.

3. THE BP NEURAL NETWORK MODEL FOR THE PREDICTION OF THE STRUCTURE OF THE REGIONAL COMPREHENSIVE TRANSPORT SYSTEM [2]

One province is one of the richest provinces and whose economy increases the fastest. In this province, the development degree of the transportation system is also in the leading position of our country. In its integrated transport network, railway, highway and inland river undertook more than 90% of the freight transportation volume and turnover. The development and the turning of the position of these kinds of transport methods have determined the trend of the province's integrated transport system development.

According to the investigation, it is the level of the economy development, the structures of three industries, the routes and mileage of transport method and some other factors that greatly affect the volume of each transport method and the structure of comprehensive transport. Therefore, we set up a BP neural network for 8×50×8 by using data from 1985 to 2002. X={GNP (x^1), the mileage of railway(x^2), highway mileage(x^3), inland river rank channel mileage(x^4), oil pipeline track mileage(x^5), the first industry ratio(x^6), the second industry ratio(x^7), the third industry ratio(x^8)} are set as input, and Y={railway freight transportation turnover(y^1), highway freight transportation turnover(y^2), inland river freight transportation turnover(y^3), pipe freight transportation turnover(y^4), railway freight transportation volume(y^5), highway freight transportation volume(y^6), inland river freight transportation volume(y^7), pipe freight transportation volume(y^8)} are set as output. In this model, V^{ij}(i=1,2,…8;j=1,2,…50) indicate the weight between the input layer neuron and the implicit layer neuron, and W^{jk}(j=1,2,…50; k= 1,2,…8) indicate the weight between the implicit layer neuron and the output layer neuron which reflect the homologous transportation volume and transport structure in the condition of different economy development situation, industry structure and transport mode development.

In the algorithm of the BP neural network, we usher in the time factor which is donated by t (T==1, 2,…, S, S is the amount of time sequence sample=18) and the predicted amount which is donated by F. Firstly to each t, we practice the BP network as above algorithm principle getting the N×L sequences (V^{ij}) t and L×M sequences (W^{jk}) t (t=1, 2,…, s). In consideration of the long period forecast, for the higher precision, we must study and grasp the regulation of change and trend of $V^{ij}$and W^{jk}. So we can adapt a certain trend extrapolate (such as two times smooth method in index), and make the prediction of every (V^{ij}) t sequence and (W^{jk})t sequence in the period of No.(S add F). This procedure has created BP neural network forecasting model for the period of N0.S+F. To each input variable of the model, we can adapt some prediction method separately (such as increasing the flat out method) to predict the period of No. S adds P, and at last we can make use of this model to make prediction of all output variables in the period of No. (S+F). Model algorithm is shown in the figure 2.

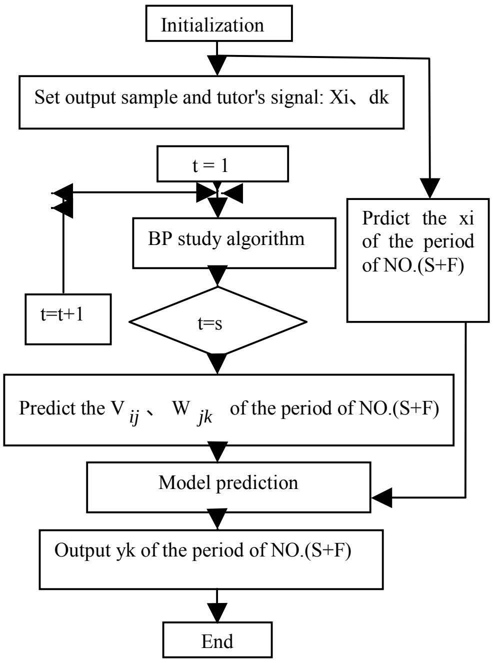

Initialization
Set output sample and tutor's signal: Xi、dk
t = 1
BP study algorithm
Prdict the xi of the period of NO.(S+F)
t=t+1
t=s
Predict the V ij 、 W jk of the period of NO.(S+F)
Model prediction
Output yk of the period of NO.(S+F)
End

Fig2 BP neural network forecast algorithm

To check the prediction precision of the network model we take the S=13, F=1~ 5. We use the real data from 1985 to 1997 as the input sample and tutor's signal (desire output), and we use the real data of the reason variable from 1998 to 2002 as the input of the model. And then we get the output of predicting the result variable of the past five years. The average error between the network model output value and the real value is smaller than 5%.The precision of the forecast is very high .These show that the model has simulated the mapping well between the input and output, and the result is credible. So it can be put into practice.

4. THE PREDICTION OF THE STRUCTURE OF THE REGION COMPREHENSIVE TRANSPORT VOLUME AND TURNOVER [3]

According to plan of the development of the economy and comprehensive transportation, we adopt the BP net work model to forecast the freight transportation volume and turnover of railroad, highway, inland river and pipe of 2005 and 2010 .The result is shown in the table 3-1Table 3-1.

Table 3-1 The forecast of regional integrated transport structure by neural networkUnit: Hundred million ton kilometer, Ten thousand ton

Year	Railroad freight transportation turnover	Highway freight transportation turnover	Inland river freight transportation turnover	Pipe freight transportation turnover	Railroad goods volume	Highway goods volume	Inland river goods volume	Pipe goods volume
2005	450	400	550	80	6000	70000	24000	2000
2010	500	550	700	100	7000	95000	30000	2500

5. THE ANALYSIS OF THE FORECAST RESULT

The forecast result shows that with the continuously increasing of GNP and the improvement of transport network condition, the province’s amount of integrated freight transportation volume and turnover rise continuously. Railroad and inland river are still main freight transportation way of long distance. And the ratio of highway rises continuously.

The province’s national economy remains increasing rapidly in the passed twenty years. In the aspect of industrial structure adjusting, the ratio of the first industry is descending continuously, and the second industry stands still, while the third develops greatly. But because the economy is still in the early development in industrialization step, the ration of the third industry can’t be in the leading position very soon.

Although the amount of freight transportation of the unit production value have step-down, and because the ration of the second industry remains in the leading position in a long period of time ,the freight transportation will still increase constantly in a certain period ,but the speed of increasing will slow down gradually .As for the structure of product ,it turns out to be short ,small, thin ,small amount ,most species and of higher addition. As for the transportation structure, though the ratio of mass freight transportation which mainly is the primary product descends a little, it still remains the main goods. At the same time ,the development speed of the processing industry and the third industry exceed that of the traditional heavy industries .On one hand, the processing industry and the third industry have the stronger acceptance capability to the transport price ,on the other hand ,the more rapid and vivid transport mode is needed to match for meeting the request of fastening the speed of capital .Therefore ,as for the freight transportation structure ,the railway and inland river which are suitable for the mass freight transportation are still the main mode .At the same time, the ratio of highway increases continuously, becoming the combining part of the big path freight transportation. This kind of developing characteristics of the province’s comprehensive transportation system is normal and reasonable. It matches the expectation of today’s economic theory in transportation and also fundamentally matches the general feature of the change of the world’s comprehensive transportation system.

6. CONCERNING THE CONCLUSIONS OF THE BP NEURAL NETWORK APPLICATION

1) Since the BP network can stand for a high nonlinearity mapping system with arbitrarily precision, it can be either a system of single input and single output or a system of much input and much output .As for the quantitative analysis of the complex system which is of many factors or can’t be expressed by the typical type, BP network has the apparent advantages .And the most of actual problems are the complex nonlinearity system, so the BP neural network is suitable in a large range.

2) The neural network has the intelligent learning functions, and it can adjust itself continuously according to the actual situation. This is the function which the ordinary quantitative analysis systems can hardly have.

3) It is very easy to set up the BP network model which does not need any idealization premise hypothesis. And the machine operation is convenient while the margin of error is low.

4) Although the BP network is suitable for the short-term estimate, it can also be suitable for the long-term estimate after its self-adapting in algorithm and error correction.

7. REFERENCES

[1] Shitong Wang et al *Neural Fuzzy Network System and its application.*. Beijing :Beijing University of aeronautics and astronautics Press，1998. P37~P48

[2] Cenreng Yuan et al *Artificial neuron network and application*. Beijing, Tsinghua University Press .1999. P95~P108

[3] Traffic statistical almanac of Jiangsu Province from 1985 to 2002.

Huiyuan Jiang: an associate professor, doctor, major in international shipment and integrated logistic

Jiang Li: a postgraduate, major in international shipment and integrated logistic

Using HTABP Algorithm to Determine Number of Hidden Units in NN *

Zhang Xi[1], Pan Hao[1], Xiong Shanqing[2]
[1] School of Computer Science and Technology, Wuhan University of Technology
[2] School of Information Engineering, Hubei University of Technology

ABSTRACT

Based on Heuristic Terminal Attractor Back Propagation (HTABP) a new algorithm is put forward to determine the appropriate number of hidden units in a multilayer feed forward neural network. The algorithm is decided by the time-varying gain γ_c of HTABP and the normalized error function instead of the selection by trait and error.Simulation results show that the algorithm is scientific and effective.

Keywords: Neural Network; Hidden Unit; self Growing Learning

1. INTRODUCTION

It has been proved that a network with as few as one hidden layer and appropriate hidden units is capable of arbitrarily accurate approximation to any real-valued continuous functions over a compact set. However, there is hardly a rule or method to determine the appropriate number of hidden units so far. People usually determine the appropriate number of hidden units by trial and error. As a result, it may take a long time to get a network with proper size.

In recent years, several researchers have tried to solve the problem of determining the optimal number of hidden units. For example, Tenorio and Lee [1] introduced a self-organizing neural network structure which can construct the network itself for system identification application. Fu [2] discussed the effect of the number of hidden units and input units by experimental results. In this paper, we introduce a new algorithm called "self learning Algorithm" to determine the appropriate number of hidden units. This algorithm are developed from the HTABP Algorithm [3], which can finish learning process in finite time reach the global minimum of error function and guarantee to converge faster than the BP algorithm. We propose a time-varying factor together with the total error function for making the decision of adding or deleting a hidden unit. The time-varying factor used is important to the HTABP and can reflect the status of learning process. .

2. HEURISTIC TERMINAL ATTRACTOR BACK PROPAGATION

2.1 TERMINAL ATTRACTOR

The terminal attractors are equilibrium points of a dynamic system where the Lipschitz condition is violated. The Lipschitz condition guarantees the existence of a unique solution for each of initial conditions. Hence, a solution evolving from an initial condition can not intersect the corresponding equilibrium point, and therefore, the time of approaching the equilibrium points is always infinite. If the Lipschitz condition is violated, the fixed point becomes a singular solution which is interested by all the attracted transients. Dynamic systems with terminal attractors can reach terminal attractors in finite time.

2.2 HTABP

The HTABP algorithm is a new learning algorithm for multilayer neural networks based on the concept of terminal attractor and the back propagation algorithm. The key point is the introducing of time varying gains γ_c to the weight update law of BP algorithm. The concept of the gains γ_c is that it will depend on the total error function E and the derivative of E with respect to weights such that a terminal attractor for energy transients is formed at the equilibrium point of zero, we briefly explain the key concepts.

Define the error function E to be

$$E = \frac{1}{2}\sum_{m}\sum_{l}\left(\overline{V_l^m} - V_l^m\right)^2 \tag{1}$$

Where $\overline{V_l^m}$ and V_l^m are the target output and the actual output for the lth neuron at output layer corresponding to the mth pattern. Assume there are p processing elements used for implementing the neural network. The time evolution of weights is now given as

$$\frac{dT^c}{dt} = -\gamma_c \nabla F^c E \qquad (c = 1,2,\Lambda, p) \tag{2}$$

Where T^c means the weights vector governed by processing element c and γ_c is the adaptive gain of processing element c . γ_c Can be formulated by

$$\gamma_c = E^k \big/ P // \nabla T^c E //^2 \quad \left(c = 1,2,\Lambda, P, 0 < k < 1\right) \tag{3}$$

Note that if $\gamma_c = 1$, the weight update law degenerates to be the BP algorithm.

The HTABP algorithm is presented as the following:

- ●Define E to be normalized error function
- ●Assume there are P processors
- ●Initialize all variables: weights, input pattern and targets.

Input the parameter Error tolerance.

```
Max pass. Pass
While ( $E$ > Error tolerance and pass < max pass) do
  {Calculate forward to find the network outputs $V_l^m$
    Calculate backward to find $\delta_j^m$ and $\gamma_c$ , $c = 1,2,3,\Lambda, p$
    If $\gamma_c > 1$ update weights by
    $T_{ji}^c(n+1) = T_{ji}^c(n) + \mu\gamma_c\delta_j^m V_i^m(n)$
    Else update weights by
  $T_{ji}^c(n+1) = T_{ji}^c(n) + \mu\delta_j^m V_i^m(n) + \alpha\Delta T_{ji}^c(n-1)$
      Pass=pass+1}
```

* Funded by national "Fifteenth –year plan" key project （2001BA307B01-02-01）

3. SELF LEARNING ALGORITHM

To simplify the problem, a network with one hidden layer is considered. The purpose of the algorithm is varying the number of hidden units such that the network can learn to approximate a given function with the implementation cost as least as possible. The flow chart of the algorithm is shown in Fig.1.

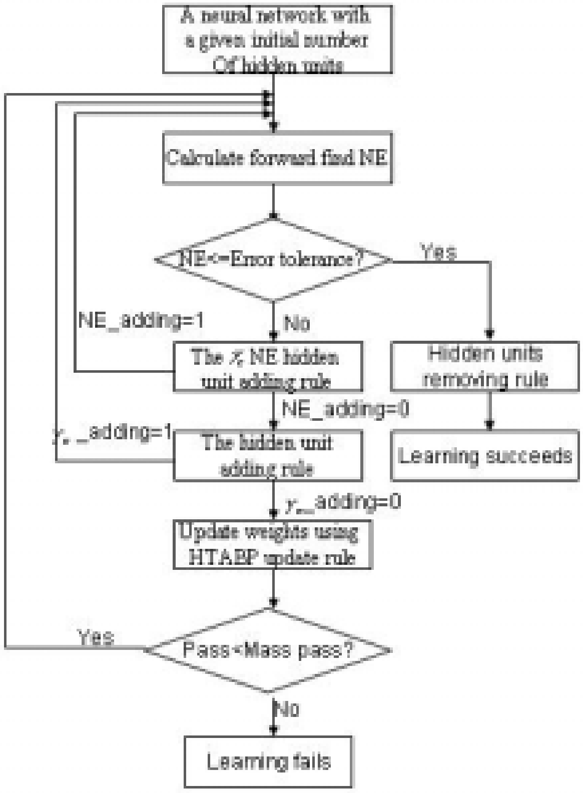

A neural network with a given initial number Of hidden units
Calculate forward find NE
NE<=Error tolerance?
Yes
No
NE_adding=1
The γ_c NE hidden unit adding rule
Hidden units removing rule
NE_adding=0
γ_c _adding=1
The hidden unit adding rule
Learning succeeds
γ_c_adding=0
Update weights using HTABP update rule
Yes
Pass<Mass pass?
No
Learning fails

Fig.1 The flow chart of the SLA

3.1 THE γ_c HIDDEN UNIT ADDING RULE

A very large value of γ_c or a surge of the gains γ_c over the threshold means the trajectory of error transients comes into a flat region of the error surface. It may still can be tolerated and the learning can be either continued find the optimal set id weights or restarted by adding a new hidden units to escape from current embarrassing situation. Since the actions are not distinguishable by γ_c, one way to choose which actions can be resort to a random process. We define the parameter prob to be $1 - e^{-E/Temp}$. Prob is the probability to add a new node to the network which will decay as E decreases but will not decay too fast because Temp becomes small during the learning process. The γ_c hidden unit adding rule can be expressed as follows:

```
Input: $\gamma_c$, $N$, Minerr
        Output: $\gamma_c$_adding d = 1   (add a new hidden unit)
                $\gamma_c$_adding d = 0   (do not add a new hidden unit)
        If ( $\gamma_c$ > Threshold)
          { $prob = 1 - \exp^{(-N/Temp)}$
              Temp = 2*Minerr
              if (prob < random( ))
{ raise Threshold by appropriate quantity
                $\gamma_c$_adding =0}
else $\gamma_c$_adding =1 }
          else $\gamma_c$_adding =0
```

Fig.2 The γ_c hidden unit adding rule

3.2 THE E HIDDEN UNIT ADDING RULE

If there is a deficit in hidden units, the network can not realize a desired function. In this case, the system governed by HTABP update law may set into a situation that the normalized error is bounded and satisfied N > Error tolerance and oscillates around the equilibrium point or situation for a long time. We consider that E will not reduce any more. Parameter count is proposed to characterize this situation. As count exceed a given value max count, it is reasonable to conclude that learning process can not reduce error function any more, so we add a new node and restart learning.

3.3 THE HIDDLE UNIT REMOVING RULE

As soon as the algorithm finds the optimal set of weights, we save the set of weights in a temporary location and then remove the last added node and its corresponding weights. The learning process is continued and the N hidden unit adding rule is applied to make decision, if the deleted node should be recovered: if the learning process can still converge, E > Error tolerance, we remove one node again: otherwise, we recover the previous set of weights from the temporary location and stop the algorithm.

4. SIMULATION RESULTS

In order to comparing the self learning algorithm with BP algorithm, we have done four experiments totally. (Results are shown in Fig3). In the fourth experiment, seventy input and output modes have been trained by self learning algorithm and traditional BP algorithm separately. As a result, only we try 462 times and use 6-14-3 network can we successfully achieve training.we find it takes longer time to use traditional BP algorithm and change hidden units artificially. However, we use self learning algorithm and achieve the same result as 6-11-3 network by 200 times. The time decreases largely comparing with that of using traditional BP algorithm.

Example	Initial Temp	Initial Threshold	Max Count	Error tolerance	Initial number of hidden node	Final number of hidden node
(i)	0.03	9000	2000	0.005	1	11
(ii)	0.03	9000	1500	0.005	1	4
(iii)	0.03	9000	2000	0.005	1	4
(iv)	0.03	9000	2000	0.005	1	3

Fig.3 simulation results

5. CONCLUSION

In this paper, we develop a new algorithm. This algorithm can adjust the number of hidden units autHTABP and the normalized minimum error function. The simulation results show that this algorithm is effective in find the proper number of hidden units with the algorithm; we need not to estimate the number of hidden units by trial and error any more. We believe that the algorithms are helpful in determining a proper size of network automatically. The criterions to increase hidden units are the time-varying gain γ_c of.

6. REFERENCES

[1] M.F. Tenorio and W.T. Lec. “Self-organizing Network for optimum super vised learning”. IEEE Trans on Neural network. Vol.1 No.1. March.1997.

[2] L.M.Fu. “Analysis of the Dimensionality of Neural Network for pattern Recognition”. Pattern Recognition. Vol.23 No.10 pp 1131〜1140. 1996.

[3] Sheng De Wang and Ching hao Hsu. “A self Growing learning Algorithm for Determining the Appropriate Number of Hidden Units”. Neural Network. Vol.2 pp 183〜192. 1991.

[4] Jian Li cheng. “Neural Network system Theory”. Xi´an Electron Technology University Publishing House, 1990.

[5] Pan Hao, Zhong Luo. A New Neural Network Model and Its synthesized algorithm. Journal of HuBei Technology Institute, Vol.3 pp 52〜55, 1999.

A Study on Neural Network Based on Contractive Mapping Genetic Algorithm

Zou Chengming, Tong Qiwei, Yang Hongyun, Yuan Jingling
The Computer Science &Technology Department, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: zoucm@mail.whut.edu.cn **Tel:** 027-87211983

ABSTRACT

BP algorithm for multilayer feedforward neural network depends on value of weights of neural network, which will not converge or converge to local minimum. A new method employing contractive mapping genetic algorithm to learning of multilayer feedforward neural network was proposed in this paper. The algorithm provides an effective method of global convergence of BP network by introducing fixed point, which not depends on the value of weight.

Keywords: Neural Network; Genetic Algorithm; Contractive Mapping; Fixed Point

1. DEFINITIONS AND CONCEPTS

Definition 1: X is supposed to be a non-empty aggregate, If d(x,y) is a real number in which parameter x and y are picked up from X， and content the conditions below:

Non-negative: d(x,y)≥0，if and only if x=y, d(x,y)=0
Symmetry: d(x,y)=d(y,x)
Triangle inequality d(x,y)≤d(x,z)+d(z,y)， z∈X
then d(x,y) is called measurement or distance of x and y，and X is called measure mental space or distant space of d as (X,d) or X.

Definition 2: X is supposed to be a real or complex linear space. If ‖x‖ is a real number in which parameter x is picked up from X， and content the conditions below:

(1) ‖x‖≥0，if and only if x=0, ‖x‖=0;
(2) ‖αx‖=|α|‖x‖， in which α is a real number;
(3) ‖x+y‖≤‖x‖+‖y‖， x,y∈X.
then ‖x‖ is called formatted number of x，and X is called linear space or space for short.

Definition 3: (X,d) is supposed to be a measurement space, spot line $\{x_n\} \subset X$.{xn} is called Cauchy spot line or basic line, in the condition that given ε >0,non-negative N exist if m,n>N then d(xm,xn)< ε. measurement space X is called self-contained measurement space if every Cauchy spot line converge.

Definition 4: self-contained linear space is called Banach space by the means of measurement is abducted by formatted number in linear space, viz d(x,y)= ‖x-y‖.

Definition 5: X is supposed to be a non-empty aggregate in mapping $T: X \to X$. x* is called immobile point of mapping T in the condition that $x^* \in X$, then $Tx^* = x^*$.

Definition 6: X is supposed to be a linear space in mapping $T: X \to X$ .T is called contractive mapping of mapping T in the condition that x,y∈X，then $\|Tx - Ty\| \le \alpha \|x - y\|$ in which 0< α <1.

Theorem 1(Banach immobile point theory)There is only one fixed point during contractive mapping in the Banach space.
Contractive mapping genetic neural network BP(CMGANN)algorithm

BP algorithm for multilayer feedforward neural network depends on value of weight of neural network. Convergence speed and precision of network can be obtained if the appropriate value of weight is chosen, or the network will not converge. Furthermore, the convergence speed of BP algorithm is very low, proper result can’t be obtained even after thousands of iterative calculation. Therefore, method to improve convergence speed and precision should be developed.

As shown below, a non-linear mapping F:Rn→Rm is applied in BP neural network, in which (xi, yi)i=1,2…,N are supposed to be samples of neural network learning, only one hedding layer is supposed in the network(more hiding layer are acceptable), the number of hiding cell is N, and inspiriting function is f. Then the mapping of the network is [1~4]:

$$y = \sum_{i=1}^{N} c_i f(w_i * x - \theta_i) \qquad \text{Eq(1)}$$

$w = (w_1, w_2, ..., w_N)$ denote the weight between input layer and hiding layer， $c = (c_1, c_2, ... c_N)$ denote the weight of hiding layer between output layer， $\theta = (\theta_1, \theta_2, ... \theta_N)$ denote hiding cell thresh value. The target of network training is:

$$\min E(\omega) = \sum_{j=1}^{M} | y_j - \sum_{i=1}^{N} c_i f(w_i x_j - \theta_i) |^2 \qquad \text{Eq(2)}$$

As shown above, Eq (2) is a extremely non-linear mapping, so the network probably can not converge by general training method after a long time or converge to local minimum. Therefore, contractive mapping genetic neural network BP(CMGANN)algorithm is provided to get the solution of Eq(2).

X is supposed to be the space of group P[5~6]，in which the group size is fixed. Each group is made up of n units, viz $P = \{p_1, p_2, ..., p_n\}$ ， and each unit is made up of weight and thresh value ,viz $p_i = (c_i, w_i, \theta_i)$ ， thus training of neural network can be performed by get minE(p). E(p) is taken as estimate function eval(p).If eval(pi)<eval(pj)，unit pi is better

than unit pj. $Eval(P)=\frac{1}{n}\sum_{i=1}^{n} eval(p_i)$ is the estimate function of group P. Obviously, eval(p)≥0，Eval(P)≥0.

Measurement is defined as below:

$$d(P_1,P_2)=\|P_1-P_2\|=\begin{cases}0 & P_1=P_2\\ Eval(P_1)+Eval(P_2) & P_1\neq P_2\end{cases}$$

(X,d)is a measure mental space，for the reason that it content the situations below:

(1)Given group P1 and P2，d(P1,P2)≥0，if and only if P1=P2，d(P1,P2)=0;
(2)d(P1,P2)=d(P2,P1);
(3)d(P1,P2)+d(P2,P3)=Eval(P1)+Eval(P2)+Eval(P2)+Eval(P3)≥Eval(P1)+Eval(P3)=d(P1,P3).

Moreover, a measurement space (X,d) is self-contained, since there are only finite units for each group, and every Cauchy sequence $P_i(i\to\infty)$ converge. Parameter k exists for every Cauchy sequence so that, given n>k, Pn=Pk. The measurement space (X,d) is Banach space.

Mapping $T:X\to X$ can be defined as T(P(t))=P(t+1) which is a simple iterative in genetic algorithm. Since group P(t) is more advanced than group P(t+1)(situation without improvement is not taken into account),viz Eval(P(t))>Eval(P(t+1))=Eval(T(P(t))).Therefore,

$$\begin{aligned}&\|T(P_1(t))-T(P_2(t))\|=Eval\,(T(P_1(t)))+Eval\,(T(P_2(t)))\\ &<Eval\,(P_1(t))+Eval\,(P_2(t))=\|P_1(t)-P_2(t)\|\end{aligned}$$

and

$$\|T(P_1(t))-T(P_2(t))\|\le\alpha\|P_1(t)-P_2(t)\|,\alpha\in[0,1)\,.$$

So mapping T is contractive mapping. As shown in theorem 1, P* is the only fixed point. Obviously, $P^*=\lim_{i\to\infty}T^i(P(0))\cdot$ P* is the only fixed point in the group space which is searched by contractive genetic algorithm. P* has no relationship with initial group P(0).Conclusion that fixed point P* can be obtained when every unit has the same global smallest value can be drew from the definition of Eval (P).

So no matter how much initial weight value of w, c and thresh value θ are ,a global optimized solution can be obtained by this algorithm.

2. SIMULATIVE CALCULATION

Neural network is used to approach the function below to approve the validity of the algorithm:

$$\begin{aligned}f(x,y)&=-20\times\exp\,(-0.2\times\sqrt{(x^2+y^2)/2})\\ &-\exp\,((\cos\,(2\pi x)+\cos\,(2\pi y))/\,2)\end{aligned}$$

If x=0, y=0，f(x,y)=－22.71282 is the minimum of the function. A 3 layer neural network with 5 the hiding neural is built to approach f(x,y).2 neural is built as the input layer，and 1 neural is built as the output layer. BP algorithm and Contractive mapping genetic algorithm which provide in this paper are applied separately to train the neural network.

Floating-point coding of chromosome don’t have problems such as the limit of weight and precision which generated by binary system coding. The crossing method taken as the arithmetic crossing via v1, v2 is supposed to be 2 chromosomes for crossing, then we can obtain v’1= v1+(1-a) v2, v’2= v2+(1-a) v1，in which a∈(0,1). Heterogeneous aberrance is applied, viz given father v, and if element xk is chosen to aberrance, then the offspring are v’=[x1, …,x’k,…xn], x’k= xk+(xUk- xk)*r*(1-t/T)b and x’k= xk－(xk- xLk)*r*(1-t/T)b are expressions to calculate x’k which is chosen in a random way. xUk and xLk denote the maximum and minimum value of xk. r denote the random real number in [0, 1]. T denotes the maximum algebra，and t denotes current algebra b denote the certain heterogeneous parameter. The size of stirp group is chosen as 10，the maximum algebra as 1000，aberrance rate as 0.1，crossing rate as 0.3.

The error of neural network is less than 0.1 after 1000 iterative by Contractive genetic algorithm and f(0,0)=－22.718276.On the other hand, precision is not meet after 1000 iterative by BP algorithm and f(0,0)=－18.53422.As showed in the result of training by BP algorithm, the network converges to local minimum.

3. CONCLUSION

The algorithm provided in this paper is proved to be valid by the simulative calculation above. CMGANN algorithm provides an effective method of global convergence of multilayer feed forward neural network, and the iterative process is shortening patently. Iterative process can be shorten by some other methods, such as to restrict the error in a scale or a fixed training algetra, or improve the inherit selection、crossing、aberrance operators by the means of applying enlightening information in genetic algorithm.

4. REFERENCES

[1]. Zhong Luo, Tang Chao, Xie Weiping et al. A Modified Back-propagation Algorithm [J]. Journal of Wuhan University of Technology, 1995，17(2):91~93.

[2]. Pan Hao，Zhong Luo，Chen Jie. Probing Modification of BP Neural Network Learning-Rate[J]. Journal of Hubei Polytechnoic University, 1997, 12(2):1~4.

[3]. Pan Hao，Chen Jie，Zhong Luo.Discussion about BP Neural Network Structure and Selection of Samples Training Parameters [J]. Journal of Hubei Polytechnoic University, 1997, 12(3):1~4.

[4]. Zhong Luo, Liu Lisheng, Zou Chengming et al. The Application of Neural Network in Lifetime Prediction of Concrete[J]. Journal of Wuhan University of Technology, 2002,17(1): 79~81.

[5]. Chen Guoliang，Wang Xifa，Zhuang Zhenquan et al. Genetic Algorithm & Application[M]. Posts&Telecom Press，2001.

Zou Chengming is an instructor, who is working at School of Computer Science and Technology,Wuhan University of Technology He got the doctor degree from Structure Engineering, Wuhan University of Technology in 2003.His research interests are neural network, distributed computation.

Tong Qiwei is a Member of Intelligent Technology and System Lab, Postgraduate Student of School of Computer Science and Technology, Wuhan University of Technology. She graduated from Wuhan University of Technology in 2002 with specialty of civil engineering. Her research interests are in Artificial Intelligence, Intelligent Calculation, Expert System, E-commence and distributed parallel processing.

Study on Logistics System Safety
Based on Neural Network and Fuzzy Probability

Li Bo Chen Dingfang Zhang Xiaochuan Li Wenfeng
Institute of Logistics Engineering,Wuhan University of Technology,Wuhan,Hubei,China
Email: whlibo@mail.whut.edu.cn **Tel:** +86-027-62590816

ABSTRACT

Based on man-machine the fault tree of Logistics System safety is obtained. When calculating Logistics system accident probability , the traditional fault tree can’t handle the imprecise incidents efficiently, this paper wields the related knowledge of fuzzy probability and fault tree to simulate. A neural network model is used to distribute the system reliability due to date being fuzzy, nonlinear and inclusive of noise. It is no doubt that this will make the probability calculation for Logistics system accident more reliable and accurate.

Keywords: Safety,Neural Network,Fuzzy probability, Fault tree, Logistics.

1. INTRODUCTION

Logistics has been now the main tendency of the development of various countries all of world as the technology of IT and internet are developed and the transportation is liberty in the world. The logistics system are needed better factors by modern market, that is: shorter period of predicting, lower cost of storing, quicker responding speed , higher lever of system optimization, better serving, more transparent process of material flowing and safer system, there have been much research to the preceding parts, but has few to the safety.

The safety of logistics system is studied based on man-machine, the pure probability method is difficult to solve it for fundamental event probability are not easily got accurately, system modeling is not precise due to human, dependent failure, common cause failure, system changeable and the ergodicity of working process of system. The probabilities of fundamental events in logistics system fault tree are unfixed or fuzzy, so it is difficult to analysis logistics system by traditional fault tree. Due to date being fuzzy, nonlinear and inclusive of noise, the distinguishing characteristic of pattern decided is indefinite, neural network is suitor than traditional model to distribute the system reliability,the parallel processing, adaptability, self-organization are also interested in to it.. We study the safety of logistics system by fuzzy fault tree (FTA) and a neural network model here.

2. ESTABLISHING FUZZY FAULT TREE

The fuzzy fault tree has been established to the logistics center of Hubei Jiuzhoutong pharmaceutical Co, Ltd (national GSP admitting enterprise, located in logistics region in Wuhan economics tap district).

Although there are a lot of factors responsible for safety of logistics, and with some fortuitous factors. The fault tree in Figure l only includes the main fundamental events that are selected from many incident statistical data, due to which ninety percent incidents happened.

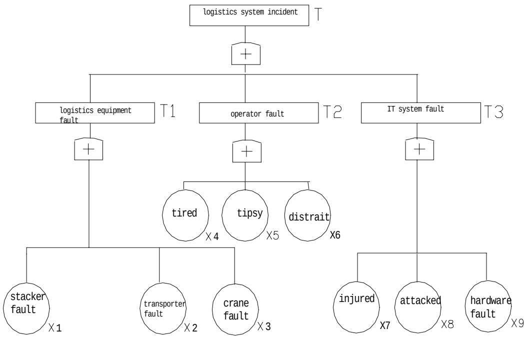


Figure1 Safe fault tree of logistics system

FTA analysis procedures:

Step1 choose top accident, use various logical gates to build a fault tree in which only “AND” and “OR” gates are included at last.

Step2 all events are classified into ones with statistics data, ones without statistics data and fuzzy ones.

Step3 probabilities of every fundamental event are obtained from experience, design handbook or expert, including of precise one, language one or fuzzy one.

Step4 according to certain rule, the precise data, language data and fuzzy one are all transferred into triangle fuzzy numbers.

Step5 minimal cut sets and the probability of top incident happening have been achieved based on the fuzzy fault tree.

Step6 results are used to different cases according to different practical problems.

3. ANALYSIS AND SIMULATION

The happening probabilities of fundamental events in fuzzy fault tree can be obtained from materials or experiential data, that are accurate. The fuzzy data and language could be used to appraise the fuzzy happening probabilities of fundamental events by experts in case of having no statistical data or to fuzzy events. The happening probabilities of fundamental events may have many data pattern: precise ones, language ones or fuzzy ones, before using them to analyze the safety by a fault tree, the amounts must be resulted in the same form. The triangle fuzzy numbers are used widely belongs to its linear property functions and simple handled. Various probabilities can be resulted in the triangle fuzzy numbers [1][2]. The method how to do is described in the following passages.

A precise probability p can be turned into a triangle fuzzy number z=(p,p,p).A nontriangle fuzzy number q, such as a normal fuzzy number, a LR fuzzy number, a trapezoid number, and so on, can be turned into a triangle one (a , m , b) after obtaining its centre of property function m. According to references [3] the relationship between a, m, b is:

$$m-a=b-m=0.556m \tag{1}$$

The formulas to find the centre x_1 of a triangle fuzzy number (a, m, b) and the one x_2 of a trapezoid number (a, b, c, d) are given below[3].

$$x_1=\frac{1}{3}(a+m+b) \tag{2}$$

$$x_2=\frac{c^2+d^2-a^2-b^2-ab+cd}{3(c-b+d-a)} \tag{3}$$

The languages are often used to express complex phenomena or uneasily to describe. In safety analysis. there are approximate probability numbers to approach to the language, in Chart 1.

Chart 1 The language and its probability numbers

lowest	low	lower	medium	higher	highest
0.0005	0.001	0.002	0.005	0.007	0.01

The formulas dealing with triangle fuzzy numbers are used in fault analysis based on fuzzy fault tree after all kinds of probability are resulted in the triangle fuzzy numbers. A triangle number p is expressed by three elements a, m, b, recording p= (a, m, b).Stipulate fuzzy number p_1=(a_1, m_1, b_1), p_2=(a_2, m_2, b_2), The fuzzy addition $\oplus$, the fuzzy subtraction Θ and the fuzzy multiplication $\otimes$ are differently defined:

$$\underline{p}_1 \oplus \underline{p}_2=\left(a_1+a_2,m_1+m_2,b_1+b_2\right) \tag{4}$$

$$\underline{p}_1 \Theta \underline{p}_2=\left(a_1+a_2,m_1+m_2,b_1+b_2\right) \tag{5}$$

$$\underline{p}_1 \otimes \underline{p}_2=\left(a_1\cdot a_2,m_1\cdot m_2,c_1\cdot c_2\right) \tag{6}$$

The And fuzzy calculus and the or one are discussed only because a fault tree can only include And calculus and Or one in the end (all other logic calculus's can be translated into And ones and Or ones).

The And calculus and Or calculus used in tradition fault tree analysis are given below.

$$q_{AND}=\prod_{i=1}^{n} q_i \tag{7}$$

$$q_{OR}=1-\prod_{i=1}^{n}(1-q_i) \tag{8}$$

The probability of incident i is q_i,this is precise. The And fuzzy calculus and the Or one are below according to fuzzy mathematics.

$$\underline{q}_{AND}=(a_{AND},m_{AND},b_{AND})=\prod_{i=1}^{n}\underline{q}_i= \tag{9}$$
$$\underline{q}_1\otimes\underline{q}_2\otimes\Lambda\otimes\underline{q}_n=\left|\prod_{i=1}^{n}a_i,\ \prod_{i=1}^{n}m_i,\ \prod_{i=1}^{n}b_i\right|$$

$$\underline{q}_{OR}=(a_{OR},\ m_{OR},\ b_{OR})=1\,\Theta\prod_{i=1}^{n}(1\,\Theta\,\underline{q}_i)$$
$$=1\,\Theta\prod_{i=1}^{n}(1\,\Theta(a_i,\ m_i,\ b_i))=$$
$$\left|\left|1-\prod_{i=1}^{n}(1-a_i)\right|,\ \left|1-\prod_{i=1}^{n}(1-m_i)\right|,\ 1-\left|\prod_{i=1}^{n}(1-b_i)\right|\right| \tag{10}$$

The probabilities of every fundamental event in safety fault tree of logistics system in Figure 1 are given in Chart 2.

Chart 2 The probabilities of every fundamental event

fundamental event	happening probability description	translation result
X1	low	
X2	0.003	$\begin{pmatrix}0.003,0.003,\\0.003\end{pmatrix}$
X3	0.004	$\begin{pmatrix}0.004,0.004,\\0.004\end{pmatrix}$
X4	0.002	$\begin{pmatrix}0.002,0.002,\\0.002\end{pmatrix}$
X5	0.001	$\begin{pmatrix}0.001,0.001,\\0.001\end{pmatrix}$
X6	0.005	$\begin{pmatrix}0.005,0.005,\\0.005\end{pmatrix}$
X7	normal fuzzy number $(m=0.01)$	$\begin{pmatrix}0.0009444,0.001,\\0.0010556\end{pmatrix}$
X8	0.005	$\begin{pmatrix}0.005,0.005,\\0.005\end{pmatrix}$
X9	lowest	$\begin{pmatrix}0.0004722,0.0005,\\0.0005278\end{pmatrix}$

According to calculation rules of fuzzy numbers above, middle incidents can be got: PT1-logistics equipment fault= $(0.007925,0.007981,0.008036)$,P_{T2}-operator fault= $(0.007983,0.007983,0.007983)$,P_{T3}-IT system fault= $(0.014862,0.015443,0.016023)$,logistics system fault as the top incident, its fuzzy probability is obtained at last, P_T=$(0.030472,0.031098,0.031722)$.It is determined that the probability of top incident happening is from 3.05%---3.17%,approximate value is 3%. So the system reliability is 97% approximately.

Minimal cut sets have been achieved based on the fuzzy fault tree.

$$T = T1 + T2 \ + T3 = \sum_{i=1}^{9} xi \tag{11}$$

And nine minimal cut sets, $\{x1\}$, $\{x2\}$,……., $\{x9\}$.

After obtaining the probabilities of every fundamental event in Chart 2, can use probability importance of every fundamental event to reflect its effect on top incident, various probability importance of fundamental event $I_g(i)$ is fixed below.

$$I_g(i) = \partial g \Big/ \partial g_i = \frac{\sum_{i=1}^{9}(1-g_i)}{(1-g_i)} \tag{12}$$

The probability importances of every fundamental event are given in Chart 3.

Chart3 The probability importance of every fundamental event

fundamental event	probability importance
X1	0.9699
X2	0.9718
X3	0.9728
X4	0.9708
X5	0.9699
X6	0.9738
X7	0.9787
X8	0,9738
X9	0.9694

Arrange in order,
$I_g(7) > I_g(6) = I_g(8) > I_g(3) > I_g(2) > I_g(4) > I_g(1) = I_g(5) > I_g(9)$

4. RELIABILITY DISTRIBUTING

The three middle incidents, logistics equipments fault, operator fault and information system fault, play different parts in the safety of logistics system, the composition of operator branch system can not be quantities due to its fuzziness so reliability are distributed to the three branch system by weighted mansard= RT1· RT2· RT3=0.97, fixxing RT3=0.999(largest), then determining RT3=1.01RT2, so RT2=0.98911, RT1=0.98166.

It is considered having no advantage to have more than two middle layers from experience. The more middle layers, the more complicated the calculating process of error passing afterwards, the amount of training time will increase drastically and minimal part error will also increase ,the network will cave in the process of seeking minimal part error .Probability is allotted to logistics equipment branch system by a 3-8-1 structure neural network in algebra distribution[4]. Expected probability, fundamental system complexity, importance are used as input joints of perceptions, the probability of fundamental system is used as output joint.

Algorithm process:

Step1 initiating, all weights are given random small numbers, first numbers are assumed to estimate one.

Step2 trainning data set, including of input vector o_{pi} and expected output o_{pj} are supplied.

Step3 real output o_{pj} is obtained below

o_{pj} =f($\sum w_{ij} o_{pi}$), w_{ij} are weights connecting neuron i with neuron j, o_{pi} is current input of neuron j, o_{pj} is output.

$$f = \frac{1}{1 + e^{-(x-\theta)}} \tag{13}$$

x--independent,
θ--constant

Step4 weights are adjusted from output node to hidden layers according to below formula

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_{pj}O_{pj} + a[w_{ij}(t) - w_{ij}(t-1)] \tag{14}$$

a--momentum coefficient, 0<a<1; δ_{pj} --error in node j;t--number of hidden layers

η --gain factor, η >0. η is larger at start of learning and less while approaching the optimal site preventing from weights shocking repeatedly.

$\delta_{pj} = \mathrm{O}_{pj}$ (1- O_{pj})(O_{pj} - T_{pj}), j is output node
or $\delta_{pj} = \mathrm{O}_{pj}$ (1- O_{pj}) $\sum_k \delta_{pk} w_{kj}$, j is hidden node

T_{pj} --ideal output, δ_{pk} --error in node k, w_{kj} --weights connecting neuron k with neuron j

Step5 return to Step2 calculating repeatedly until error is less than fixes one, expect error is not more than 0.1%.

The results of probability distribution and the relevant data of logistics equipment branch system are given in Chart4.

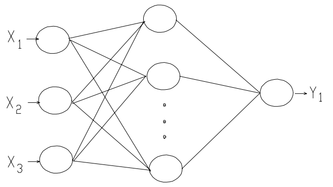


Figure 2 Probality distribution model of logistics equipment branch system

Model establishment and training are all based on Matlab5.3, the training and simulation of neural network model is easily carried out by its neural network box. Data are standardized before training neural network, then the data are changed into an order having import sequence, Trainlm calculus is used to train sample, Logsig function is applied in hidden layer and Purelin function in Output layer in neural network. Altering speed study method is used in training; expected error in studying is 0.4%.

Chart4 Logistics equipment branch system data and distribution probability

fundamental system	X1	X2	X3
Constituent quantities	12	27	9
importance	0.8	1.0	1.0
complexity	0.31	0.57	0.12
distribution probability	0.977	0.993	0.969

5. CONCLUSIONS

It is proposed that safety of logistics is studied based on man—machine system, the fault tree is obtained under the special environment .The happening probabilities of fundamental event in the fault tree is resulted in the same form, the event are described in various fuzzy numbers, language ones and precise ones, all of them show that simulation is suitable and valued, the branch system complexity is reflected and reliability are distributed reasonably for using a neural network. The model and method under above specified case can also be used commonly.

6. REFERENCES

[1]. D. Singer, "A Fuzzy Set Approach to Fault Tree and Reliablity Analysis"[J], Fuzzy Sets and Systerms. 34. 1990. 145~155

[2]. J. A. B. Geyrmay and N.F.F.Ebecken, "Fault Tree Analysis; A Knowledge-Engieering Approach"[J], IEEE Transactions on Reliability. Vol, 44, No.1, 1995 March

[3]. Dong Yuge,Zhu Wangyu ,Chen Xinchao,"Fuzzy Fault Tree Analysis and Its Application"[J],Journal of Hefei University of Technology.Vol,19,No.4,Dec.1996. 35~40

[4]. Xu Yong,Hou Chaozhen,Yang Guosheng, "Neural Network Model of System Reliability Assignment"[J],Systems Engineering and Electronics,Vol,23,No.1,2001. 90~93

li Bo is a vice professor and a head of department of Industrial Engineering of Institute of Logistics Engineering, Wuhan University of Technology. From 1980 to 1984 he studied diesel engine in the power department of Wuhan Water Transportation Engineering College and was graduated with an B.A. degree in engineering; In 1988 he was enrolled to study in the mechanical engineering department of Wuhan Institute of Technology as a postgraduate on machine, he had been studying there for 2.5 years and finished the courses with master's degree in engineering. He has published one book, 16 Journal papers. His research interests are in machine design, distributed parallel processing, and logistics engineering.

Research on a Back-Propagation Neural Network Based Q Learning Algorithm in Multi Agent System

Lin Ouyang
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430063, China
Email: linyoweb@sohu.com

Qingping Guo
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430063, China
Email: qpguo@public.wh.hb.cn

Santai Ouyang
Department of Electric and Information Engineering, Hunan Institute of Technology
Xiangtan, Hunan, 411104, China
Email: santai_ouyang@mail1.hnie.edu.cn

ABSTRACT

Following the development of the artificial intelligence, the research of reinforcement learning of multi agent and the neural network become more and more prevail. The Q learning algorithm, as a kind of reinforcement learning, is a kind of online learning method. Following increasing of the scale of the problem, the exploration space becomes too enormous to deal with by the traditional Q learning algorithm. The neural network, as a kind of self-organization, self-adaptive and supervised method on learning, can hide the inner continuous connection between the input and the output of problem. The combination of the neural network with Q learning algorithm, which called back-propagation neural network based Q learning algorithm (BPNNQ), can reduces the exploration space remarkably, by take advantage of the neural network and the Q learning reinforcement learning methods. How to avoid falling into local optimal solution is another difficult problem in machine learning. Through the using of the Boltzmann distribution strategy in the BPNNQ algorithm, the locale optimal solution is solved to a certain extent.

Keywords: Multi agent, Neural Network, Q Learning, Reinforcement Learning

1. INTRODUCTION

The concept of agent appeared in 70's of last century. Following the rapidly developing of agent, agent-based systems technology has applied widely in various fields, and generated lots of excitement in recent years, because it becomes a new paradigm for conceptualizing, designing, and implementing software systems.

1.1. Agent

But what is agent?[1] Perhaps the most general way in which the term agent is used is to denote a hardware or (more usually) software-based computer system that enjoys the following properties:

Autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state.

Social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language.

Reactivity: agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the internet, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it.

Pro-activeness: agents do not simply act in response to their environment, and they are able to exhibit goal-directed behavior by taking the initiative.

For some researchers, particularly those working in AI, the term 'agent' has a stronger and more specific meaning than that sketched out above. These researchers generally mean an agent to be a computer system that, in addition to having the properties identified above, is either conceptualized or implemented using concepts that are more usually applied to humans. It is characterize an agent using mentalistic notions in AI field, such as knowledge, belief, intention, and obligation.

1.2. Multi Agent

The single agent-based system is too simple to adapt the complicated and flexible problems in the real world. Because of the complication and flexibility of the problems, Multi agent system (MAS)[2] becomes more and more prevail. It attracts more and more experts and people, who work in distributed artificial intelligence (DAI), machine learning (ML) and so on, to developing it.

Multi agent consists of a group of agents. The agents are considered to be autonomous entities, such as software programs or robots. The agents in this group can collaborate and cooperate each other for a common goal (e.g. an ant colony), or pursue their own interests (as in the free market economy).

The characteristics of MAS are that

- Each agent has incomplete information or capabilities for solving the problem and, thus, has a limited viewpoint;

- There is no system global control;
- Data are decentralized; and
- Computation is asynchronous.

Multi agent can be used in many problems in lieu of a single agent difficult or can not to deal with. It can make a complex learning task simpler or to achieve better performance.

Following the developing of multi agent system, multi agent learning becomes prevail in machine learning field too. Usually we use Q learning, which is the most broadly used in reinforcement learning of machine learning, in multi agent learning. But because of the enormous searching space of problem solving in multi agent learning, the convergence of Q learning algorithm is too difficult or impossible. In this paper, we introduce a neural network based Q learning algorithm to solve this problem.

2. REINFORCEMENT LEARNING

2.1. Reinforcement Learning

Reinforcement learning[3][4] is learning what to do, how to map situations to actions, so as to maximize a numerical reward signal. In reinforcement learning, the learner is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. The framework of reinforcement learning[5] is represented by figure 1.

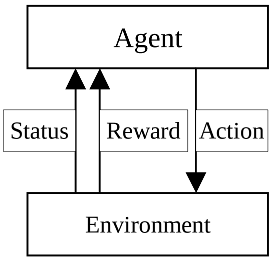


Fig. 1 The Framework of Reinforcement Learning

Reinforcement learning is an online learning technique and it’s relevant to dynamic programming which usually to be used to solve the conventional optimal control problems. The problem solved by reinforcement learning is that, how an autonomic agent, who can perceives the environment around, selects the optimal action to achieve the target of task through learning.

Reinforcement learning is very different from supervised learning studied in most current research, such as machine learning, statistical pattern recognition, and artificial neural networks. Supervised learning is learning from examples provided by some knowledgeable external supervisor. But supervised learning is not adaptive for interactive problems which are often difficult or impractical to obtain examples of desired behavior that are both correct and representative of all the situations in which the agent has to act.

2.2. Markov Decision Process

In reinforcement learning, the external real world is modeled as a discrete time. It can be formalized to an action sequence. The agent in this learning should learn an optimal control policy through exploring the whole environment states. A reward function can be defined, and we can set different values to different action selected in different environment states, the value is immediate payoff. Following the action sequence of agent, we can get the maximum reward value in the whole environment state spaces. The action selection of learning control policy looks like the problem of function.

The task of reinforcement learning is to maximize the long-term discounted reward per action. In the most cases, actions may affect not only the immediate reward, but also the next situation and, through that, all subsequent rewards. These two characteristics, trial and error search and delayed reward, are the two most important distinguishing features of reinforcement learning.

Usually, the reinforcement learning of multi agent formalized either to a determinative Markov decision process (MDP)[9], or to a non-determinative Markov decision process, according to the model of external real world.

The task of agent is to learn a policy $\pi: S \rightarrow A$, it selects a next action a_t according to the state s_t currently observed, that is $\pi(s_t) = a_t$. In order to select a optimal policy π, the discounted cumulative reward of the action selected by the agent should be the maximum. We can define the discounted cumulative reward value $V^{\pi}(s_t)$ as a equation:

$$\begin{aligned} V^{\pi}(s_t) &\equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \\ &\equiv \sum_{i=0}^{\infty} \gamma^i r_{t+i} \end{aligned} \qquad (1)$$

In this equation, the reward sequence r_{t+i} is created by using the policy π iterative to select the action from the state s_t (such as, $a_t = \pi(s_t)$, $a_{t+1} = \pi(s_{t+1})$, etc.). The value of discounted factor γ, which is a constant and determine the proportion of delayed reward and immediate reward, is between 0 and 1 ($0 \leq \gamma < 1$). If γ is 0, only the immediate reward is taken into account, and if γ value approximate 1, the delayed reward becomes more and more important than the immediate reward. We can use other kinds of rewards in different environments, such as finite horizon reward ($\sum_{i=0}^{h} r_{t+i}$), average reward ($\lim_{h \rightarrow \infty} \frac{1}{h} \sum_{i=0}^{h} r_{t+i}$). The finite horizon reward takes account of the reward within finite h steps, and the average reward takes account of average reward of the whole life cycle of agent. And now we can us π^* to indicates the optimal policy:

$$\pi^* \equiv arg_{\pi} \max V^{\pi}(s), (\forall s) \qquad (2)$$

So the value function of the optimal policy is $V^{\pi^*}(s)$, and we indicate it as $V^*(s)$ briefly.

That is to say, the optimal action under the state s is to make the sum of the immediate reward and the optimal V^* value of immediate next state to be the maximum. We can indicate it as follow:

$$\pi^*(s) = arg_a \max[r(s, a) + \gamma V^*(\delta(s, a))] \qquad (3)$$

Here $r(s, a)$ means the immediate reward and $\delta(s, a)$ means immediate next state. Obviously, the agent can get the optimal policy by learning V^* if the immediate reward function *r(s, a)* and the state convert function $\delta(s, a)$ can be perfectly defined.

3. Q LERNING

Q learning is a recent form of Reinforcement Learning algorithm that does not need a model of its environment and can be used on-line. Therefore, it is very suited for repeated games against an unknown opponent.

3.1. Q Learning Algorithms

Q learning algorithm[6] works by estimating the values of state-action pairs. The value *Q(s, a)* is defined to be the expected discounted sum of future rewards obtained by taking action *a* from a state *s* and following an optimal policy there after.

$$Q(s, a) \equiv r(s, a) + \gamma V^{*}(\delta(s, a)) \qquad (4)$$

Equation (3) can be rewrite as follow:

$$\pi^{*}(s) = \arg_{a} \max Q(s, a) \qquad (5)$$

Equation (5) indicates that even if the immediate reward function *r(s, a)* and the state convert function $\delta(s, a)$ can not be perfectly defined or we lack of the knowledge of the function *r(s, a)* and the function $\delta(s, a)$, the agent can select the optimal action through learning function *Q(s, a)*. In other words, the agent only needs to take account of each action *a* under current state *s* to select the optimal action *a* which make the *Q(s, a)* value maximum.

Once these *Q(s, a)* values have been learned, the optimal action from any state is the one with the highest *Q* value. After being initialized to arbitrary numbers, usually we initialize into 0, *Q* values are estimated iteratively as equation (6) where $\hat{Q}(s, a)$ is estimated value of *Q*.

$$\hat{Q}(s, a) \leftarrow r(s, a) + \gamma \max_{a'} \hat{Q}(s', a') \qquad (6)$$

The Q learning algorithm can be described as follows:

```
Q learning algorithm
Initialize value 0 to $\hat{Q}(s, a)$ for each state-action pairs
Observe current state s
Repeat:
    Select an action a and execute it
    Receive the immediate reward of r
    Observe the new state $s'$
    Update $\hat{Q}(s, a)$ according to:
        $\hat{Q}(s, a) \leftarrow r(s, a) + \gamma \max_{a'} \hat{Q}(s', a')$
    $s \leftarrow s'$
```

3.2. The Convergence of Q Learning Algorithm

This algorithm is guaranteed to converge to the correct Q values if the environment is stationary and depends on the current state and the action taken in it; called Markovian, a lookup table is used to store the Q values, every state-action pairs continues to be visited. This exploration strategy does not specify which action to select at each step.

The convergence of Q learning algorithm is testified in mathematics as follows:

The convergence of Q learning algorithm
Hypothesis:
The agent who is learning Q function is in a determinative MDP which has boundary reward, that is: $(\forall s, a)|r(s, a)| \le c$.
The discounted factor γ, $0 \le \gamma < 1$.
$\hat{Q}_n(s, a)$ *means the* $\hat{Q}(s, a)$ *after the nth refresh.*
And every state-action pairs can be accessed frequently infinitely.
Then:
For each state-action pairs, $\hat{Q}_n(s, a)$ *will converge to* $Q(s, a)$ *when* $n \to \infty$.
Testify:
We define $\Delta_n = \max_{(s, a)} | \hat{Q}_n(s, a) - Q(s, a) |$
and $s' = \delta(s, a)$

$$| \hat{Q}_n(s, a) - Q(s, a) |$$
$$= |(r + \gamma \max_{a'} \hat{Q}_n(s', a')) - (r + \gamma \max_{a'} Q(s', a'))|$$
$$= \gamma |\max_{a'} \hat{Q}_n(s', a') - \max_{a'} Q(s', a')|$$
$$\le \gamma \max_{a'} | \hat{Q}_n(s', a') - Q(s', a')|$$
$$\le \gamma \max_{(s'', a')} | \hat{Q}_n(s'', a') - Q(s'', a')|$$

that is: $| \hat{Q}_{n+1}(s, a) - Q(s, a) | \le \gamma \Delta_n$
$\Delta_{n+1} \le \gamma \Delta_n$
So when $n \to \infty$, $\Delta_n \to 0$
So For each state-action pairs, $\hat{Q}_n(s, a)$ *will converge to* $Q(s, a)$ *when* $n \to \infty$.

3.3. Non-determinative Q Learning Algorithm

In most cases, the reward function and the action convert function may have an uncertain output with a probability, such as, robot systems with noises. In this instance, *r(s, a)* and $\delta(s, a)$ can be treated to a probability of state *s* and action *a*. We call it non-determinative the Markov decision process. The policy value should be redefined to the expectation of the discounted cumulative reward as follow:

$$V^{\pi}(s_t) \equiv E[\sum_{i=0}^{\infty} \gamma^{i} r_{t+i}] \qquad (7)$$

And the *Q(s, a)* can be redefined as follows too:

$$Q(s, a) \equiv E[r(s, a) + \gamma V^{*}(\delta(s, a))] \qquad (8)$$
$$\equiv E[r(s, a)] + \gamma E[V^{*}(\delta(s, a))] \qquad (9)$$
$$\equiv E[r(s, a)] + \gamma \sum_{s'} P(s'|s, a) V^{*}(s') \qquad (10)$$
$$Q(s, a) \equiv E[r(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q(s', a')] \qquad (11)$$
$$\hat{Q}_n(s, a) \leftarrow (1-\alpha_n) \hat{Q}_{n-1}(s, a) + \alpha_n[r + \gamma \max_{a'} \hat{Q}_{n-1}(s', a')] \qquad (12)$$

Where

$$\alpha_n = \frac{1}{1+visits_n(s, a)} \qquad (13)$$

In equation (13), the parameter $visits_n(s, a)$ means the total access times of the state-action pair *(s, a)* within *n* times repeats. If the value of α_n is 1, equation (12)[12] convert to the determinative leaning. In this instance, the value of $\hat{Q}(s, a)$ can converge to $Q(s, a)$ too[7].

4. BACK-PROPAGATION NEURAL NETWORK BASED Q LEARNING

In the case of multi agent environment, the exploration space of state-action pairs becomes more and more enormous following the increasing of num of agents. The time of Q learning will become so long that it’s unpractical. In order to solve this problem, we combine neural network and Q learning together.

Neural network[10] is a kind of inducing learning essentially. Through the repeating learning of a great deal of instances, the inner self-adaptive processing modifies the weights of the connection between neurons. Finally, the weights of neural network converge to a range of stability. The process of modifying of weights is the process of learning. It’s very different from traditional artificial intelligence.

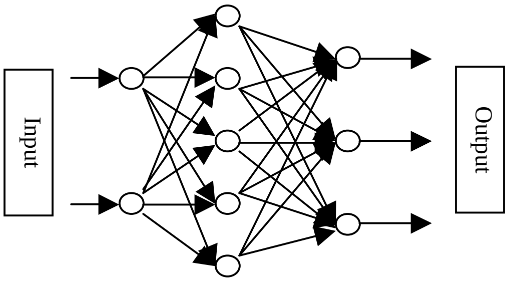


Fig. 2 Back-Propagation Neural Network

Some kinds of neural networks have been reported recently. Especially in automatic control field, the radial base function neural network is combined the reinforcement learning. Radial functions are simply a class of functions. In principle, they could be employed in any sort of model (linear or nonlinear) and any sort of network (single-layer or multi-layer). Generally, the radial base function neural network is more suitable for automatic control field, like the reinforcement learning in RoboCup[13].

The Back-Propagation Neural Network (BPNN)[8] is a kind of single direction propagation neural network who has multi layers (Fig 2). Except the input and the output layer, there are one or more middle layers called hidden layers In BPNN. There are no connections between the neurons of the same layer, and usually the sigmoid function is chose as the pass function.

Both the BPNN and the RBFNN are forward feedback neural network. They can get the function approximator of arbitrary continuous nonlinear function both. For comparison, the RBF network uses Gaussian function, whose input range is limited to a distance scale, as the transfer function of neural network, but the BP network uses sigmoid function, which has infinite input scale. Their characteristic feature of RBF network is that their response decreases (or increases) monotonically with distance from a central point. The center, the distance scale, and the precise shape of the radial function are parameters of the model, all fixed if it is linear, or can alternate if it is nonlinear. So the application of RBF network is limited.

On the other hand, following the increasing of the count of agents, the state space increasing rapidly, and quantity of neurons of middle layer in RBF network will increase more rapidly than in BP network. So we choose BP network to combine Q learning algorithm in our research.

The BPNNQ algorithm is composed of several modules, which we call IPM (information perceiving module), LM (learning module) and ASM (action selecting module). Each agent works individually and simultaneously, and interacts through the effect of environment. The framework of BPNNQ illustrated as figure 3.

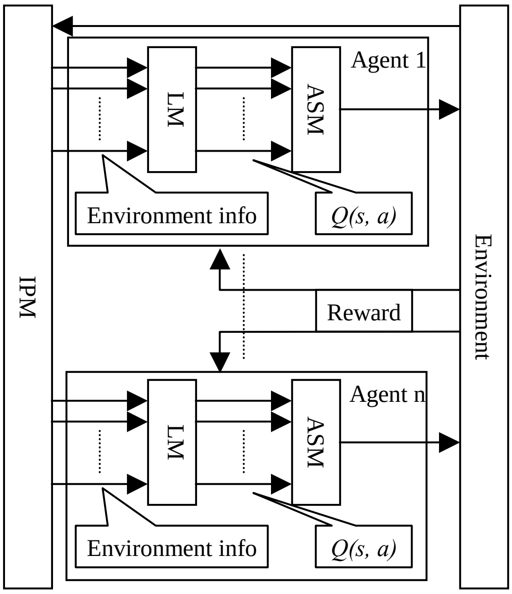


Fig. 3 The framework of BPNNQ learning algorithm

IPM: In BPNNQ learning algorithm, the agents get environment information by IPM. Each agent only has partial environment knowledge including the information of the other agents’ states. And after each action of each agent, the environment state will be changed, and a reward value $r(s, a)$ returned.

LM: The LM uses the algorithm of BPNN to learn knowledge. The input values of LM are partial environment state outputted by IPM and the output values of LM are the Q values.

ASM: The agents select an action by ASM. The ASM compares the input value, which is the sum of Q value and the immediate reward value, and then produces output, an action, to act on the environment under the condition of the Boltzmann distribution strategy.

In LM module, the inputs are environment state including other agents’ states, and the outputs are $\hat{Q}_n(s, a)$, the *n*th iterative value of $Q(s, a)$.

We define the error value as ΔQ:

$$\Delta Q = \hat{Q}_n(s_n, a_n) - \hat{Q}_{n-1}(s_{n-1}, a_{n-1})$$
$$= r_{n-1} + \max_a \hat{Q}_n(s_n, a) - \hat{Q}_{n-1}(s_{n-1}, a_{n-1}) \quad (14)$$

The weights of BPNN modified through the gradient descent method, which defined as equation (15).

$$w(k+1) = w(k) + \eta(k)(- \frac{\partial \Delta Q}{\partial Q})|_{w=w(k)} \qquad (15)$$

In equation (15), *w(k)* means the weights of the *k*th iteration, and $\eta(k)$ is alternative learning rate that can be modified by man or system to control the velocity of the weights modification, and its value is between 0 and 1.

If ΔQ falls into limited range, that means the $\hat{Q}_n(s, a)$ has converge to *Q(s, a)* approximately, then the system stops back propagation, and the weights reach to a stable value, otherwise, the system will continue to modify the weights of BPNN to fall into convergence by equation (15).

Usually, in order to avoid trapping into the local optimal, some kinds of global optimal algorithms is adopted. Here we choose one kind, the Boltzmann distribution strategy, to solve this problem. The ASM selects an action by Boltzmann method to ensure sufficient exploration while still favoring actions with higher value estimates. The Boltzmann distribution strategy is based on probability[11].

$$prob(a_j) = \frac{e^{Q(s,\, a_j)/T}}{\sum_{a_j} e^{Q(s,\, a_j)/T}} \qquad (16)$$

The randomicity becomes larger when the parameter *T* becomes larger, vice versa. The probability becomes larger when the *Q(s, a)* value becomes larger, vice versa too. In order to control the probability, the T value can be set by hand or automatic by system.

5. CONCLUSIONS

In the BPNNQ algorithm researched in this article, each agent of multi agent system learns knowledge individually according the partial environment knowledge. The agents cooperate and collaborate to achieve the common goal implicitly under the environment state. Each agent in MAS selects an action, which has the maximum Q value, and gives effects to the environment. Through the action and the effect, other agents' info has been included into the environment. In another words, the environment state results from the total agents' effect and the initial environment state. In this case, the agents' information is represented implicitly in environment, so the exploration space decreased greatly by the using of NN. Through the combination of the neural network and the Q learning algorithm, the BPNNQ algorithm overcomes the shortcoming of the traditional Q learning algorithm, which cannot deal with the great scale multi agent system because of the enormous exploration space. On the other hand, through the Boltzmann distribution strategy, which effects the action selection of agent by probability, the algorithm avoids falling into the local optimal space, which usually occurs in reinforcement learning.

6. REFERENCES

[1]. Seel, N, Agent Theories and Architectures, PhD thesis, Surrey University, Guildford, UK, 1989.

[2]. M. P. Singh, Multi-Agent Systems: A Theoretical Framework for Intentions, Know How and Communications, Springer-Verlag, 1994.

[3]. R. S. Sutton, Reinforcement Learning Architectures for Animats. In J. –A. Meyer and S. W. Wilson, editors, From Animals to Animats 1: Proceedings of the First International Conference on The Simulation of Adaptive Behavior, The MIT Press/Bradford Books, Cambridge, 1991, pp. 105-124.

[4]. Leslie Laelbling, Michael L Littman, Reinforcement Learning: A Survey, Journal of Artificial Intelligence Research , Vol.4, No.1, 1996, pp.237～285.

[5]. Tom M. Mitchell, Machine Learning, McGraw-Hil Companies, Inc, 1997.

[6]. Watkins C., Learning from Delayed Rewards, Thesis, University of Cambridge, England, 1989.

[7]. Watkins C., Dayan, P., Q-learning, Machine Learning, 1992.

[8]. S. Horikawa, On Fuzzy Modeling Using Fuzzy Neural Networks with BP Algoritym, IEEE Trans, Neural Networks, No.2, 1992.

[9]. Yang Gao, Zhihua Zhou, Jiazhou He, et al, Research on Markov Game-based Multi Agent Reinforcement Learning Model and Algorithms, Journal of Computer Research & Development, Vol.37, No.3, 2000, pp.257-263.

[10]. Wang Xingce, Zhang Rubo, Gu Guochang, Research on Multi-agent Team Formation Based on Reinforcement Learning, Computer Engineering, Vol.28, No.6, 2002, pp.15-16, 98.

[11]. Wang Xingce, Zhang Rubo, Gu Guochang, Research on Dynamic Team Formation of Multi Robots Reinforcement Learning, Journal of Computer Research and Development, Vol.40, No.10, 2003, pp.1444-1450.

[12]. Hong Bingrong, Piao Songhao, Multi-robot Cooperation Based on Conflict Resolution, Journal of Harbin Institute of Technology, Vol.35, No.9, 2003, pp.1053-1055.

[13]. Luo Qing, Li Zhijun, Iqbal Nadeem, et al, Study on Radial Basis Function Networks Based Reinforcement Learning in Robot Soccer, Journal of System Simulation, Vol.14, No.8, 2002, pp.1094-1097.

Ouyang Lin, male, is a master candidate in School of Computer Science and Technology, Wuhan University of Technology. His research directions are Artificial Intelligence and computer network.

Guo Qingping is a Full Professor and a head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He graduated from Wuhan University in 1968; from Huazhong University of Science and Technology in 1981 with specialty of wireless technology. 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 the UK Royal Society (1994), 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, co-chair of DCABES 2002, and will be the chairman of DCABES 2004. 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-commence.

Ouyang Santai, male, is a senior engineer and a head of PLC laboratory of Hunan Institute of Technology. He graduated from Wuhan Institute of River Traffic Engineering. His research interests are PLC, electric automatic control and automatic detection.

Recursive Neural Networks and its Application in Forecasting The State of Metal Oxide Arrester*

Zhou Long
Electrical Information & Engineering Department, Wuhan Polytechnic University
Wuhan,Hubei Province Postcode 430023, P.R. China
Email: zhoulong123@sohu.com **Tel:** +86(0)27)83950230

ABSTRACT

A method is put forward to forecast the Metal Oxide Arrester's (MOA's) state based on the recursive neural network. By compared with the real example, it is better than the usual BP algorithm. In order to realize the intellectual diagnosis and forecast of the MOA's state, we provide an effective method.

Keywords: recursive neural network, Metal Oxide Arrester (MOA), forecast.

1. INTRODUCTION

The MOA is important extra-voltage protection equipment; its liability directly decides the running safety of electric network. Because of the system voltage's effect, there is some leak current passing the MOV, therefore, it is important to supervise and diagnose the MOA on line, it will provide a method to avoid the malignant accident.

Many research results and working experience indicate that the majority of MOA's fault symptoms put up to be the increase of resistance leak current, and when it is under the effect of the working voltage in a long term, the deterioration of MOV is regular, whose phenomena are the changes of resistance leak current. In general, these changes are nonlinear. The Reference [3] provided a method using BP network to forecast the resistance leak current, and it resolves the nonlinear forecast problem of resistance current relatively effectively, and provides a better method for pre-management and pre-examination of MOA.

Because the BP network is a static neural network, it doesn't think enough about the relative relations among samples in forecast process. The static network only carries out nonlinear mapping between input and output, without dynamic feature. But the problem of MOA's state forecast belongs to the problem of dynamic system identification and object tracking, and there is affinity between input and output and among samples. Therefore, dynamic neural network will greatly improve the forecast performance. In this paper, we put forward a method to forecast the MOA's state based on the recursive neural network. And the result indicates that this method is effective.

2. RECURSIVE NEURAL NETWORK AND ITS ALGORITHM

The recursive neural network is formed based on the sequential system idea. The hidden units and output units' neurons in BP network are regarded, as a sequential system, and the output of each neuron will all influence the output of the neuron behind, and the neuron is only influenced by the neuron before itself.

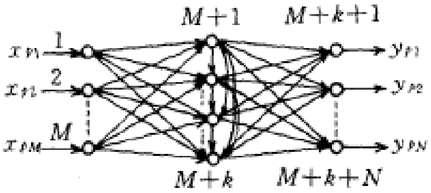
Recursive neural network is a dynamic neural network, the basis of dynamic neural network learning algorithm is sequential partial differential coefficient, then here's the brief introduction of it.

Suppose that $\{z_1, z_2, \Lambda, z_i, \Lambda, z_j, \Lambda, z_n\}$ is a set of variables, if z_i is only the function of variable set $\{z_1, z_2, \Lambda, z_{i-1}\}$, then the variable set is called sequential set. In order to be distinguished from the common partial differential coefficients, $\partial^+ z_j \big/ z_i$ represents the z_i's partial differential coefficients towards z_i, and the constant is $\{z_1, z_2, \Lambda, z_{i-1}\}$, and the variable is $\{z_1, \Lambda\ z_j, \Lambda, z_n\}$, the sequential partial differential coefficients have two properties defined as follow:

1) $$\frac{\partial^+ z_{i+1}}{\partial z_i} = \frac{\partial z_{i+1}}{\partial z_i} \qquad (1)$$

2) $$\left.\begin{aligned} &when \cdot j > i, \frac{\partial^+ z_j}{\partial z_i} = 0 \\ &when \cdot j > i+1, \frac{\partial^+ z_j}{\partial z_i} = \frac{\partial z_j}{\partial z_i} + \sum_{k=i+1}^{j-1} \frac{\partial^+ z_j}{\partial z_k} \bullet \frac{\partial z_k}{\partial z_i} \end{aligned}\right\} \qquad (2)$$

In the dynamic recursive neural network, the relationship among the training samples is shown between the neurons of the network structures, that is the connection weight. The typical three-layer recursive BP neural network structure is shown as Figure 1, the basic structure is the same as BP network, and the difference is only the introduction of the transverse connection between the hidden layer and the output layer.



Input layer hidden layer output layer

Figure 1 three-layer recursive BP neural network structure

* This work is supported by National Natural Science Foundation under Grant 79970025 of China, Foundation of Hubei Provincial Department of Education under Grant 2003ABA053, The Youth Chenguang Project of Science and Technology of Wuhan City in China, and Scientific Research Team of Wuhan Polytechnic University under Grant 03-T-01.

The algorithm of recursive neural network is same as that of BP, also including the forward calculation and the error back propagation. The error function E is defined as the DMS between the expected output and the actual output

$$E = \frac{1}{2}\sum_{p=1}^{p}\sum_{j=1}^{N}\left(y_{pi} - y_{pj}\right)^2 \tag{3}$$

Where y_{pj} is the actual output of the network; d_{pj} is the expected output; N is the neuron number of the output layer; P is the number of the training samples.

First, each neuron of the input layer, hidden layer and output layer in the network should be labeled with number, in order to make them a sequential system, just as the Fig. 1. As for some sample P , the algorithm is following.

The forward calculation process in the network is:

- The output of the input layer node equals to its input.
- The input and the output of the hidden layer node is decided by equation (4):

$$\left.\begin{aligned}
&net_{pk} = \sum_{i=1}^{k-1} W_{ki} O_{pi} + b_k \\
&M + 1 \le k \le M + k \\
&O_{pk} = f\left(net_{pk}\right) = 1 \big/ \left[1 + \exp\left(-net_{pk}\right)\right]
\end{aligned}\right\} \tag{4}$$

The input and the output of the output node are respectively as following:

$$\begin{aligned}
&net_{pj} = \sum_{k=M+1}^{k-1} W_{jk} O_{pk} + b_j \\
&M + k + 1 \le j \le M + k + N \\
&y_{pj} = f\left(net_{pj}\right) = 1 \big/ \left[1 + \exp(-net_{pk})\right]
\end{aligned} \tag{5}$$

Where W_{ki}, W_{jk} are respectively the connection weight between hidden node K and input node i , and between output node j and hidden node K ; b_k and b_j are the thresholds of corresponding nodes.

The error back propagation process is:

As to the connection weight W_{jk} and threshold b_j between output node and hidden node, according to the equation (1) and equation (2) and its definition, we can get that

$$\left.\begin{aligned}
&\Delta_p W_{jk} = -\eta \frac{\partial^+ E}{\partial W_{jk}} = -\eta \sum_{p=1}^{p} \frac{\partial^+ E_p}{\partial net_{pj}} \bullet \frac{\partial net_{pj}}{\partial W_{jk}} = \eta \delta_{pj} O_{pk} \\
&\Delta_p b_k = \eta \bullet \delta_{pk}
\end{aligned}\right\} \tag{6}$$

Where η is the learning rate, E_p is the error function of the sample P , that is

$$E_p = \frac{1}{2}\sum_{j=1}^{N}\left(y_{pi} - y_{pj}\right)^2 \quad ,$$

if $j < r \le M + k + N$ then

$$\delta_{pj} = \frac{\partial^+ E_p}{\partial net_{pj}} = \frac{\partial E_p}{\partial net_{pj}} + \sum_{r=j+1}^{M+k+N} \frac{\partial^+ E_p}{\partial net_{pr}} \bullet \frac{\partial net_{pr}}{\partial net_{pj}}$$

When, $j = M + k + N$ then

$$\delta_{pj} = \frac{\partial \ E_p}{\partial net_{pj}} = \frac{\partial E_p}{\partial net_{pj}} \bullet \frac{\partial y_{pj}}{\partial net_{pj}}$$
$$= (d_{pj} - y_{pj}) \bullet y_{pj} \bullet \left(1 - y_{pj}\right)$$

When $M + k + 1 \le j < M + k + N$,

$$\frac{\partial net_{pr}}{\partial net_{pj}} = \frac{\partial net_{pr}}{\partial y_{pj}} \bullet \frac{\partial y_{pj}}{\partial net_{pj}} = W_{rj} y_{pj} (1 - y_{pj}) \quad ,$$

so

$$\delta_{pj} = (d_{pj} - y_{pj}) \bullet y_{pj} \bullet \left(1 - y_{pj}\right) +$$
$$y_{pj}\left(1 - y_{pj}\right) \sum_{r=j+1}^{M+k+N} W_{rj} \delta_{pr}$$

- As for the connection weight W_{kj} and the threshold b_k between hidden node and input node, we can get that

$$\left.\begin{aligned}
&\Delta_p W_{kj} = \eta \delta_{pk} O_{pi} \\
&\Delta_p b_k = \eta \delta_{pk} \\
&where \cdot \delta_{pk} = O_{pk} \bullet \left(1 - O_{pk}\right) \sum_{j=k+1}^{M+k+N} W_{jk} \delta_{pj}
\end{aligned}\right\} \tag{7}$$

The training learning process mentioned above indicates that the algorithm of recursive neural network embodies the idea of time sequence, and thinks a lot about the relation among neurons.

In order to test the validity of the forecasting method, the relative EMS is introduced to measure the precision of forecast, that is

$$E_f = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left[\left(X(K) - Y(K)\right) / X(K)\right]^2} \tag{8}$$

Where E_f represents the relative EMS of forecast, n is the number of test points, $X(K)$ represents the actually tested value of the nonlinear sequence, $Y(K)$ is the forecasting value of the sequence.

3. THE STATE FORECAST OF MOA

The actual running theories and experience indicate that the main reason, which causes the increase of MOA's resistance leakage current, is aging and moistening of MOV. The deterioration process is a gradually changing process, the phenomenon is that the current value of resistance leakage current is somewhat related to that of the foretime, this relation is the basis of forecast. Because the relation is nonlinear, it is reasonable to use recursive neural network to carry out forecasting and fitting.

As for the forecast of nonlinear time sequence, we should use the neural network to train the sequence value of past time section by section, and build the relationship among sequences into the neural network model, then it will be made able to forecast, therefore, the number of neurons of the input layer and hidden layer will effect the forecast precision. In this paper, we use a three –layer BP network structure with a hidden layer, it has three input nodes, one output node, five hidden nodes. As is shown in Fig.1, the inputs are resistance leak current of continuous time sequence, the values behind are outputs, after the training and learning according to the recursive neural network algorithm, we can carry out the state forecast of MOA, from the analysis mentioned before, we can conclude that the more layers and nodes are, the slower speed of convergence is. In order to speed the learning convergence, we often use the equation below:

$$\Delta W(t) = -\eta \frac{\partial E}{\partial W(t)} + \alpha \Delta W(t-1) + \beta \Delta W(t-2) \qquad (9)$$

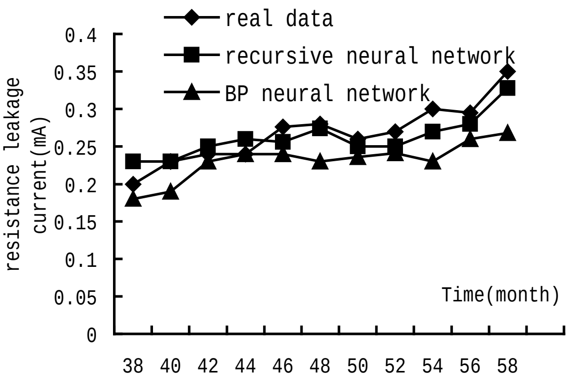


Fig.2 comparison between the actual value and forecast value of resistance leak current

Where $\Delta W(t)$ is the weight revision value of the network; η is the learning rate; α is the parameter of one-order momentum$(o < \alpha < 1)$; β is the parameter of two-order momentum $(\beta < 0)$; t is the iterative times of learning$(t = 0,1,\Lambda)$, because the network structure is not complex, one-order momentum is only added to revise the weight, $\alpha = 0.6$, $\eta = 0.8$, the samples are also the data mentioned in paper $[7]$, the results are shown in Fig.2.

In order to further compare them, we use the equation (8) to compute the forecast precision of the two methods, forecast precision of normal BP algorithm is 0.12865, but that of recursive network is 0.08457 , the result indicates that recursive network is more adapted to the state forecast of MOA.

4. CONCLUSIONS

In this paper, a method is put forward to forecast the MOA's state based on the recursive neural network. The result indicates that recursive network is more adapted to the state forecast of MOA. Because the running state of MOA is closely related to the system voltage and the environment, the state forecast method affected by multi-factors should be further considered.

5. REFERENCES

[1] Li Xueshi et al., The finding analysis on high voltage metal oxide arrester's condition. Insulators and Surge Arresters, 1992(2).

[2] Linyi, The accident analysis and prevention on 110kv and upwards of MOA in China. Power System Technology,1995.(3).

[3] zhoulong et al., The foreast method of MOA's condition based on neural network, High Voltage Apparatus,1997(3).

[4] Werbo P.Ph Dissertation D.Committee on Applied Math. Harvard Univ. Cambridge, MA,1974.

[5] Ku C C,Lee K Y.System,identification and control using diagonal recurrent neural networks.Proc American Control Conf.Chichago,545~549,1992.

[6] Parlos A G et al., Application of the Recurrent Multilayer pereeption in Modeling Complex Process Dynamies, IEEE Trans. Neural Networks,5(2),2003

[7] Huazhong electrical power testing institute, The analysis on spoil reason of 500KV metal oxide arrester in Gezhouba.1994(11).

Zhou Long is a Full Professor , vice dean of Electrical Information & Engineering Department, Wuhan Polytechnic University. He graduated from Huazhong University of Science and Technology in 1998 and got the doctor degree. He has published two books, over 30 Journal papers. His research interests are in distributed parallel processing, fuzzy theory, neural network, grey system and its application.

Multiagent-Based Partner Selection of Dynamic Alliances in Inter-organizational Collaborative E-commerce*

Wang Jie [1 2], Shi Xingguo [2], Zhong Weijun [2]
[1] Laboratory of System Engineering
Department of Intelligent Science and Engineering, Nanjing University of Technology
Nanjing, Jiangsu 210009, China
[2] School of Economics and Management, Southeast University
Nanjing, Jiangsu 210096, China
Email: aiky903@126.com shixg@ec.js.edu.cn

ABSTRACT

Collaborative e-commerce has been considered as an effective waterway for underpinning advanced inter-organizational relationships. Dynamic alliance is one of novel organization structures that bring together individual entities located in an open and distributed environment temporarily for a specific goal, and combine their core competencies to improve the agility and flexibility of alliances in the global market. Agent technology provides a suitable enabler for achieving aims of dynamic alliances. Partner selection is a critical step in the success of a dynamic alliance. In this paper, a multiagent-based model to support the formation of dynamic alliances is proposed. The required attributes of the agents in the proposed model are explained. To address how to form a dynamic alliance rapidly and efficiently, the key issues in the process of partner selections are presented and discussed.

Keywords: Partner Selection, Dynamic Alliance, Intelligent Agent, Inter-Organization Collaboration, Electronic Commerce

1. INTRODUCTION

E-commerce encompasses business processes directly and indirectly related to the buying, selling and trading of products, services and information via Internet. In comparison with traditional commercial activities, the e-commerce transactions are carried out more rapidly and less costly in CSCW environment. Today inter-organizational collaboration and communication has become the kernel for B2B e-commerce, where business activities and processes are not limited within a single organization but involved in several cooperative organizations. Generally speaking, relationships between organizations are more complex and fluid than those between businesses and consumers since they involve the adoption of similar standards with respect to communications and collaboration, as well as joint information technology investment [1].

In particular, one of the main aims of B2B e-commerce is to significantly improve the supply chain by facilitating more efficient and agile procurement processes. As a consequence, new ways of working, new forms of organizations and new business models are emerging, such as electronic marketplace, electronic procurements, strategic sourcing, online auctions, strategic alliances, virtual enterprises, integrated supply chain networks and so on.

The remaining part of the paper is organized as follows: firstly related concepts and techniques are discussed. Then in the section 2 an MAS-based model of partner selection using e3-value methodology is introduced. Section 3 describes the agents and their attributes in detail. In the section 4 we give a detailed description of the process of partner selection of dynamic alliances. Finally conclusions and future research work are given.

Collaborative e-commerce

According to a white paper from Morgan Stanley Dean Witter [2], there have been three major phases in the evolution of e-business technology. EDI (Electronic Data Interchange) [3] networks represented the first phase of B2B e-commerce. The second phase was basic e-commerce, where retailers sell their products through their websites. Phase three of e-business, currently in progress, is in the form of communities of commerce, which bring trading partners with related interests together into a common community (i.e., an exchange) to match buyers and sellers and provide other services to serve their interests.

Currently, the distinct trend in the development of e-commerce, inter-organizational cooperation realized by advanced distributed computing and communication technologies, leads to the forth phase, collaborative e-commerce which enables enterprises going beyond the geographical, industrial and cultural boundaries to form virtual alliances with other enterprises to achieve the common business goals. Given the success of such collaboration, enterprises will increase the amount of resources accessible, gain the complementary abilities and become more competitive in the global market. Simultaneously due to the rapidly changing characteristics of the market, the structure and duration of virtual alliances also vary.

As noted in [4], inter-organizational processes have two distinguishing features. Firstly, the resources needed for a process cannot be assigned centrally as they reside in different organizations. Secondly, the organizations involved in a process have a certain degree of autonomy meaning that no central authority has control over all the members. In a nutshell, dynamic, autonomy, provisionality, interaction and distributed cooperation are among the key attributes of collaborative e-commerce. These new features imply that traditional e-commerce technologies, which require a direct communication between the potential buyer and seller, are not completely effective in open and distributed environments.

* This work is supported by National Science Foundation of China under the grant No. 70171025 and Research Project Grant of JiangSu under the grant No. 02KJB630001.

Dynamic alliance and its lifecycle

As a new form of collaborative e-commerce, dynamic alliance can be described as a goal-oriented and commitment-based group of semi-autonomous geographically distributed entities with a limited lifetime. Such independent entity that can be an organization, individual, human beings or a software agent attempts to maximize its own profits as well as contributes to defining and achieving the overall goals of the alliance. The entities that constitute the DA are the partners of the DA. The partners cooperate to fulfill a set of specific goals. Once the goal is achieved, the alliance can either disseminate or evolve by changing its goal. Unlike traditional organizations, DAs do not have a rigid and permanent organizational structure. Consequently, the success of the DA is strongly dependent on the commitment, the performance and the delivery capabilities of its partners. For example, an individual company may collaborate with several partner companies that provide related products so that each of them need only provide the services or products in which they specialize, but, when taken together, the DA can provide a broader range of offerings. [11]

The lifecycle of DA are composed of four stages that are strategic market planning, alliance formation, alliance operation and management, and alliance dissemination. Due to its volatile and resilient characteristics, one of the most important stages in the lifecycle of the DA is the formation of the DA. Customer requirements or market demands stimulates to form a new dynamic alliance, and in most cases, an initiator is responsible for its formulation. Since DAs have a limited lifetime, they need to be formed rapidly enough to meet the deadlines of the goals. An important activity in the formation of the DA is the selection of partners.

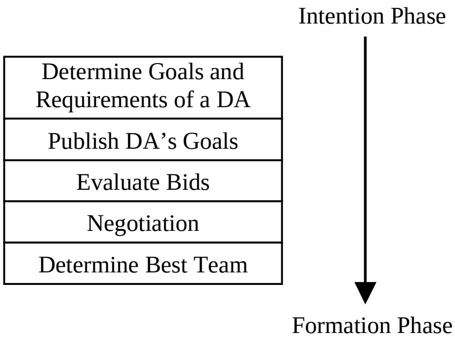


Fig.1 The Formation of a Dynamic Alliance

Fig. 1 shows the formation stage of a DA within a lifecycle context. Before a DA is formed, it's concepts and goals have to be defined. The requirements from the customer sets the requirements for the DA team and in order for the DA to be able to deliver to its customer, the right team has to be formed. During the formation stage of a DA, the individual entities compete and negotiate to become the partners of the DA. When the DA is formed, the partners that have been selected constitute the DA and work together to deliver to the customer.

The formation of a DA involves a selection process based on multiple variables such as organizational fit, technological capabilities, relationship development, quality, price, and speed [5] In this paper, we focus on partner selection in a DA.

Multiagent system in dynamic alliances

Agents are computer systems capable of performing tasks autonomously in complex and dynamically changing environments, without the direct intervention of users. Most agents possess some degree of at least one of the following properties:[6]

1. Mobility: A computer code is *mobile* if it is capable of making copies of itself, from one site to another, over a network.
2. Intelligence: the capability to interpret, learn and improve. Every form of intelligence which could improve the agent performance, excluding social intelligence
3. Agency: The ability to interact with other agents, ranging from "naive" to "strategic" exchange of messages such as negotiation skills.

A Multiagent System (MAS) consists of a group of agents that combine their specific competencies and cooperate in order to achieve a common goal. Efficient cooperation as well as coordination procedures between agents endows a MAS with a capability higher than the sum of capabilities of individual agent. Martinez et al. [7] propose a multiagent control system that consists of three kinds of controller agent: product agents (which manage the activity associated with each product), activity agents (which autonomously manage an entire manufacturing activity), and resource agents (which manage their own operative functions and propose service offers to activity agents). Together, these agents use and control the other entities in the system in order to achieve the overall aims. The MASSYVE project focuses on the use of multiagent systems in agile scheduling in a virtual environment [8].

Considering the similarity of characteristics between agents and partners of dynamic alliance, MAS technology is a natural paradigm for modeling individual organizations in dynamic alliances. Different organizations are represented by intelligent agents with distinct attributes and variables. In a DA, a primary agent performs the process of partner selection on the basis of specific rules.

2. E^3-VALUE BASED MODEL OF PARTNER SELECTION

Base on the relevant analysis in the first section, it is clear that partner selection is a both fundamental and cardinal step in the whole life cycle of a dynamic alliance. So in the paper, we focus on the selection process in DA formations.

An important part of an e-commerce application development is to design an e-business model. Such a model should show the business essentials of the e-commerce business case to be developed.

Partners are selected by their skills and resources to fulfill the requirements of the DA. Since all the partners have to work as a team, these requirements must address not only the single partners of the DA, but also how the partners fit into a team [9]. In developing a model of partner selection, there are at least the following three issues that must be taken into account:

1. An organization agent that is considering whether or not to join a DA must determine the conditions under which it is profitable.
2. An initiator agent must be able to recognize circumstances in which it should initiate VO formation.

3. The initiator agent must determine the best combination of business partners.
4. To determine the kind of information that is flowing among the different entities is helpful to design the agents and the communication and collaboration among the agents.

Many attempts have been taken on the area of ontology and model development for e-commerce system design and implementation. e3-value methodology provides an ontology to conceptualize and to visualize an e-business idea. Its concepts include actor, value activity and value objects and value exchange. [10] On basis of e3-value ontology, a model for partner selection is developed as shown in Fig 2.

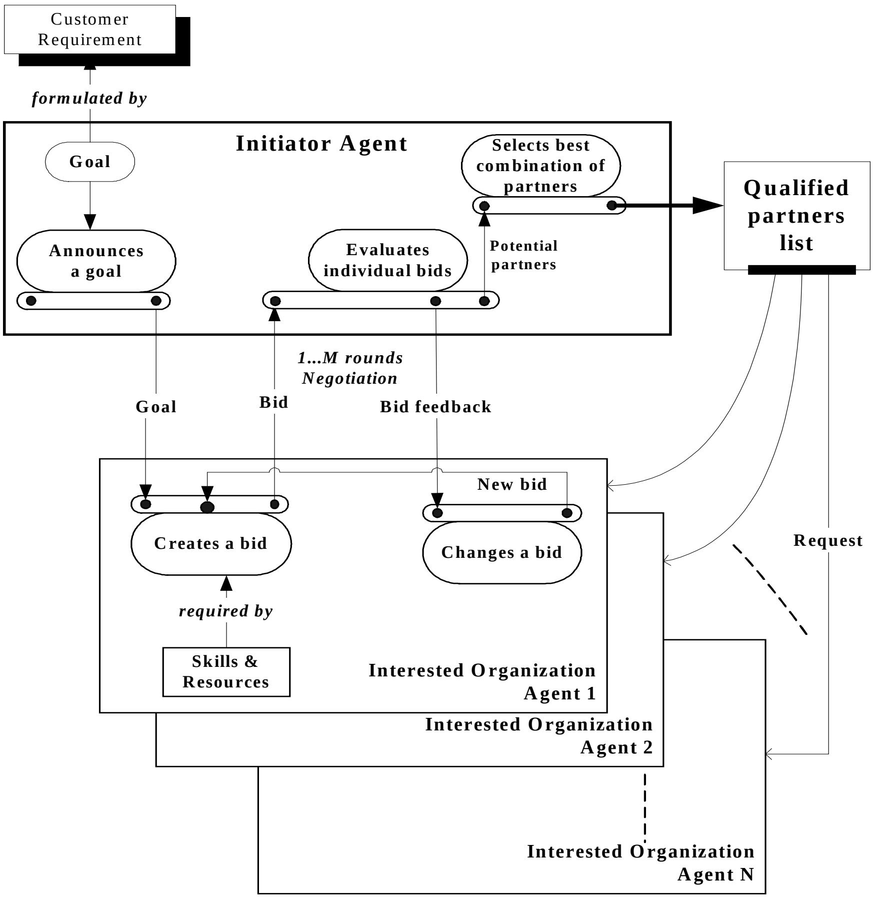


Fig.2 An e3-value based MAS Model for Partner Selection of a dynamic alliance

The model shows entities and how they relate to each other. It includes several intelligent agents representing independent organizations that are interested in participating a DA. A special agent, named initiator agent, acts as the role of initiator to start and manage the selection process by using the internal activities such as goal announcement, bid evaluation, best combination selection. Multiple rounds negotiation mechanism is adopted to determine potential partners from interested partners. Then an optimal combination of a alliance is created.

The information flows among agents are represented by value object, including goal, bid, feedback, request, skills and resources, partner list. The relationship of objects is denoted in predicate calculus.

3. AGENT MODEL DECRIPTION

The agents are described by a set of attributes and these attributes form the basis for the evaluation of the agent as a partner in the DA and during the selection of the DA team. Instead of considering the complete attributes of an agent, here we only consider the attributes that are required for the agents to propose a bid and negotiate to become a partner.

Attributes of agents

The main attributes of an agent that are used to recognize it are its name and the IP address. In addition to these, an agent has goals and an ontology for communication. As shown in Fig.2, the agent has relationships with several other entities such as activities, actors and value objects. The relationships are represented as predicate calculus clauses as Table 1.

Table 1 Attributes of an Agent

Predicate Caculus	***Description***
has (agent, set of goals)	An agent has goals
assigned_to (goal)	An agent is assigned to a goal
assigned(activity)	An agent is assigned to an activity
has(agent, set of skills)	An agent has skills
performed_by(role, activity, agent)	An agent performs a role in a DA
member(team, agent)	The affiliation of an agent
available(agent,start time, end time)	The time period an agent is available for work
cost_per_hour(agent, price)	The cost an agent expects to be paid for each hour of work.
Risk(agent, cost)	The risk to accept an agent into a DA (the cost for accept an agent-the commitment breaking cost)

XML description of an agent is as follows:

```
<?xml version="1.0" ?>
   <organization agent>
      <agent id="agent101">
      <name>Ford Auto</name>
      <description>automobile
manufacturer</description>
      <IP_address>202.119.248.66</IP_address>
      <skill>programming</skill>
      <cost_per_hour>44.95</cost_per_hour>
      <available_start>2004-10-25</available_start>
      <available_end>2005-10-25<available_start>2000-
10-01</available_end></available_end>
      <commitment_breaking_cost>500</commitment_
breaking_cost>
      <cost_for_accept>2000</cost_for_accept>
   </agent>
   <agent id="agent201">
      ……………..
   </agent>
………………….
</organization agent>
```

An agent representing the Initiator Agent is described by the attributes shown in Table 2.

Table 2 Attributes of an Initiator Agent

Attribute	***Description***
Goal	The goal of a DA
Availability	The time period that the partners are required
DA requirements	The skills and other information that are required by the DA and the constraints on these attributes
Deadline	The closing date of bids

The skills of an agent are described as a attribute set that has constraints. Each agent has one or more skills, each of which can be described by the attributes: skills, number of years of experience, and skill rating. Examples of some constraints for these attributes are the minimum no. of years of experience that is required for a skill or the lowest acceptable level of skill rating for a skill.

4. PARTNER SELECTION OF DYNAMIC ALLIANCES

Selection process

As shown in Fig.2, selection process consists of four phases. The first phase is goal formulation, by an Initiator Agent, a goal of the DA on the basis of customer requirement. Then Initiator Agent publishes or announces the goal of the DA to all the other organization agents. It follows the second phase, individual bid formation, where if an organization agent finds participation into the DA is profitable and has intention to compete with other agents, it will become an Interested Agent. The third phase is individual bid evaluation, when interested agents send their bids, which match their skills and resources, to the Initiator Agent respectively. The Initiator Agent has the activity of evaluation to select proper bids according to some constraints by a certain specific algorithm. Interested Agents, whose bids are qualified if the goals are reached and a minimum constraint is matched, now become Potential Agents. The output of the evaluation phase is a rank of all the bids. The bids that are disqualified are informed of their failure. The best set of partners selected by considering individual bids may not necessarily be the best team. Then Initiator Agent prepares to negotiate with the Potential Agents.

Therefore, a fourth phase, optimal team selection, is included where the Potential Agents are considered as a team during the selection. The first three phases are mainly on individual agents while the fourth phase emphasizes on partner teams and the evaluation is based on different criteria from those for the individual bids.

Exchange of value objects

Value objects are exchanged between the Initiator and other agents. A new intention to establish a DA is published in order to invite interested agents to bid. The announcement should contain information including the goals of the DA, the skills and resources required for the DA, the time period of the DA, and the deadline for bids.

Interested Agents respond to the announcement by submitting a bid, which contains the information on the goals of the partner and list of attributes and their values.

In addition to the above information, both the announcement and the bid will carry the name, identification and address of the sender and receiver agents.

Bid evaluation

Given the fact that agents have many attributes as described above, generally bid evaluation and negotiation is a multiple-criteria decision problem. The evaluation methods suggested in the related literatures can be classified into two categories: qualitative and quantitative. The traditional methods include analytic hierarchy process, multi-objective programming, linear programming or mixed integer programming, activity based costing and Bayesian classifiers. Recently new techniques have been applied to address the problem, such as neural network, fuzzy set and machine learning.

The weighting model, one of the most popular methods, is applied here in order to give an example. Each attribute is weighted and a multi-attribute utility function is defined as follows:

Utility Value = Σ (attribute value * weight)

First, the attribute values need to be normalized before the calculation. Each qualified bid is checked to decide whether they meet the constraints or not. If the values do not meet the constraints, then they are assigned the value zero.

```
/* calculate utility value */
for (int i=1; i<=num_of_agents; i++)
{
        for (int j=1; j<= agent(i).num_of_attributes; j++)
        {
                UtiVal(i)=0;
                AttiVal=agent(i).attribute(j);
                if attiVal not meet limit(j) then attiValue=0;
                else UtiVal+=attiVal*weight(j);
        }
}
```

The utility values are calculated for all the qualified bids and the values are ranked, where the highest utility value is at the top. This list is then submitted to the Initiator Agent. The Initiator can choose the best (highest ranked) Potential Partners. Given the different conditions, the utility function can be changed by choosing a different set of attributes and by changing the weights that are assigned to the attributes.

Team Selection

When selecting qualified partners from a global resource pool, how to determine the optimal team under a particular condition? The initiator agent must determine the best combination of business partners. The concept of a team is an important point in forming a DA as the partners have to collaborate and work together in a team to achieve the goals of the DA. Therefore, we consider the selection of the team as a separate phase in the partner selection and consider the attributes of a team rather than the attributes of an individual in the utility function to determine the best team of partners.

The selection of the best team can be based on several criteria and the best team may not always be the team that consists of the highest ranked Potential Partners. For example, a DA may have constraints such as a total budget that the Initiator can pay its partners. There may be other such constraints such as the total cost of the partners in a team and the total risk of having the partners in the team.

5. CONCLUSIONS AND FUTURE RESEARCH

This paper is a step towards incorporating agent technologies into the collaborative electronic commerce. Compared to other models in the literatures, the model presented in the paper is the better one if considering the partner selection from viewpoint of both business and process aspects. Not only are value objects denoted, but the necessary processes and activities are also included in the model. More work need to be done towards negotiation and collaboration models of intelligent agents. In the meantime, there is a need for further investigation and research on multi-variable team selection that is the same significant as individual bid evaluation. We believe that a multiagent architecture is a proper solution of collaborative e-commerce applications. The further enhancement of the model will be done based on real industry cases.

6. REFERENCES

[1] M. Subramani, E. Walden, “Economic Returns to Firms from Business-to-Business Electronic Commerce Initiatives: An Empirical Examination,” Proc. 21st Int’l Conf. Information Systems, pp. 229-241, 2000.

[2] C.Phillips, M.Meeker, “The B2B Internet Report: Collaborative Commerce” Morgan Stanley Dean Witter, April 2000

[3] [Ada98] Adam, N., Adiwijaya, I., and Atluri, V., “EDI through a Distributed Information Systems Approach,” Proceedings of the 31st Hawaii International Conference on System Sciences, Kohala Coast, Hawaii, USA, January 1998, http://www.computer.org/ proceedings/hicss/8233/8233toc.htm

[4] Weigand H., van den Heuvel W. and Dignum F.: “Modeling Electronic Commerce Transactions – A Layered Approach”, Third International Workshop, The Language Action Perspective on Communication Modeling, eds. G. Goldkuhl et.al. 1998.

[5] J. Sarkis and R.P. Sundarraj, “Evolution of Brokering; Paradigms in E-Commerce Enabled Manufacturing,” Int. J. Production Economics, vol. 75, 2002. pp. 21-31

[6] O’Hare G. M. P., Jennings N. R., Foundations of Distributed Artificial Intelligence, John Wiley & Sons, 1996

[7] M.T. Martinez, P. Fouletier, K.H. Park, and J. Favrel, “Virtual Enterprise—Organisation, Evolution and Control,” Int’l J. Production Economics, vol.74, 2001.pp. 225-238

[8] R.J. Rabelo, L.M. Camarinha-Matos, and H. Afsarmanesh, “Multi-Agent-Based Agile Scheduling,” Robotics and Autonomous Systems,vol.27,1999. pp. 15-28

[9] D.E. O’Leary, D. Kuokka, and R. Plant, “Artificial Intelligence andVirtual Organizations,” Comm. ACM, vol. 40, no. 1, Jan.1997. pp. 52-59

[10] J. Gordijn. Value-based Requirements Engineering: Exploring Innovative e-Commerce Ideas. PhD thesis, Free University Amsterdam, NL, 2002. via http://www.cs.vu.nl/~gordijn/thesis.htm

[11] L. Meade, J. Sarkis, and D. Liles, “Justifying Strategic Alliances: A Prerequisite for Virtual Enterprising,” OMEGA: The Int’l Management Science, vol. 25, no. 1 , 1996.pp. 29-42

Agile Reconstruction Methods Based on Agent in Distributed Database *

Qu YouTian Xu Hong
Institute of Computer Science Studies, Zhejiang Normal University
College of Information Science and engineering, Zhejiang Normal University
Jinhua city, Zhejiang province, 321004, P. R. China
Email: quyt@mail.zjnu.net.cn, **Tel:** 0086-0579-2283410

ABSTRACT

Collaborative design has become a research topic mainly in the domain of Architecture Engineering and Construction (AEC) . Data sharing is a basic requirement in collaborative AEC design. In distributed database environment, a new information infrastructure and data sharing method by agent based agile database reconstruction is proposed. The method can retrieve all partners' data information and can reconstruct a virtual distributed database with it, intelligently, flexibly and effectively. And it makes the agent based agile software development possible for the virtual enterprise.

Keywords: Agile Reconstruction; Agent; Distributed Database; Collaborative Design; Virtual Enterprise

1. INTRODUCTION

Agent technology is a recently emerging branch of distributed artificial intelligence research [1]. There are widespread and growing concerns on the potential of Agent technologies which can be used to build the next generation of advanced software systems. The concern is especially focused on domains such as Internet-based applications and problems requiring advanced user interactions and flexible collaborations. The term Agent is gaining considerable attention in the software engineering and information technology community. So-called intelligent, or autonomous, agents are reported capable of many advanced forms of behaviors. Advocates of agents claim many benefits from the use of agents. Agents and related technology have implications for software engineer*ing. As autonomous software entities, agents play important roles in software systems [2]. The advanced agent communication mechanisms further represent ways to structure societies of agents that need to exchange information and cooperate [3]. Agents are therefore something a software engineer must begin to consider as the technology transitions from the research labs to products.

Collaborative design has become a research topic mainly in the domain of Architecture Engineering and Construction (AEC). Many prototype systems as well as mechanisms have been proposed to support collaborative design [4]. Current database systems in collaborative AEC design have two architectures: the central model and the distributed model. The central architecture has a virtual central database that is accessible by multiple partners. An example is the DICE system (Distributed and Integrated environment for Computer-aided Engineering) developed at MIT (Sriram *et al.*, 1991). DICE can be envisioned as a network of computers and users, where communication and coordination is achieved through a global database and a distributed control mechanism. Current AEC design systems only put emphasis on the collaborative designs. However, in agile manufacturing, approaches are needed to support agile partner collaborations not only in product design, but also in process planning and resource management. In AEC domain, building global databases and maintaining distributed database static constraints are the main approaches. Our work provides an integrated framework based on agents, which support collaborative product development across domains and companies. Constraints among distributed partner databases are maintained dynamically, flexibly and intelligently.

Data sharing is a basic requirement in collaborative AEC design [5]. Sometimes, data sharing among all partners can not realize simply, new technologies should be produced to solve it. In this paper we propose a data sharing method which reconstruct distributed database based on Agent.

2. DATA AND AGENT ARCHITECTURE

2.1 Three-Tier Data Architecture

Today's state of the art distributed database systems are logically organized as multi-tier, component-based software systems [6]. Typical generic approach is the three-tier architecture [Bers96] in which distinction among the following three tiers exists:

- Database management tier
- Process management tier
- User Interface tier

Similar to this three-tier software architecture the three-tier data architecture for distributed database systems is proposed here. Three-tier data architecture is shown in Figure 1:

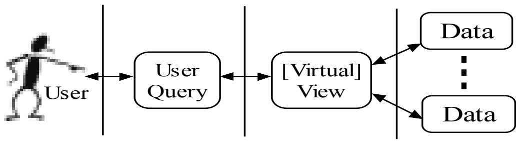


Figure 1 Three-tier data architecture

- Tier shown on the right is the Data tier. It represents the data stored in the distributed environment i.e. different computers and which is available through network infrastructure. Data tier will also be referred to as Data Source(s).
- Middle tier is the Virtual View tier. The goal of virtual view tier is to hide the geographical and logical distribution of data among Data sources, by providing a new and unique view over that data. Virtual view is created corresponding to specific requirements of some application. There can be a lot of virtual views over single group of distributed data sources.
- Tier on the left is the User Query tier. Through

* This work has been supported by the National Science Foundation under grant M603245 and key specialty of computer application, ZNU

this tier user can ask ad-hoc queries over one or several virtual views.

All of the three tiers in three-tier data architecture have to handle data. Each tier needs to access the data from the previous tier and finally the user will access the data from user query (client) tier. Thus, each tier serves the data to the next tier. Data model used to represent the data which is served from tier x to the tier x+1 is called Access Data Model. Access data model will be used here to classify the existing distributed systems.

2.2 Utility-Based Agents

Goals alone are not really enough to generate high-quality behaviors [7]. For example, there are many action sequences that will get the Agent to its destination, thereby achieving the goal, but some are quicker, safer, more reliable and effective, or cost cheaper than others. Goals just provide a crude distinction between “success” and “fail” states, whereas a more general performance measure should allow a comparison among different world states (or sequences of states) according to exactly how successful they would make the agent if they could be achieved. Because “success” does not sound very scientific, the customary terminology is to say that if one world state is preferred to another, then it has higher utility for the agent.

Utility is therefore a function that maps a state onto a real number, which describes the associated degree of success. A complete specification of the utility function allows rational decisions in two kinds of cases where goals have trouble. First, when there are conflicting goals, only some of which can be achieved (for example, speed and cost), the utility function specifies the appropriate trade-off. Second, when there are several goals that the agent can aim for, none of which can be achieved with certainty, utility provides a way in which the likelihood of success can be weighed up against the importance of the goals.

An agent that possesses an explicit utility function therefore can make rational decisions, but may have to compare the utilities achieved with different courses of actions. Goals, although cruder, enable the agent to pick an action right away if it satisfies the goal. In some cases, moreover, a utility function can be translated into a set of goals, so that the decisions made by a goal-based agent using those goals are identical with/to those made by the utility-based agent. The overall utility-based agent structure appears in Figure 2.

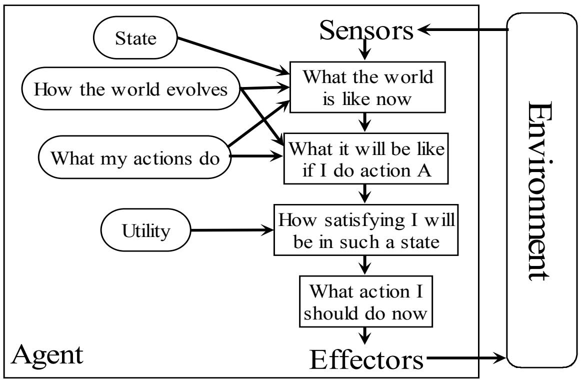

State
Sensors
How the world evolves
What the world is like now
What my actions do
What it will be like if I do action A
Utility
How satisfying I will be in such a state
What action I should do now
Agent
Effectors
Environment

Figure 2 A complete utility-based agents

An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors. An Agent has encoded bit strings as its percepts and actions [8]. Our aim in this paper is to design agents that do a good job of acting on their environment.

In our opinion, an agent is a piece of software that detects and reacts on changes in its environment. An agent is able to operate autonomously, whereby its goals are explicitly given. A multi agent system is a software system build by a number of agents which communicate in order to solve more complex problems.

According to those definitions potential applications must meet three criteria for applying multi agent technologies:

- . natural distribution,
- . dynamic world, and
- . complex interaction.

3. THE INFRASTRUCTURE OF AGENT BASED DISTRIBUTED DATABASE

Dynamic collaboration is product-oriented which is created at the beginning of product’s life cycle and dissolved while the market opportunity disappears by the organizer [9]. Different from traditional enterprise where organization structure is strictly hierarchical, a dynamic collaborative enterprise is not only hierarchical, but also flat. The sponsor of the virtual enterprise has the responsibility and obligation to manage the collaborative enterprise. Meanwhile, all the other partners must be coordinated by the organizer. On the other hand, each partner has sufficient independence and autonomy to manage its own company with individual business goals and can have different management mode and organization method. In a virtual enterprise, it needs to exchange product data and knowledge frequently and dynamically which is segmented and distributed across the distributed sites of partners. Information sharing between distributed database and consistent information providing to distributed partners are essential to support agile product development [9].

The data hierarchy represents the agile manufacturing information in a flexible, object-oriented, hierarchical structure, so that the agile partners can dynamically share the information in different views and at different abstraction levels. The basic idea is to provide Dynamic Object Views from primitive information storage. Focusing on the object framework development, we emphasize on the hierarchical structure as well as the methods to form the data hierarchy.

These databases are connected and managed by a certain scheme so that they satisfy the agile manufacturing information system requirements as a whole unified database. The database hierarchy has the following properties:

1) Flexible and dynamic object construction and sharing schemes. The virtual enterprise data hierarchy is composed of individual partner information hierarchies, which aggregate together to form a complete enterprise information representation. The primitive information is stored at distributed partner sites, the atomic objects are constructed from local storage, the composite objects are formed by using other atomic objects and composite objects located at local or global sites. As shown in Figure 3, there are three layers in each partner site: the bottom is primitive information storage, the middle is atomic objects, and the top is composite objects. A

composite object in partner A can be formed from the atomic objects from partner A, B and C, and the composite objects from Partner A, B and C.

2) Multiple object views. Based on the distributed primitive manufacturing information, atomic objects and composite objects are built from different partner views. For example, a product is an assembly of engineering drawing and specifications in the product designer's point of view, it is an aggregation of primitive processes from the process planner's perspective, and for an MRP user, a product concerns the work schedule, the inventory and the purchase orders. Because of the partner information exchangeability, our framework allows the formation of multiple object views from the same primitive storage.

As we have discussed above, each Agents of the partners can retrieve the cooperate information intelligently which the products produce needed, about the useless or unimportant information the Utility Agent will ignore, and the Agent Federation will reconstruct the virtual database according to the information which the partner agent has retrieved.

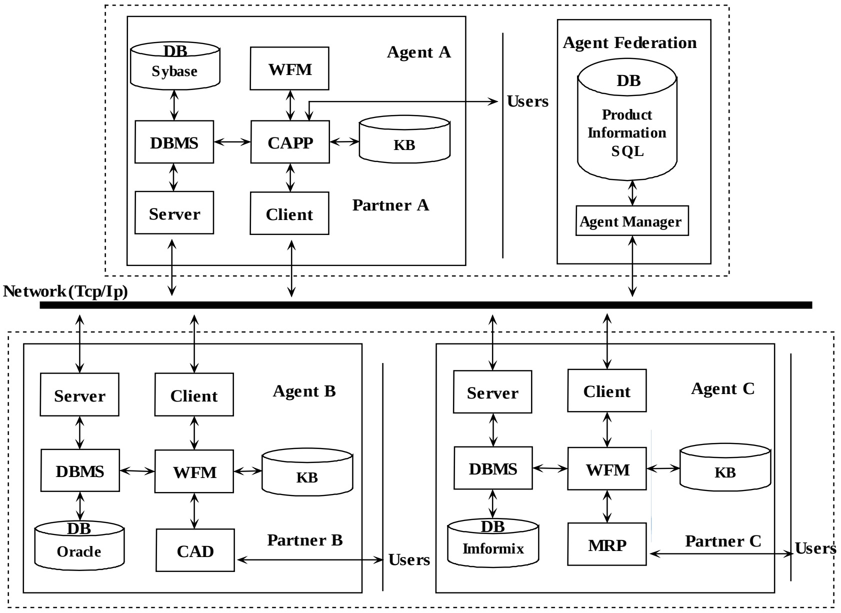


Figure 3 The infrastructure of Agent based distributed database

4. THE PROCESS OF RECONSTRUCTION

Once we realize that an agent's behavior depends only on its percept sequence to date [8], then we can describe any particular agent by making a table of the actions it takes in response to each possible percept sequences. Such a list is called a mapping from percept sequences to actions. We can, in principle, find out which mapping correctly describes an agent by trying out all possible percept sequences and recording actions which the agent does in response. And if mappings describe agents, then ideal mappings describe ideal agents. Specifying which action an agent ought to take in response to any given percept sequence provides a design for an ideal agent.

We have discussed the Partner Query Language (PQL), which is used by individual partners and reflects the logical view of an aggregated database, and the Partner Workflow Language (PWL), which is constructed by the system and reflects the logical view of relationships between distributed partner database systems. The mapping from PQL to PWL is performed by consulting the system data dictionary, and the knowledge base consisting of partner information sharing policies. By interacting with the knowledge base, the workflow manger is able to convert partner queries into workflow queries, and manage the local client to execute the workflow in a consistent way.

In the rest of the paper the special attention will be paid to the position of Agile data within distributed database systems of different types. If we have retrieved the agile data information, then we can reconstruct the distributed database easily. The utility Agent or an agent federation can do an iteration Top-down design process as following:

1) Requirement analysis
2) Conceptual design and view integration E-R representation and translation to RDB schema
3) Distribution design Data fragmentation and allocation
4) Physical design
5) Tuning

5. EXAMPLE FOR APPLICATION IN RECONSTRUCTION

Supposing that partner A undertakes a project which should be supported by partner B for the completion of the project. Because partner A has got the information about MRPII and partner B has got the information about CAD/CAM, they form a Agent organization which is also called virtual organization .Supposing that the data granularity is chosen as the Table, Agent Federation can extract n Tables from database A and database B to form a Virtual Database among which Table 1~ Table k come from partner A and Table k+1 ~ Table n come from partner B according to the requirement of the project. With the virtual database, developers can apply CBD (Component Based software Development) in the development of the MIS of virtual organization.

Of course, during the process of reconstruction, each Agent can choose its own data granularity according to the requirement of the project which can be a new relationship formed by extracting some data fields from the original Table or a virtual database of greater scale established by separate partner's own database of coarse data granularity.

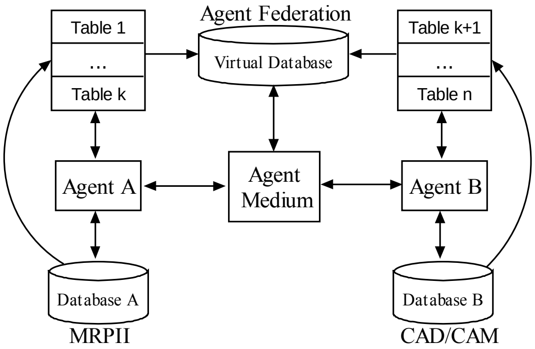


Figure 4 Agent Reconstruction application

6. CONCLUTION

Based on the information requirements in an agile enterprise, developing a virtual agile manufacturing information system depends on the establishing of the agile distributed database. Of course, the agile reconstruction methods based on agent in distributed database which we have proposed in this paper, will not fit for heterogeneous data source immediately, but we can use the recent technology such as CORBA or ODBC to solve it. However, there are some problems need to be solved. Such as how to determine the size of data granularity, and how to evaluate the effectiveness of agile reconstruction.

7. REFERENCES

[1] Brian Henderson-Sellers, , Agent-based Software Development Methodologies University of Technology, Sydney (Australia) and Ian Gorton, Pacific Northwest National Laboratory (USA)

[2] Bresciani, P. and Giorgini, P. (2002). The TROPOS analysis process as graph transformation system, in Proceedings of the OOPSLA 2002 Workshop on Agent-Oriented Methodologies (eds. J. Debenham, B. Henderson-Sellers, N. Jennings, and J.Odell), COTAR, Sydney, 1-12

[3] Bresciani, P., Giorgini, P., Giunchiglia, F., Mylopoulos, J. and Perini, A., 2003, Tropos: an agent-oriented software development methodology, J. Autonomous and Multi-Agents (in press)

[4] Liugen Song, Rakesh Nagi, Design and implementation of a virtual information system for agile manufacturing .1997, 29 (10): 839-857

[5] Debenham, J.K. and Henderson-Sellers, B. (2003). Designing agent-based process systems –extending the OPEN process framework, chapter in Intelligent Agent Software Engineering (eds. V. Plekhanova), Idea Group Publishing, 160-190

[6] Marko Smiljanić, Henk Blanken, Maurice van Keulen, Willem Jonker, Database group Distributed XML Database Systems Twente University Faculty of Informatics October 2002

[7] Artificial Intelligence: A Modern Approach by Stuart Russell and Peter Norvig, © 1995 Prentice-Hall, Inc, 31-46

[8] Cabri, G., Leonardi, L. and Zambonelli, F. (2002). Modeling role-based interactions for agents, in Proceedings of the OOPSLA 2002 Workshop on Agent-Oriented Methodologies (eds. J. Debenham, B. Henderson-Sellers, N. Jennings, and J. Odell),COTAR, Sydney, 13-20

[9] Xu Qingsong,Fan Yushun,Tao Dan, Research of an agile manufacturing information system, Systems Engineering and Electronics, Vol.11, No.3,2000,22-27

The Applications of Multi-agent in an Expert System –AVDDT*

Yang Kaiying
Wuhan University of Technology
Wuhan, 430070, China
Email: yangky@mail.whut.edu.cn **Tel:** 027-62052953

ABSTRACT

The paper introduces the applications of multi-agent distributed AI technology in diagnosis and treatment system of acute ventral disease (AVD). Based on the complex correlated analyses for the treating process, multi-agent's structure and the functions of each layer is constructed. Whereafter, the applications of blackboard model are discussed as a key part in this multi—agent system's cooperative communication subsystem.

Keywords: Multi-agent, Expert System, Acute Ventral Disease, Distributed AI.

1. INTRODUCTION

Acute ventral disease (AVD) is a kind of common disease, such as acute pancreatic inflammation, acute stomach perforation, biliary tract worm, acute appendicitis, etc. It's acute and shows complex symptoms. Sometimes it will bring great pains and life crisis to patients. Developing an expert system to diagnose and treat AVD make it possible to cure AVD rapidly and accurately.

The AVD has many special characters, so many means of diagnoses and treatments should be used simultaneously and comprehensively. The expert system (ES) to cure AVD is a typical distributed parallel system. In recent years, multi-agent technology became more and more popular, it affords an efficient way to achieve this system.

2. AGENT AND MULTI-AGENTS

In distributed computing region, people usually name a autonomous program as "Agent" which has the ability to control it's own decisions and actions .It can achieve one or more purposes based on its understanding about environment in distributed system. Agent possess four the most critical characteristics: Autonomy, Social ability, Reactivity and Pro-activeness [1].

(1) Autonomy: Agent own interior self-controlling mechanism and problem-resolving mechanism. It can make justify and act according to it's own knowledge and captured messages without intervention from outside.

(2) Cooperated ability: Agent isn't isolated but in a group with interaction. All Agent could communicate and talk with each other using a certain protocol or language, then complete a certain mission with cooperation.

(3) Reactivity: Agent possesses function of reflecting external environment, and it can identify changes in external environment and make relevant reaction.

(4) Pro-activeness: It can feel changes of surroundings, and take actions for it's aims.

Multi-agent System (MAS) is composed of many Agents. It is distributed both in structure and organization and each Agent has intelligence. So the system possess characteristic of both distributed system and AI. It represent at three points: (1) Many Agents combine to a incoherent and cooperative combination, which can finish more complicated mission than individual agent dose and achieve more complicated functions. (2) Each agent only complete some basic missions owing to each agent in it can work in cooperation. The complexity of single Agent is brought down. (3) Allocations of the MAS system are flexible. Each agent in it is autonomous. . So changes of some Agents will not influence others' functions.
MAS is available to abase the hardness of cooperating resolution about complex problem, achieve the advantages of expert system's distributed handling, and improve it's abilities to handle emergencies. It is an ideal selection to develop expert system for acute ventral disease (AVD) diagnosis and treatment (AVDDTES).

3. STRCTURE AND FUNCTIONS

Cure AVD is a complicated process, We can get basic structures and functions of this system according to analyses of the process.

3.1 Analyses of the process to cure AVD

In this Process, there are no cause-and-effect relations but more complex ones among symptoms、diseases、means of diagnose and treatments . For example, different disease could appear to similar symptom, causes of homogeneous disease and means of treatment are different too. Many means of diagnosis are usually demanded, and treatments should be made with diagnosis at the same time. The process can be abstracted as the following Fig.1:

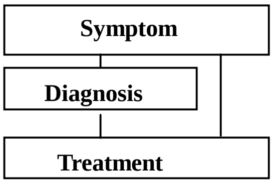


Fig. 1 handling process

Diagnosis and treatment are main parts of the process.
Diagnosis and treatment are main parts of the process. The diagnosing process that caused by symptoms of disease is the kernel of it. There are not only complex relations between symptom and disease but also ordering and parallel relations of many methods of diagnosis. **Fig.** 2:

* This paper is supported by S&T project of Wuhan, China

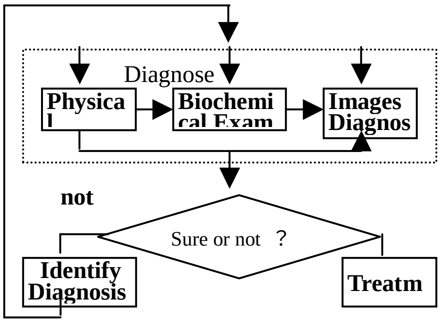


Fig. 2 The simplified diagram of the diagnosing process

Following are the relations produced from induction:
(1) The portrait relation of the symptoms diseases and the causes of diseases.
(2) Cooperation of many means of diagnoses.
(3) Discrimination of similar diseases.
Furthermore, there are complex relations not only in diagnoses and treatment but also in interior processes of treatment. AVDDTES based on multi-Agents can well realize this complicated system.

3.2 Structure of AVDDTES
AVDDTES is composed of some agents, each agent has its own knowledgebase and blackboard to realize communication and cooperation. Can be described as following **Fig.** 3:

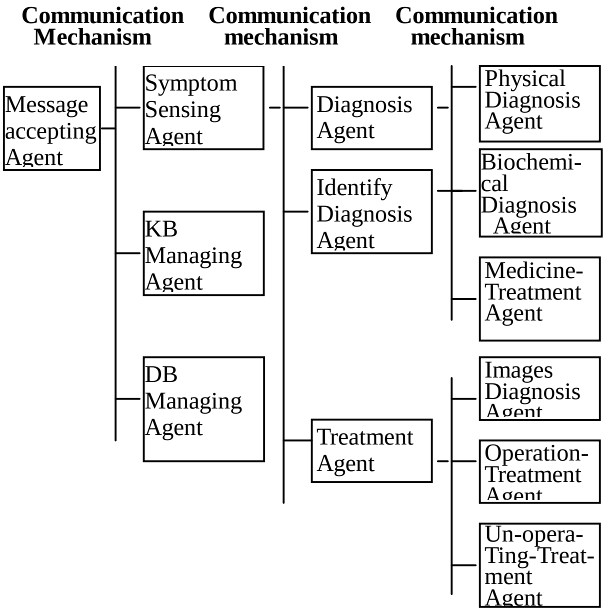


Fig. 3 The structure of Multi-Agent

(1) **Message accepting agent** mainly takes charge of the ES and interact with external environment. It is composed with Person-to-computer Interaction Agent, Date- Accepting Agent, Knowledge Accepting agent and Interpreting Agent.
(2) **DB managing agent** mainly store and manage description of messages. It stores messages from Message-accepting Agent to related Database based on it's own knowledge Library.
(3) **KB managing agent:** It mainly store and manage knowledge. The ES's knowledge Bases are classified into Public KB and Private KB. There are usually many Private KBs, So the KB managing Agent can not only supply traditional KB but also achieve functions that distribute knowledge into each Public KB or Private KB.
(4) **Symptom sensing agent:** It should be separated from Message accepting Agent because of the importance of Symptom sensing and close relations between Diagnosis Agent and Treatment Agent. It mainly accept updating dynamic symptom messages and interact with Diagnosis Agent and Treatment Agent through communication mechanism based on it's own knowledge library
(5) **Diagnosis agent:** It is the Decisive component of the system. Firstly, It divides diagnosis tasks according to messages from Symptom sensing agent. It will awake one or more agents in sub-layer and estimate the result. If an available diagnosis resolution can't be got, it will coordinate the whole task. Secondly, It needs to arbitrate conflicts in sub-layer agents.
(6) **Identify diagnosis agent:** Its functions are similar with Diagnosis Agent. Identify diagnosis agent is separated from Diagnosis Agent to abase its complexity. Except for different knowledge Database ，the main distinction between Identify Diagnosis Agent and Diagnosis Agent is that Identify Diagnosis Agent can infer more efficiently using time-ordering evidences.
(7) **Treatment agent:** It is the functional component of the whole system. It can be awaked by Symptom sensing agent or Diagnosis Agent. It cooperates with Diagnosis Agent when it's awaked by the Symptom Sensing Agent. Decisive mechanism used in sub-layer of Treatment Agent is similar with Diagnosis Agent.

4. COMMUNICATIONS AND COOPERATION MECHANISM OF AVDDTES

Agent communication is kernel technology in Agent Cooperation. Agent Cooperation is achieved by various kinds of communicating environments. Three important basic structures which are benefit for agent' communication are Blackboard Structure, Message Sending Structure and Agent communication language. Blackboard paradigm is suitable for "Multi-Agents message sharing "or " Multi-agent discussion ". According to analysis of the Blackboard Structure and characteristics of AVDDTES, the system can use communication and cooperation mechanisms that are based on the blackboard structure.

4.1 Introduction of Blackboard paradigm

Blackboard paradigm can supply multi-Agents to achieve task paralleling and distributed computing [2]. This paradigm can integrate knowledge sources in heterogeneous world. Blackboard paradigm can supply Agent Communication Paradigm from some aspects as followings: Agent parallel controlling, Agent knowledge interaction, Agent updating controlling, etc. Blackboard paradigm is mainly made up of three parts: Blackboard, knowledge source and control policy. It can be described as chart 4. KS1。。。KSn indicate knowledge sources

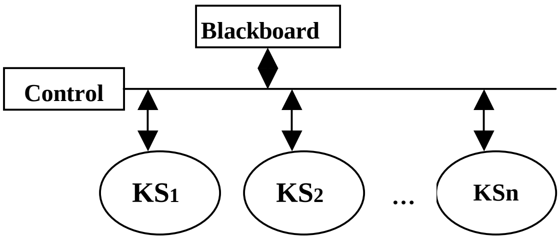


Fig.4 The blackboard structure

Blackboard is one kind of DB to store states of questions, send messages and handle methods. It record Messages needed by knowledge source and hypothesis from it, and can be shared by all knowledge sources. Then, knowledge sources can change blackboard to get the resolution of the problem. Knowledge Source is a knowledge database. They can get messages from blackboard, and express views (hypothesis, conclusion, etc), that is to say communication and cooperation should be achieved through blackboard. Controlling policy (control unit) can send messages between blackboard and knowledge source, then Coordinate program can use these messages to choose proper knowledge sources to achieve execution of Diagnosis orderly.

4.2 Communication and Cooperation mechanism in AVDDTES

Based on characteristics of blackboard, each Agent in AVDDTES has it's own blackboard structure, but Agents located in different layers have different functions in these three parts. According to the structure and functions of each agent, an agent can be divided into two classes, Leaf Agent and Non-leaf Agent (decisive Agent). For example, Diagnosing Agent is a Non-leaf Agent class, but Physical Diagnosing Agent, Biochemical Diagnosing l Agent and Image Diagnosing Agent are Leaf Agent classes. Form 3 shows the difference between these two classes:

Table 1 The difference between these two classes

Classification	Blackboard	Knowledge source	Control strategy
Leaf Agent	Store state's message, midst conclusion and its own result in logic .	Only be suitable for own knowledge which is domain knowledge	Be suitable for own control strategy in logic
Non-Leaf Agent	Store state's message, feedback message of sub-layer Agent and own inferring results .	Has domain knowledge, and control knowledge of task-decomposition, supervision, coordination	Mainly to Coordinate and arbitrate sub-layer Agents

In Multi-agent diagnosing system, we rule four response strategies:

(1) When Leaf Agent can't complete allocated tasks, it will send a cooperation request to Agent in up layer through blackboard; up-layer-Agent receive the request, estimate it, then make a response. If up-layer-Agent thinks this request is worth serving, then it will broadcast the request on blackboard to those possible agents in the same layer. Each Agent gets the requests, then analyzes the functions of this task, estimates it's own task and current states, then decides whether to accept the request or not.

(2) Leaf Agent in the same layer can't communicate and interact directly or send cooperation request with each other directly.

(3) Decisive Agent set up a system to divide task and estimate state . At first, it divide task into small pieces and allocate them to every functional Agent to execute , then estimate the executing result and coordinate tasks more efficiently. When Decisive Agent find a Agent unit can't finish task in time or can't fits some special needs, it will coordinate the whole task again, and allocate this small task or part of it to another agent. Leaf Agent must communicate with each other through up-layer-Agent.

Decisive Agent can communicate and exchange messages with each other directly, and handle cooperative task request with two agents or more. . It can greatly improve parallelism and flexibility of the system.

4.3 One example of communication and cooperation

We will explain this process using an example about the communication and cooperation between Diagnosis Agent and sub-layer-Agent, Diagnosis Agent and Treatment Agent.

(1) Diagnosis Agent is awaked by messages sent by Symptom sensing Agent in communicating area of the blackboard. Firstly, It disassemble cooperative task into small ones using private knowledge, and distribute these sub-tasks to suitable sub-layer-agents. There are two strategies of schedule: one is start sub-layer-agents one by one, another way is start two or more sub-layer-agents synchronously. It is an important problem to achieve cooperative diagnosis, and the key tech is what algorithms of task- coordinating should be chosen.

(2) There are many algorithms of coordinating strategy, such as multi-factors (symptom, willing of patients, facilities of hospitals) value-adding, dominant-decision, etc. Precedence of algorithms can be ascertained by dynamic knowledge controlling or ruled by users previously.

(3) Leaf Agent infers relative diagnosis autonomously on its allocated task, and send results back to Diagnosis Agent. Diagnosis Agent estimate and coordinate messages (when start up-layer-agents synchronously, messages back from up-layer has precedence) from assigned message area on the Blackboard. When the credibility of diagnosis conclusion reaches a particular degree, it can notice the conclusion to Treatment Agent, and finish its task.

(4) When the credibility of diagnosis conclusion is lower than the given threshold, Diagnosis Agent should ask Identify Agent for help. The communication and cooperation mechanism are similar with Diagnosis Agent and up-layer-agent .If Identify Diagnosis Agent still can't get a available result, the patient should be diagnosed by human specialist

5. DEVELOPMENT OF THE PROTOTYPE SYSTEM

From the notion of agent, we learn that Agent owns the characteristics of class in object-oriented methods (OOM) such as encapsulation, message-transfer, etc. we can use OO technology to achieve the functions of agent. Object-Relation Database offers good supports to achieve complicated Data patterns with layered structure. AVDDTES based on Object-relation Babe, sets up knowledge library and Database, use OO language Delphi as developing tool, and practice an prototype system.

6. REFERENCES

[1] Lu Ruling, Agent Communication Pattern Devise and Practice, Knowledge and Calculation Science, published by QH University, 2003.1.

[2] Wang Bin, Zhang Yaoxue, Chen Songqiao, A Communication Method of MAS Based on Blackboard Architecture, MINI- MICRO SYSTEM Vol.123, No.111, Nov. 2002.

[3] Sun Ning, Cao Yuanda, MAS-based Expert System Tool Computer Engineering, Vol.28, No.3, March. 2002

[4] Finin T, F ritzson R,M cKay D, M cEntire R. KQML as an agent communication language〔C〕. C IKM '94, ACM P ress, November1994.

[5] Hyacinth S., Nwana & Divine T. N dumu. A perspective on software agents research. See http: //agents. umbc. eduˆintroductionˆhn2dn2ker99.pdf

Yang kaiying is a Full Professor of Applied Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. She graduated from Xi'an Jiaotong University in 1970, and from Shanghai Jiaotong University with specialty of Computer Science and Technology in 1982. She has published three books, over 20 Journal papers. Her research interests are in application of distributed database and artificial intelligence technology in data processing.

Role-oriented Multi-Agents Approach to Optimize for Grid Resource Allocation*

Qian Wang **Debao Xiao**
Computer science department, Central China Normal University
Wuhan, Hubei, China
Email: icedlitchi@163.com **Tel.:** +86(0) 27 67866108

ABSTRACT

Managing resources in large-scale distributed systems – computational grids, is a very complex process. The computational grids resources are heterogeneous and their properties can vary over time. An approach adapt to computational grids environment is presented there. It is based on the role-oriented agents, where each role agent is modeled as a BDI agent. Those agents are autonomous (intelligent) processes, capable of communication with other agents, interaction with the world, and adaptation to changes in their environment. This article also describes the process of Resource allocation, which is one of a key technology of resource management in computational grids. And several optimizing strategies of resource allocation are discussed.

Keywords: Computational grids; resource allocation; BDI (Belief-Desire-Intention); agent

1. INTRODUCTION

With the proliferation of the Internet, the possibility of aggregating vast collections of computers into large-scale computational platforms comes. A new computing paradigm known as the Computational grid has been developed in recent years. It is designed to entirely share the Internet resources, such as computing resource, information resource, memory resource, knowledge resource and communication resource etc. So Internet is connected to be an enormous computer with the integration and cooperation of the resources. However, the resources management in computational grid environment is very complex, because of the distribution, heterogeneity, and dynamic variety of the resource[6]. An approach adapting to the resource management in computational grids environment is presented in this article, which is based on the role-oriented multiple BDI agents.

Resource allocation is a challenge problem in the management because of the following[3]. First, a key implication of the distributed nature of resource allocation is that control is distributed in multiple agents; yet these multiple agents must collaborate to accomplish the tasks at hand. Second, another implication is that agents face global ambiguity- an agent may know based on the results of its local operations, that some task is present. However, it may not be able to locally determine exactly which task is present. The agents must collaborate to determine which one of the many possible tasks is actually present and needs to be done. Third, different tasks may require the same resources and thus, resource contention may occur. In these situations, agents must take care to allocate critical resources to appropriate tasks –allocating a critical resource incorrectly may lead to situations where some other tasks must go unperformed. Finally, the situation is dynamic so a solution to the resource allocation problem at one time may become obsolete when the underlying tasks have changed. The agents must have a way to express and cope with such changes in the problem. Several elementary scenarios of optimizing resource allocation are given by the later part of this article.

2. GRID RESOURCE MANAGEMENT FRAME

Some tasks or subtasks in computational grid environment usually execute on distributed resources. For resource requirements and resource statements are time sensitive and vary at all the time, resource allocation which to find a mapping between tasks and resources should be adjusted dynamically.

On the basis of the hierarchy model of grid resource management, a scenario providing the high-level uniform resource management frame is brought forward based on the role-oriented multi agents. Its frame contains three levels that are the resource level, the agent level and the user level from the bottom up[9], which is showed by figure 1.

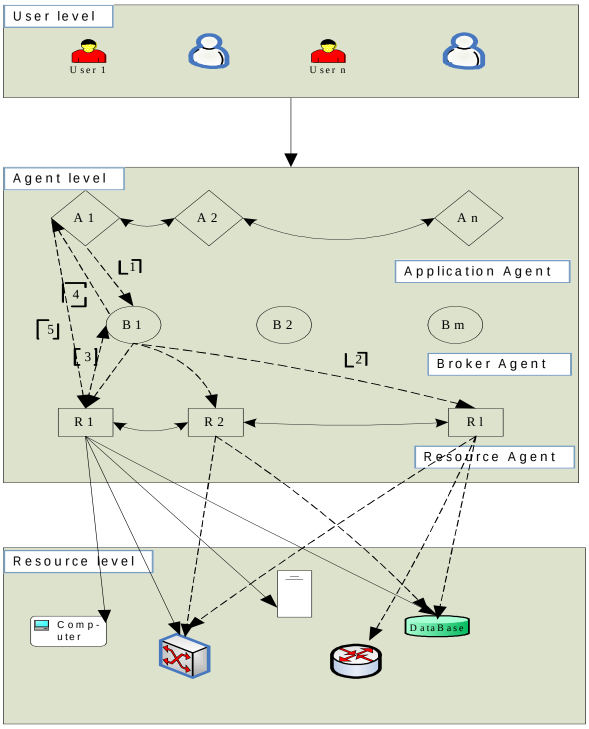


Figure 1[1]: Grid resource management frame

[*] This work was supported by the Provincial Great Science Project of HuBei (grant 2001AA104A05) and the Provincial Science Foundation of HuBei (grant 2001ABB013).

The resource level includes autonomic systems (AS) composed of numerous resources. It provides resource services to the agent level that is a middleware of computational grid.

The agent level is the kernel in the frame. Three roles agents on this level are divided by partaking different responsibility and possessing different privileges during the resource management process[10]. One is the resource agent managing the AS on resource level. One is the application agent managing particular computational grid applications in specific field or particular user requirements. Another is the broker agent also naming information service agent for providing uniform resource information services within a specific grid management domain. All of the three agents are BDI agents. They depend on this intelligent agent model to communicate and cooperate each other, attaining to manage the complex gird resources.

3. BDI AGENT MODEL

The BDI model is a popular model for intelligent agents. It has its basis in philosophy and offers a logical theory which defines the mental attitudes of Belief, Desire, and Intention using a modal logic. The central concepts in the BDI model are[1]:

- **Beliefs**: A set of information about the environment, other agents and the agent itself; *informative*.
- **Desires**: Situations the agents want to reach or maintain; *motivational*.
- **Intentions**: The currently chosen course of actions, which are the promises made by the agent to achieve some desires, and are the logical abstract to the plans; *deliberative*.
- **Plans**: Means of achieving certain future world states. Intuitively, plans are an abstract specification of both the means for achieving certain desires and the options available to the agent. Each plan has (1) a body describing the primitive actions or sub-goals that have to be achieved for plan execution to be successful;(2) an invocation condition which specifies the triggering event, and (3) a context condition which specifies the situation in which the plan is applicable.

The BDI agent consists of a belief set, and a collection of plan clauses. Each plan clause is of the form: $B_1 \wedge \ldots \wedge B_n \leftarrow S_1; \ldots; Sm$ where each B_i is a belief, and each S_i is either an action or a sub goal. Ultimately, all the plan clauses may equal a set of primitive actions, denoting as $\pi_i (X, \alpha)=(a_1(p), a_2(p), \ldots ,a_k(p))$ where each π_i is used to denote the *i*th plan clause , X to denote $B_1 \wedge \ldots \wedge B_n$, α to denote the goal which triggers π_i , and each action $a_i(p)$ corresponds to a well defined operation that an application agent can perform on a task graph ,or an resource agent on perform on its local schedule (or executing task) on this agent level. Each action within a plan results in an update on the properties of a task or a resource.

Before utilizing the BDI agent to managing grid resources, the grid resources should be described as a local goal of the different role agent using RSL. Each agent only maintain local plan library, this decentralized resource management policy adapts to the distributed and dynamic grid environment.

4. THREE ROLES AGENTS

Application agent (AA)

An application agent is responsible for managing one or more special applications. Each applications denoted as a task graph which is an ordered set, consisting of tasks *t* and arcs *l*: $TG=(t, l)$ where each arc *l* carries a label, a type, and input or output data.

Each atomic operation of application agent results in transfer of the executive unit state.

For multi tasks can be parallel executed, the application agent must decide an ordering-especially if tasks belong to different graphs can be shared. In cases where common task can be detected, such tasks are given a higher precedence (execution priority). Each application agent must therefore perform a static analysis of a task graph to determine whether tasks can be shared across graphs.

Each application agent maintains a plan library which can be used to change the ordering of tasks, the discovery of common tasks, the decomposition of tasks, and the aggregating of tasks. In this way, it optimizes the resource allocation, minimizing the execution time of an application for application agent.

Resource agent (RA)

A resource agent is responsible for managing the autonomy system submitted from the resource level. It also processes the given resource description and monitors the state of the resource, making these parameters available to broker and application agents.

Similar to an application agent, each resource agent maintains a plan library based on the type of resources it managed, which include scheduling operations on the resource, re-ordering a given schedule, and pre-empting executing tasks to adjust the resource allocation. Based on its current state (beliefs), the resource agent makes one or more plans active, and executes these in order to achieve its goal of improving resource utilization.

Broker agent (BA)

The Broker agent also naming the information service agent offers services such as a certificate granting service, a matchmaking service, and resource discovery service etc. Hence, the broker agent restricts interaction between the application agent and the resource agent based on access criteria, such as control access to resource agents from one or more administrative domains, access to tasks from resource agent based on the types of resources currently available, and based on periods of access for particular types of resources. Resource discovery is a elementary precondition of the resource allocation, which is charged by the broker agent that collects the information on the capability and utilization of resources that is advertised by resource agent in particular administrative domains. Moreover, the broker agent provides the usable resource information to the application agent based on some policies.

Co-operations of the multi agents

Co-operations among the multi agents by the interactions are very necessary during the managing resources process in the complex grid environment[11]. Sometimes in order to solve the confliction of the resource allocation, the agents undertaking a particular role even need to be altruistic.

The interactions among the three BDI agents are to change the mental situation (beliefs, desires, intentions) by the behaviors such as notification, requirement and response etc. During once interaction, at first, the receiver shapes a new belief state based on the message received, then forms new desires and new intentions, which triggers some behaviors to complete the next interactions. After multi interactions, agents may form team-promise and union intentions, which prepares for the cooperation of the multi agents.

5. OPTIMIZE GRID RESOURCE ALLOCATIONS

In this management frame, the agents undertaking different role participate different competitions to optimize the resources allocation. For examples, the application agents compete for the resources to minimize the execution time for a task graph that they manage. Similarly, the resource agents compete for the tasks aiming to maximize their utilizations over all available applications[1]. Resource agents compete for tasks from application agents based on the particular capabilities of resource agents.

Grid resource allocation is to find the mapping between tasks and resources. At first the broker agents make the resources discovered, collecting information list of the usable resources that is helpful to optimize resource allocation, then a given task for the application uses this list to make the resource mapping, which is showed by the figure 1(step 1 to step 4 is resource discovery; step 5 is application mapping).

The whole process and strategies adopted of resource allocation to a given task are as follows:

1) The application agent checks whether the given task is similar or identical (have the same input or output) to others. If similar or identical tasks exist, the application agent constructs a merged task and allocates it. Otherwise, the application agent sends a requirement message to the broker agent to get the information list of potential resources to which the task can be allocate it.

2) The broker agent gathers the information list of resources which match the application criteria, sometimes even include those that are not currently available. If the particular or only suitable for running this type of task is not available at that time the resource request is made. The broker agent must therefore determine whether to return resources currently available, or the best matching resources.

3) Once the application agent has received the resource information list from the broker agent. It selects a resource from the list at random, and sends it a request. If the resource is free and accepts the allocation, the task is transferred to the resource, and execution of the task commences. Otherwise, none of the resources on the list are able to accept the allocation at present time, i.e. all are busy. In this case there are several strategies that the application agent can pursue to continue resource allocation[1]:

I. The application agent may submit a new request to the broker with weaker requirement, i.e. it may indicate that it is willing to accept compute servers with less memory.

Ⅱ. The application agent may ask the broker to send it a list of all resources, regardless of availability. Based on this list, the application agent determines which other application agent owns the task that is keeping the resource busy, and requests the application agent in question to release the resource. This interaction involves a negotiation between the application agent (AA_1) requests the resource (RA_i), and one (AA_2) that currently owns the running task on the resource. The scenarios of AA_1may adopt are:

① AA_1 possessing of higher PRI for its task preempt the RA_i. Based on the belief set for RA_i, and its current plans, not only the existing schedule on RA_i to execute the task of AA_2 is aborted, or the all requests from AA_2 are ignored.

② AA_1 request AA_2 to release RA_i actively and reserve RA_i for it. If AA_2 agrees (this assumes that agents are altruistic, or at least cooperative), it will make a preemption request, and pass a reservation token to RA_i to enable AA_1, then schedule its task on RA_i. The release/reserve protocol for resource RA_i is shown by figure 2.

③ AA_1 request a higher priority level from a broker agent, and use this as a means to preempt RA_i. This strategy is different with the first one, which would not require a direct interaction between the application agents in order to resolve direct conflicts.

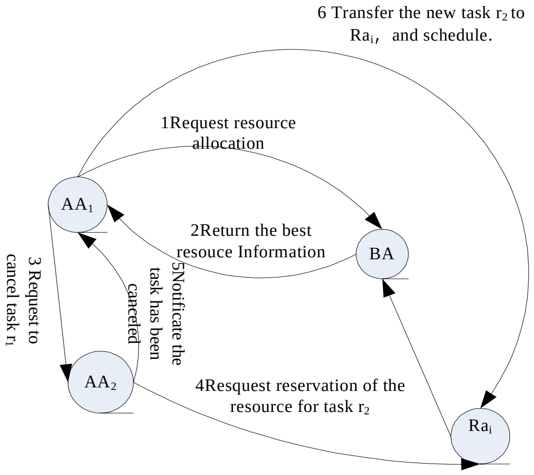

6 Transfer the new task r_2 to Ra_i, and schedule.
1Request resource allocation
AA_1
2Return the best resouce Information
BA
3 Request to cancel task r_1
5Notificate the task has been canceled
AA_2
4Resquest reservation of the resource for task r_2
Ra_i

Figure 2: The release/reserve protocol for resource (AA_1 wishes to execute a task r_2 on a given resource RA_i which is currently running task r_1 for AA_2)

6. CONCLUSIONS

An approach based on BDI model and role-oriented agents frame to manage the resources in the complex and distributed grid environment is more intelligent and more cooperative. The use of BDI behaviors enables new application or resource agents to be added, leading to existing agents adapting their behaviors. The BDI model is most appropriate to the grid resources allocation, because it enables each agent to model a site-specific administrative policy (for both RA and AA agents), and Broker agents may undertake different plans based on their priorities. The BDI model is also useful in that it can allow agents to enter/leave the environment dynamically, and for all other participants to adjust their plan libraries accordingly[1].

Generally speaking, multi-roles BDI agents cooperate with each other by interactions; they facilitate to optimize the grid resource allocation and solve the conflicts in the resource allocation etc.

7. REFERENCES

[1] [Jennings, 2001] N. R.Jennings. An Agent-based Approach for Building Complex Software Systems. Communications of the ACM, 44(4): 35-41,2001

[2] [Frey et al., 2001] J.Frey, T.Tannenbaum, M.Livny, I. Foster, S. Tuecke. Gondor-G: A Computation Management Agent for Multi-Institutional Grids, HPDC-10, IEEE Press, August 2001

[3] Pragnesh Jay Modi, ”Distributed Resource Allocation: Formalization, Complexity Results and Mappings to Distributed CSPs”, http://www-2.cs.cmu.edu/~pmodi/papers/modi-cp01-extended.pdf

[4] Omer F.Rana, Michael Winikoff, Lin Padgham, James Harland, Applying Conflict Management Strategies in BDI Agents for Resource Management in Computational Grids, http://goanna.cs.rmit.edu.au/~winikoff/Papers/acsc02.pdf

[5] B.J. Overeinder, “Multi-Agent Support for Internet-Scale Grid Management”

[6] Zuomin Luo etc, “A Survey of Grid Computing and Key Technologies”, Computer Engineering And Applications, 2003.30. (in Chinese)

[7] Zhi-hui Du, Yu Chen, Peng Liu, Computational Grid, Bei Jing: tsinghua University Books Inc. Pub.2001-11.

[8] Kang Xiao-Qiang,”Multi-Agent Interaction Based on BDI”, Chinese Journal of Computers, Vol .22 No.11, Nov 1999. (in Chinese)

[9] Hongqiang Cao etc, “Approach to allocate resources for grid based on a market mechanism”, Journal of Computer Research and Development, 2002. (in Chinese)

[10] http://www.actionsoft.com.cn/news/2003,12,00002.php (in Chinese)

[11] Xin-Yu Liu etc,”A multi-agent dynamic cooperating model based on BDI framework and its application ”, Journal of Computer Research and Development, Vol.39, No.7, July 2002. (in Chinese)

Debao Xiao is a Full Professor, doctoral supervisor and a head of Computer Network & Communication Lab at department of Computer Science, Central China Normal University. He graduated from Huazhong University of Science and Technology in 1969 with specialty of wireless technology, and he was a visiting scholar of MCRLab Ottawa, Canada (1993-1994). His research interests are in network management, network security and grid computing.

Qian Wang a M.D. candidate at department of Computer Science, Central China Normal University. Her current research interests include network management and grid computing.

Study on Multi-Agent Based Environment for Long-Distance Collaborative Learning on Internet

Ruolin Ruan[1, 2]
1. School of Information Engineering, Xianning College
Xianning, Hubei,　437005，China
2. College of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: rlruan@163.com　**Tel:** 0715-8339923

ABSTRACT

In the last few years, web-based collaborative learning is a very important area in study of long-distance learning; the problem of individuation and intelligentized learning is still the key problem of the current long-distance collaborative learning system. The paper import Agent technology of artificial intelligence in study of process for long-distance collaborative learning, and the paper study on multi-Agent based environment of collaborative learning on Internet. In the process collaborative learning, collaborative learning is achieved either between the student and the tutor (expert) or inside a group of learners interacting with the tutor. The system comprises an artificial tutor that tries to partially replace the human one during the students' interactions with the system. The architecture of the system is a multi-Agent one, human and artificial Agents collaborating together to achieve the learning task.

Key words: Long-Distance Learning, Virtual Reality, CSCL, Web CL, MAS

1 INTRODUCTION

With the development of computer network technology and multi-media technology, the development of long-distance education stepped a new phase (the third generation long-distance education). The third generation long distance education is the modern long distance education that is based on the satellite, television network and the computer network. And its advantage is that the teaching and learning environment is more open, agility, diversification and individuation. Application and popularization of Internet provided the new opportunity for the third generation long distance education. The long-distance collaboration learning is an important research realm in modern long-distance education. Internet owns many advantages, such as infinite time and space environment; the diverse multi-media information presents the form, super-medium information buildup method, friendly interface and interaction, and so on. They provided the valid environment for development of collaborative learning on Internet. The research of based on Internet collaborative learning has made the very big progress. There are some software of the collaborative learning has been manufactured the success, and get the application, but they still exist some problems in individuation and intelligence of collaborative learning process. The paper will import the technique of Agent to construct a long-distance collaborative learning system in order to study problems in individuation and intelligence of collaborative learning process.

2 CSCW (Computer Support Cooperative Work)

2.1 CL (Collaborative Learning)

The concept of collaborative learning (CL) refers to an instruction method in which students at various performance levels work together in small groups toward a common goal. The students are responsible for one another's learning as well as their own. Collaborative Learning is based on the study theory of constructivism and humanism. Its basic elements include the collaborative learning groups, the member, the tutor and the environment of collaborative learning. [1, 2] Collaborative Learning encourages active student participation in the learning process. It encompasses a set of approaches to education, sometimes also called cooperative learning or small group learning. The competition, individual learning and collaborative learning are three organized form of student study. The study is a process to obtains the knowledge, the constructivism study theories thinks, the knowledge is not from the teacher, but, under a certain context (social culture background) and the help of the other (include the tutor and the study colleague), the knowledge is from the sense-making by making use of the necessary study data. Because learning is a process of the sense making, it is particularly important for the learners to establish the context that benefit to the sense making. In above three kinds of studies organized form, the competition learning regards the classmate as the enemy, the surroundings classmate is ignored in individual learning, and it cannot exert the context's effect. |Collaborative learning take classmate as the study resources, and established the good study context for the learner, it is in the person of constructivism. [3, 4, 5]

2.2 CSCL (Computer Supported Collaborative Learning)

CSCW (Computer Support Cooperative Work) is the scientific discipline that motivates and validates groupware design. Put another way, it is a science that involves describing how to develop groupware applications. *CSCW* is also concerned with the study and theory of how people work together, and how groupware affects group behavior. [6] The goal of *CSCW* is to study how people work together both in small groups and large organizations, how *Groupware* applications influence those groups and organizations and how *Groupware* applications can improve or enhance communication between group members and coordination inside organizations.

CSCL is a research area of *CSCW*; it is the collaborative learning, which makes use of the computer technology (the multi-media and the network technique). By dint of the information network various types, on the one hand, it can realize the long-distance and interaction school teaching, interaction discussion and tutorship, and people can use the excellent teaching resources with the not summit to and horary restrict and the region; on the other hand, a group team collaborating together to achieve certain teaching or learning task in the different sites. [7, 8]

The traditional classroom environment is established under the collective tuition mode, the student just accepts the study with the collective form, and that is to say the traditional classroom environment is not for the service of collaborative learning. But the *CSCL* takes place in the environment of collaborative theories and computer technology, it can bring into play the advantage of the collaborative learning, and it can realize the form of collaborative learning which is not developed in the traditional classroom environment. Compare with the traditional *CL*, the *CSCL* broke through every variety box off barrier in the traditional school education, and realized the consecution of the time and the space, at the same time, interaction to become to control more and easily, the role of learner can also proceed to conceal, the role of teacher took place the basic change, what they want to control is not only the logic list of the content of course and the goal of reasonable arrangement, but is more the circumstance of collaborative learning and the programming design of the process of study.

3 WEB CL (Web-Based Collaborative Learning)

Students learning via *CSCL* technology need guidance and support on-line, just as students learning in the classroom needs support from their instructor. Educational researchers and technologists developing *CSCL* tools agree that group members do not necessarily have the social interaction skills they need to collaborate effectively. They recognize that students need practice, support, and guidance in learning these skills. Web-based Collaborative Learning is called *WBCL* on the abroad, but in here, we called it as the *Web CL*, and it is the collaborative learning by using the network technology. Say from this meaning, the *Web CL* is a subclass of the *CSCL*. A lot of excellent system web-based collaborative learning has been developed, and they have the vast applied foreground, such as, the *Web CL*^{TM} (Web-Based Cooperative Learning) is a network teaching system for completely supporting collaborative learning, which is developed by the laboratory of e-learning of Peking Normal University, and it offers study style measure, cent set, study, exchanges cooperation, the evaluation of study effect and collaborative performance etc. It is based on the constructivism study theory and system theory, it can well develop the learner’s subjective and motile characteristic and establish the good study environment for the learner, such it is easy for learner to construct their knowledge system.

4 MULTI-AGENT BASED THE PROCESS OF LONG-DISTANCE COLLABORATIVE LEARNING

4.1 MAS (Multi-Agent System) [9]

The research of web-based Collaborative Learning is already more mature, and a lot of concrete products have been got the extensive application in domestic and international, but the most of them do not have the individuation and intelligentized characteristic. With the development of the artificial intelligence technology, the technology of Agent is the investigative hotspot in *DAI* (Distributed Artificial Intelligence) area in the last few years, and it already got the extensive application in computer science many areas.

In the past few years, multi-Agent systems (*MAS)* have become a very active research area that has connections to many other areas, both inside and outside of computer science. Agent is an inner driving software entity which has autonomy, self-adaptability, collaborative and intelligence characteristic, it can do to use for the oneself and the environment and it can do adaptability reaction to the environment. The single an Agent primarily used for realizing the native mission, also used for to go forward an information search in the net. The *MAS* is composed of a group Agents that are independent and collaborative work, Agent is its basic constitute unit and the entity of the independent movement. In the *MAS*, each Agent mutually negotiates and cooperates in order to achieve certain common task. In the *MAS*, each Agent can according to the variety of load and the circumstance of the others Agent to harmonize the own behavior, and it can make the reasonable arrangement and redressal for the realization of the goal and the usage of the resources in order to avoid the clash.

In the research of Agent, people import many concepts of psychics and human behaviorism, so the Agent has the very strong intelligence and the humanized color, and it is very easy to maintain and expand. Therefore, we apply the Agent to e-learning, and we will change the interactive entity into the Agent that has individuation and intelligence characteristic order to improve the intelligence of the system, such we will create a good study context, and attain to the result of drawing on student's interest, developing the individuation education and improving the effect of the teaching. Then, we will discusses multi-Agent based the virtual environment for long-distance collaborative learning.

4.2 Study on Multi-Agent System Based Process of Collaborative Learning

The architecture of the collaborative learning system is a multi-Agent one, human and artificial Agents collaborating together to achieve the learning task. There are several Agents in the system, as depicted in Figure 1. [10, 11] A learner in the collaborative learning environment is endowed with his own digital personal Agent. A dedicated window can be activated when the user wants to interact with his Personal Agent (*PA*), but the *PA* may react and make the learner aware of its presence when relevant information has to be communicated. The *PA* is responsible for monitoring the user’s actions, for creating the learner’s history and preference profile, and for entering in dialogue with the *PA*s of other learners. The tutor in the collaborative learning system has also his personal Agent, built on similar principles as the other *PAs*, but without the function of creating the learner’s history. This uniformity in treating the personal Agents facilitates communication and coordination in the system.

The learning system has Information Agent (*IA*) which is responsible of retrieving and filtering information from specified sources that can range from the on-line course materials to the entire Web. The filtering criteria may be specified by the learner when calling the *IA*, may be retrieved from the preference profile of the user (provided by the user’s *PA*) or may be created as a combination of the two.

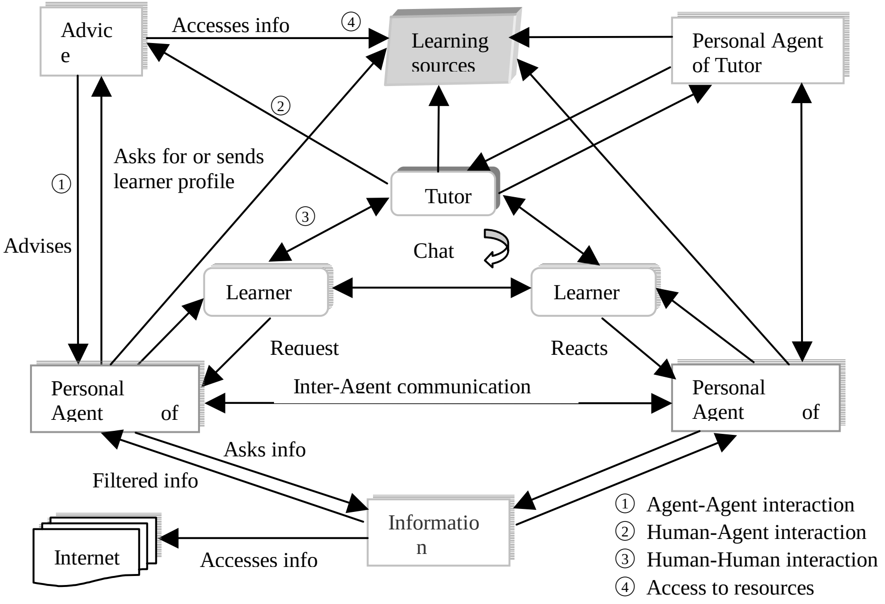


Fig. 1 Multi-Agent based process of collaborative learning in the system

Another Agent in the system is the artificial tutor or the Advice Agent (*AA*). The *AA* is capable of assisting the students when the first and second method of teaching is applied. When the student is browsing the in-line course material, depending on the answers to the questions and exercises inserting in the course but also considering the learner's history, the Advice Agent may suggest what parts or relevant topics of the course the student has to revise. The *AA* can also assist students in the system during the application of the second method of teaching namely problem demonstration. It can participate to the initial settings of the problem parameters by inspecting its internal problem representation and it can contribute partially to the interpretation of results. This last action depends on its already acquired knowledge. When making suggestions about how to set initial parameters, the Advice Agent will advice the student more or less, depending on the learner's profile developed until the time being. The Advice Agent is capable of learning from the human tutor how to contribute to the problem parameter settings by applying a simple learning from examples algorithm. Several more sophisticated machine learning algorithms and deep representation of the problem solution will be developed in the future to increase the competence and efficiency of the artificial tutor.

The interface between the learner and the *AA* is realized also by the *PA* of the learner. The *PA* is responsible for calling the *AA* to be active part of the learning process, and thus to assist the student. In the same time, the *AA* will query the *PA* of a learner on the student profile, so as to tailor its advises according to the abilities of the learner. For the time being, the learner profile refers mainly to the learner's history and preferences. The learner's history comprises, as most important parts, details on the student-system interaction and the level of achievement on self-assessments and assessments. The student-system interaction details refers to the learning methods selected during past interactions, peers learners in the group, how often the learner required the *AA* to be part of the learning process, etc. We are considering extending the learner profile with several other features and we are currently working on a model of a group profile in learning. This group profile will be used to correlate the activity of the Advice Agent with the level of knowledge of the group. The group profile will be also used by the human tutor who may, in this way, get a better understanding of the group with which he is working.

A further development of the system will be dedicated to conceive and build an intelligent planning Agent that will be able to assist students in problem solving by collaboratively participating to the synthesis of the problem solution.

4.3 Composing of System Interface and Supporting of key technology

The system interface will comprise several learning and communication areas to support both individual and collaborative interactions. There are two types of areas in the user-system interface: one type is represented by private areas that are to be viewed and managed only by the individual user connected to the system, while the other is the common type which corresponds to areas that are shared by all users connected in the system in synchronous mode or where an individual user can visualize previous collaborative experiences. For example, the Private Work Area may be used by the user for private work, private notes, etc.; the Private Learning Area may be used for browsing course notes or for assessment. The Common Area for Parameter Settings is used to collaboratively tune a solution for a specific problem instance, the Common Chat Area for informal discussions among members of the group, and the Common Area for Analysis to collectively analyse problem solutions. A Coordination Area is used for showing and managing group interaction in the system, namely: who is participating, who is inputting data of a form or another, who is in control of the system at a given moment.

The hierarchical architecture of the learning system should comprise the application layer (the facilities of collaborative learning and Agent technology), the middleware layer (Java and CORBA) and TCP/IP. The multi-Agent paradigm is used both as a model to support intelligent human-computer interaction during the learning task and as a supporting technology for building the system. The multi-Agent model that supports the system is an instantiation of the general multi-Agent framework.

5 CONCLUSIONS

The individuation and intelligence long-distance collaborative learning is an important problem for the modern long-distance education. To increase the efficiency of learning and to commit the student, the new educational approach of "learner-centered", "participative learning", or "problem-based" learning has began to show its merits, the individuation and intelligence study according needs will become mainly character of the education mode in this century. Through the research of the *CSCL*, *Web CL* and the technology of multi-Agent, the paper applied the theory and technology of multi-Agent to the environment of collaborative learning in order to resolve the individuation and intelligentized problem, and gave the relation among the learner, tutor and Agent, lastly, it put forward a multi-Agent based the virtual environment model for long-distance collaborative learning.

In the future, with the development of multi-Agent and mobile-Agent technology, we can apply mobile-Agent technology to the environment for long-distance learning.

6 REFERENCES

[1] Anuradha A. Gokhale. Collaborative Learning Enhances Critical Thinking. http://www.sdedu.net/

[2] http://www.wcer.wisc.edu/nise/cl1/CL/intro.asp

[3] Melissa Bergen, Jörg Denzinger, Jordan Kidney. Teaching Cooperation in Multi-Agent Systems with the help of the ARES System. http://www.cs.ubc.ca/wccce/Program03/papers/Denzinger/denzinger_et_al.html

[4] Amy Soller. Supporting Social Interaction in an Intelligent Collaborative Learning System. International Journal of Artificial Intelligence in Education (2001), 12, to appear

[5] Amy Soller, Kwang-Su Cho, Alan Lesgold. Adaptive Support for Collaborative Learning on the Internet. http://sra.itc.it/people/soller/documents/its2000/ws2-poster-5.htm

[6] CSCW & Groupware. http://www.cs.tcd.ie/

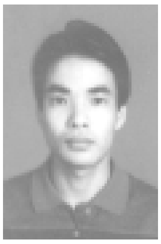
[7] Jianhua Zhao, Kedong Li. Collaborative Learning and Collaborative Learning Mode [J]. China Educational Technology. 2000, 10 (in China)

[8] Yongcheng Gan. Web-Based Collaborative Learning and study of Application of CSCL [J]. Educational Technology Research. 2002, 4 (in China)

[9] Zhongzhi Shi. Intelligence Agent and its application [M]. Science Publishing House.2000 (in China)

[10] Adina Magda Florea. A Virtual Environment for Collaborative Learning. http://rilw.emp.paed.uni-muenchen.de/

[11] Adina Magda Florea. An Agent-based Collaborative Learning System. http://www.dsp.pub.ro/



Ruolin Ruan is a graduate student and an engineer of school of information engineering, Xianning College, China. He graduated from Huanggang Normal College in 1996; from Huazhong University of Science and Technology in 2002 with specialty of computer application technology. His current research interests are in artificial intelligence, multi-Agent system, long-distance learning and computer multi-media technology.

A Heuristic Algorithm for Agent-based Task Scheduling in Grid Environments

Ding Shunli [1,2] **Yuan Jingbo** [1,2] **Ju Jiubin** [1]

[1] **College of Computer Science and Technology, Jilin University, Changchun, Jilin, CHINA**

[2] **Department of Computer Engineering, Northeast University at Qinhuangdao, Hebei CHINA**

Email: dingsl@163.com

ABSTRACT

Resource management and task scheduling is a crucial problem (the core of all the questions) in Grid Environments. The aim of task scheduling is to take full advantage of grid resource and execute user's task request as early and quickly as possible. This paper introduces an agent-based resource management model and agent structure and its function. On the basis of local resource's adopting the strategy "First Come First Served" to the task, we put forward a task scheduling heuristic algorithm using a technique of task advertisement and discovery. Agents are organized into a graph and the heuristic algorithm is based on multi-agent cooperation, to ensure this methodology achieves the goal of task scheduling.

Keywords: task scheduling, agent, graph, heuristic, request discovery

1. INTRODUCTION

The infrastructure of the grid is an open, complex software system. Multi-agent technology, as one of the ways to overcome the challenges in the development of the grid, can better ensure system Scalability and Adaptability. Service has been accepted as the most important concept in this distributed system development, and service discovery is therefore considered an essential part in many distributed system infrastructures.

There are several solutions that currently address issues of grid resource management and scheduling. These include Globus [11], Legion [15], NetSolve [12], Condor [8], Ninf [13] and Nimrod/G [10]. While many of these projects utilise query-based mechanisms for resource discovery and advertisement [7], this work adopts an agent-based approach. The agent-based system allows agents to control the query process and make resource discovery decisions based on their own internal logic rather than rely on a fixed function query engine.

An agent-based approach provides a clear high-level abstraction and a more flexible system implementation [14]. Multi-agent systems have recently been introduced in grid development and resource management. This work [2] includes a model for distributed awareness and a framework for the dynamic assembly of agents for the monitoring of network resources.

An agent-based resource management and task-scheduling model is described in [1]. The centre of the model is an algorithm that implements resource advertisement and discovery and can better solve resource scheduling in large-scale system. In this algorithm, task is first submitted to an agent, and then with the aid of discovery mechanism under agent hierarchy, an available resource is found. In this process, it is likely that some circumstances as follows will appear: when a task found an available resource on some agent, but the resource had been found by other tasks before this moment and resource state was not modified timely, the resource was available on the surface now but was actually unavailing (inefficacy), the process of discovery has to proceed with the search. In the extreme, this process will forever repeat but cannot find available resource. In view of the above mentioned, this article describes a task-based advertisement and discovery mechanism to better solve the problem.

The rest of this paper is organized as follows. In Section2, we briefly discuss the related work. We describe the agent-based system architecture in the grid in section 3. In section 4, the agent function is described. A task scheduling heuristic algorithm is describe in section 5. Algorithm analysis is included in section 6 and the paper concludes in section 7.

2. RELATED WORK

In this section, we survey representative research on resource management for Grid computing systems. An agent-based grid service discovery has been presented in [9]. This paper mainly describes grid service discovery architecture and basic design, how service request agents locate specific grid service agent by submitting requests to the Grid Service Manager with descriptions of required services in the network, and how Grid Service Agents dynamically register their services in Grid Services Manager. Agent-based grid management is also used in [3], where centralized broker/agents architecture is developed for management of massive data storage systems. Grid management in this work uses a hierarchy of homogenous agents that can be reconfigured with different roles at run time. An agent-based methodology has been developed for the management of large-scale distributed systems with highly dynamic behavior [5]. The system consists of a hierarchy of homogenous agents where each agent can be considered both a service provider and a service requestor. Multiple homogeneous agents are organized into federated groups, which have capabilities of service advertisement and discovery.

A resource discovery based on the peer-to-peer model has been proposed [4], which consists of a few request-forwarding algorithms in a fully decentralized architecture accommodating heterogeneity and dynamism in resource. Work on resource discovery in large-scale distributed systems has initially focused on the functionality of the resource discovery protocol and on appropriate visualization [6], with emphasis on protocol specifications and resource representations.

Among existing models, the object of advertisement and discovery service is resource, but there is a deficiency in these services, namely, communication delay may result in outdated information. For example, the agent will trigger a discovery process to find a resource that can execute the task when application request is submitted. If the available resource information is found in upper layer, but the resource may be executing other task now and the operation that modify state information has not yet been finished, this discovery results in a feint, that is to say, discovery is successful but the request cannot be met. In the extreme circumstances, maybe a discovery process finds repeatedly such resources all the time. Ultimately, the request can never be met. We analyze carefully the goal of resource management and scheduling and find the goal ensures that each task will be fulfilled as soon as possible, using available resources. Therefore, we don't want to see the cases mentioned above in resource management. To solve the problem, we bring forward a service advertisement and discovery strategy using task request as the object. The strategy avoid above-mentioned problem and can attain the goal. Another problem lies in the papers above is this, the organization of agents is hierarchical structured, as the utilization of resources changes with time going by, that is, the sub-tree with the current agent as the root will invalidate if the agent invalidates. And thus result in the invalidation of resources, which is what we do not want to see. If the agents are organized in a graph, this problem is solved. Agent model organization based on graph is introduced in this paper, and it broadcasts tasks on this structure, heuristic searching method is adopted to find tasks, and thus this problem is solved.

3. SYSTEM ARCHITECTURE

Agent-based resource management architecture is illustrated in Figure 1. The main components include grid users, grid resources and agents.

There are different kinds of users of grid computing environment. These include grid service and tool developers, application developers and grid end users.

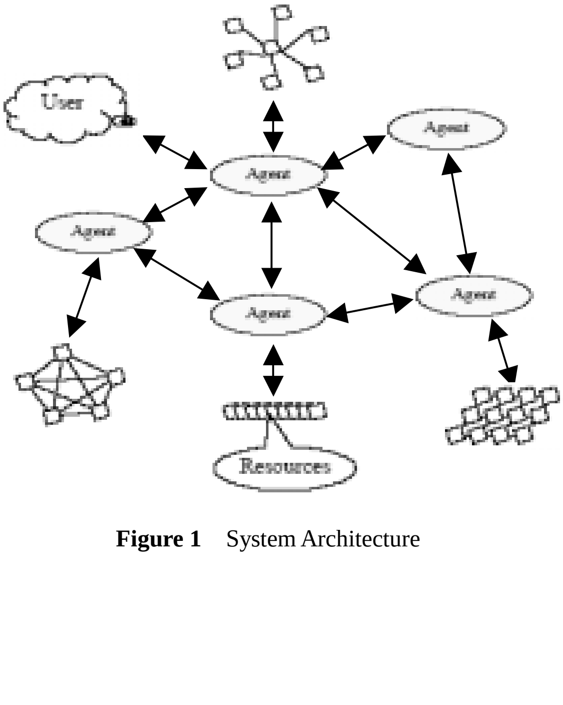


Figure 1 System Architecture

Grid resource can provide high performance computing capabilities for grid users. A resource can include Massive Parallel Processors (MPP), or a cluster of workstation or PCs. A grid resource can be considered a service provider of high performance computing capabilities. In this model, grid resources provide service to find tasks of waiting execution by the agent graph.

Agents are the main components in the system. Each agent is viewed as a representative of a grid resource at a meta-level of resource management. Therefore, an agent can be considered a service provider of high performance computing capabilities. Agents are organized into a graph. The graph of homogenous agents provides a meta-level view of the grid resources. The service information of each grid resource can be advertised in the agent graph; agents can also cooperate with each other to discover available ta.

4. AGENT

The agent system bridges the gap between grid application users and grid resources. The agent graph allows scalability to be addressed. Advertisement and discovery are processed stage by stage between neighboring agents. This feature plays an important part in system scalability. Another important factor is the capacity for agents to be able to adjust their advertisement and discovery behaviors, thus adapting to the highly dynamic grid environment.

In this paper, an agent which is considered to be both proxy of resource and proxy of task, assume most resource monitoring, task management, resource allocation, ART management and maintenance, task scheduling, task advertisement and discovery, and communication with other agents.

Each agent utilizes Agent Request Tables (ARTs) to record request information of it and other agents' .An ART item is composed of three constituent parts:

- Agent ID. This ID includes the contact information of an agent. An agent can only get ID information and contact its adjacent agents. An agent can also contact more agents and cooperate with them for service discovery.
- Task Information. Task information should contain all related information about a task. This information will be used by the agent to select the corresponding resources, and make service discovery decisions. In general, a name should be defined for each application.
- Options. Additional options can be added into each ART item to constrain agent behaviors for service advertisement and discovery.

An agent can choose to maintain different kinds of ARTs according to location in hierarchy and different task information. These include:

- T-ART (This ART). In the coordination layer of each agent, T-ART is used to record request information of local user. The local management layer is responsible for collecting this information and reporting it to the coordination layer.
- A-ART (adjacent ART). The A-ART in an agent is actually a record of the request information received from its adjacent agent.

When a new request is submitted, its agent should advertise the request information to other agents. The request information in ARTs offered by an agent can change over time. When this occurs, the corresponding request information needs also to be updated. When a request is met, it needs to advertise to cancel previous information that has been advertised into the graph. The dynamics of the system increase the difficulty of system management.

There are two methods of maintaining ART coherency - data-pull and data-push, each of which occur periodically or can be driven by system events.

- Data-pull - An agent asks other agents for their request information either periodically or when a request arrives.
- Data-push - An agent submits its request information to other agents in the system periodically or when the request information is changed.

An agent uses the ARTs as a knowledge base. This is used mainly to assist in the discovery process triggered by the submission of an idle resource. Application discovery involves querying the contents of the ARTs in the order as follows: T-ART, A-ART. If an agent exhausts the ARTs, and does not obtain the required service application request information, it can submit the resource information to its adjacent agent or terminate the discovery process and expect the beginning of the next discovery process.

5. HEURISTIC ALGORITHM

In this model, the advertisement and discovery process make use of multi-agent graph structure and information in each agent's ARTs.

Definition: Multi-agent graph are defined as follows:.

$$G = (A, E)$$

$$A = \{a_i \mid a_i \in agents\}$$

$$E = \{(a_i, a_j) \mid a_i, a_j \in V \wedge P(a_i, a_j)\}$$

In the formula defined above, G represents a graph, and A is the set of nodes in G, while agents is the set of available agents. E is the set of edges in graph G, P (a_i, a_j) denotes there is a direct link between a_i and a_j, *that is* , (vi, vj) is an edge.

In this model, an agent not only manages local resources but also receives application requests. Local management in an agent is responsible for receiving application requests and provides them to the coordination layer to store, and decide how to advert to adjacent agents. When a resource is idle, corresponding agent starts a request discovery process.

An agent can also receive many pieces of advertisement information from adjacent agents and also store the information in its coordination layer as its own knowledge. All of the request information is organized into ARTs (Agent Request Tables).

5.1 Request Discovery

Each agent has different kinds of ARTs maintained by request advertisement. An agent takes the contents in ARTs as its own knowledge, which is mainly used for request discovery. A request discovery process is triggered by the arrival of a request in an agent. A resource is usually composed of several parts:

- Resource information. These include details of idle resources. This information may be combined with information in ARTs to produce high-level performance information of corresponding requests.
- Resource performance. This includes details of resource performance information, which may be used for matchmaking for agents to make decisions on whether a resource can provide a capable service or not.
- Options. Additional options may be attached to each resource, which may include user control information for the service discovery. For example, the user may limit the time and scope of a discovery process.

An agent can act on a resource in a number of ways, for instance:

- Yes. I can provide executable task for a resource, so the discovery ends successfully.
- No. I cannot provide executable task. However, I know an agent, which may have the capability to provide such task. I can transfer the resource to it for further discovery.
- No. I have no task matching resource. However, I can transfer the resource to adjacent agents for further discovery.
- No. I have no task matching resource., and there are also no other agents that I can query, the discovery has failed.

5.2 Query ART

The process of request discovery in an agent is the process of looking in the ARTs. The general order for an agent to check different kinds of ARTs in turn is first T-ART last A-ART, which will be explained one by one below.

An agent not only manages large numbers of resources but also receives user requests in the large-scale environment. When an agent receives a local idle resource or another agent's resource, it is natural that it will check its own task requests recorded in the T-ART firstly. If a resource can find the executable task in it own territory, the discovery is successful and the information will be notified to relevant resources and tasks.

If there is no required service information in the T-ART either, an agent may then choose to look up its A-ART. A-ART records service information in local scope. Most users prefer to find an available resource located as near as possible. If the required service information is found in the A-ART, the request will be dispatched to the corresponding agent. Otherwise, additional service discovery will have to be processed.

If after looking up all the ARTs and no required service request information is found, the agent may consider submitting the request to its adjacent agent. The adjacent agent will follow the same procedure, but may maintain request information in a larger scope, thus it may be more possible to ensure available resource to provide service for grid users.

If an agent looks up all of the ARTs and does not get the required request information, and there is no other agent it can contact for further discovery, the discovery ends up failed.

From the above description, a discovery may end successfully or in a failed state. Additional options may be attached with a resource information, which may constrain the time or scope of discovery. Such kinds of options may stop and fail a discovery process.

Heuristic method is used when an agent utilizes learning methods to maintain information of adjacent nodes, when it finds an available task waiting to run. And this brings a new problem, that is, with time passing by, the number of adjacent agents increases. To solve this problem, we could set a maximum to the number of adjacent agents for each agent, when this maximum is reached, one of the adjacent agent is deleted, it could be deleted directly or using some kinds of strategy.

In next section, an instance and a formal approach will be introduced in order to make readers comprehend better the relation between advertisement process and discovery process in higher dynamical system.

5.3 Formal Approach

The rule-based reasoning is the basis in system dynamic processes. We can define corresponding rules to represent every task of discovery process. The symbols in rules are defined as follows:

Ai (i=1,……,n):one of the agents
r :a resource
t(r)and a(r) evaluation result of r in T_ART and A_ART respectively. t(r), al(r)∈{Ai (i=1,……,n), null} , null means no application request can be executed by r
Ai(r), Ai processes the request in ***r***

- ***Rule 1:*** $A_i\ (r) => A_i \rightarrow (t(r)，a(r)，*)_{Ai}$

  The request discovery process in an agent is the process of looking up the T_ART and A_ART.

- ***Rule2*** ：$(A_{this}，*，*)_{this} => TaskFound$

  If an agent is aware that it has had requests submitted by users, the request discovery is successful.

- ***Rule3*** ：$（null，adjacent，*）_{this} => adjacent\ (r)$

  If the request discovery cannot be found in the T_ART but in the A_ART, the resource will be dispatched to the adjacent agent.

- ***Rule4*** ：$（null，null，1）_{this} => A_{adjacent}$

  If an agent exhausts the ARTs, and does not obtain the request information, and the agent with parent, it will submit the resource to its adjacent agent.

- ***Rule5*** ：$（null，null，0）_{this} => NoTask$

If an agent exhausts the ARTs, and does not obtain the request information, and the agent without parent, the request discovery ends unsuccessful.

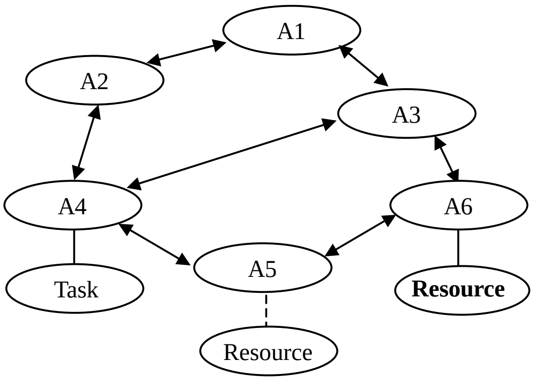


Figure 2 An Example System

The example shown in Figure 2 is a simple agent system with six nodes. Each agent maintains a T_ART and a A_ART. Consider a typical process: agent A6 provides a resource r, using this Agent graph, and finds a task submitted by Agent A4.The discovery process is described by formal representation .as follows:

$$
\begin{aligned}
A_6(r) &=> A_6 \rightarrow (null, null, 1)_{A6} && (1)\\
&=> A_6 \rightarrow A_3(r) && (3)\\
&=> A_6 \rightarrow A_3 \rightarrow (null, A_4, *)_{A3} && (1)\\
&=> A_6 \rightarrow A_3 \rightarrow A_4\ \ (r) && (3)\\
&=> A_6 \rightarrow A_3 \rightarrow A_4 \rightarrow\ \ (A_4, *, *)_{A4} && (1)\\
&=> A_6 \rightarrow A_3 \rightarrow A_4 \rightarrow TestFound && (2)
\end{aligned}
$$

The system can have more than six nodes and the requests may be changed many times. The system behavior for request discovery may be much more complex. Modeling and simulation tools can be developed to estimate the system performance.

6. ALGORITHM ANALYSYS

From the above mentioned, the key point is to match the proper request in ART with the information transmitting among agents. We adopt the principle “first come first served” to deal with the requests, in this way, one resource executes the first request which can satisfy the resource requirement. Every agent management and the request scale that it can possibly accept are not large, so the scope of the table will be respectively small, therefore the search in the ART table will be conducted according to its order. To the second problem, Depth-first –search method is adopted in this algorithm, and its time complexity is O(n+e). As heuristic information is used in selecting the adjacent agent, better efficiency is achieved.

7. CONCLUSION

The primary task of resources management and task scheduling is to make good use of available resources and execute user’s task as soon as possible. In this paper, we introduced the structure of agent and agent-based graph model, and on the basis of the latter, we introduced the concept of ART table, with the request as the objective of discovery in the discovering process. Thus utilized resource ratio can be increased and at the same time the task can be fulfilled as soon as possible, meanwhile the problem that a certain request may never be satisfied is solved as false discovery will occur if resource acts as discovering objective, heuristic information is introduced in the discovering process to achieve higher efficiency, furthermore, a brief analysis of the algorithm is carried out. Now we set out to choose a suitable setting to realize our model, and build a real system, and evaluate its practical performance.

8. REFERENCES

[1] J. Cao, D. J. Kerbyson, and G. R. Nudd, Performance evaluation of an agent-based resource management infrastructure for grid computing, in “Proc. 1st IEEE International Symposium on Cluster Computing and the Grid”, pp. 311-318, Brisbane, Australia, 2001.

[2] K. Jun, L. Boloni, K. Palacz, and D. C. Marinescu,“Agent-Based Resource Discovery”, in Proc. 9th IEEE Heterogeneous Computing Workshop, 2000.

[3] B. Tierney, A monitoring sensor management system for grid environments, Cluster Computing 4(1) (2001) 19-28.

[4] A. Iamnitchi, I. Foster, On Fully Decentralized Resource Discovery in Grid Environments, 2nd International Workshop on Grid Computing, Denver, Colorado, USA, pp.51-62, Nov. 2001.

[5] J. Cao, D. J. Kerbyson, and G. R. Nudd, “High Performance Service Discovery in Large-scale Multiagent and Mobile-agent Systems”, Int. J. Software Engineering and Knowledge Engineering, Special Issue on Multi-Agent Systems and Mobile Agents, World Scientific Publishing, Vol. 11, No. 5, pp. 621-641, 2001.

[6] Rana, O.F. Bunford-Jones, D. Walker, D.W. Addis, M. Surridge, M. Hawick, K. Resource discovery for dynamic clusters in computational grids, Proceedings 15th International Parallel and Distributed Processing Symposium, San Francisco, CA, USA, pp.759-767, Apr. 2001.

[7] K. Krauter, R. Buyya, and M. Maheswaran, “A Taxonomy and Survey of Grid Resource Management Systems”, to appear in Software: Practice and Experience, 2001.

[8] R. Raman, M. Livny, and M. Solomon, “Matchmaking:Distributed Resource Management for High Throughput Computing”, in Proc. 7th IEEE Int. Symp. on High Performance Distributed Computing, 1998.

[9] Li Changln, Li Layuan, An Agent-Based Approach For Grid Computing, Proceedings of The Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies pp. 608-611, china, Aug, 2003.

[10] R. Buyya, D. Abramson, and J. Giddy, “Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid”, in Proc. 4th Int.Conf. on High Performance Computing in Asia-Pacific Region, Beijing, China, 2000.

[11] K. Czajkowski, I. Foster, N. Karonis, C. Kesselman, S.Martin, W. Smith, and S. Tuecke, “A Resource Management Architecture for Metacomputing Systems”,in Proc. IPPS/SPDP '98 Workshop on Job Scheduling Strategies for Parallel Processing, 1998.

[12] H. Casanova, and J. Dongarra, “Applying NetSolve’s Network-Enabled Server”, IEEE Computational Science & Engineering, Vol. 5, No. 3, pp. 57-67, 1998.

[13] H. Nakada, H. Takagi, S. Matsuoka, U. Nagashima, M. Sato, and S. Sekiguchi, “Utilizing the Metaserver Architecture in the Ninf Global Computing System”, in Proc. High-Performance Computing and Networking,LNCS 1401, Springer-Verlag, pp. 607-616, 1998.

[14] N. R. Jennings, “An Agent-based Approach for Building Complex Software Systems”, Communications of the ACM, Vol. 44, No. 4, pp. 35-41, 2001.

[15] S. J. Chapin, D. Katramatos, J. Karpovich, and A.Grimshaw, “Resource Management in Legion”, Future Generation Computer Systems, Vol. 15, No. 5, pp. 583-594, 1999.

Ding Shunli is a Ph.D. student of Department of Computer Science at the Jilin University and an association professor of Department of Computer Engineering, Northeast University at Qinhuangdao. His current research interest is Distributed Computing System and Grid Computing.

Yuan Jingbo is a Ph.D. student of Department of Computer Science at the Jilin University and an association professor of Department of Computer Engineering, Northeast University at Qinhuangdao. His current research interest is Distributed Computing System and Grid Computing.

Parallel Hyperspectral Integrated Computational Imaging

Bin Dai and Robert A. Lodder*
Department of Chemistry, University of Kentucky
Lexington, KY 40506-0055 USA
***Author to whom correspondence should be addressed.**
Email: lodder@uky.edu **Tel.:** 1-859-257-9232

ABSTRACT

Modern hyperspectral imaging is able to collect extraordinary amounts of information at amazing speed. Data threaten to cause a bottleneck in genetic research, drug discovery and development, and other areas. Reducing these data from physical fields to high-level, useful information is difficult. Integrated computational imaging (ICI) is a process in which image information is encoded as it is sensed to produce information better suited for high-speed digital processors. Both spatial and spectral features of samples can be encoded in parallel in ICI. When spectral images are simultaneously obtained and encoded at many different wavelengths, the process is called hyperspectral integrated computational imaging (HICI). Lenslet arrays and masks are ideal for encoding spatial features of an image. Complex molecular absorption filters can be used as mathematical factors in spectral encoding to create a factor-analytic optical calibration in a high-throughput spectrometer. In this system, the molecules in the filter effectively compute the principal-axis calibration function by weighting the signals received at each wavelength over a broad wavelength range. One or two molecular filters are sufficient to produce a detector voltage that is proportional to an analyte concentration in the image field. Because a single detector voltage can reveal analyte concentration, HICI is able to calculate chemical images orders of magnitude more rapidly than conventional chemometric approaches. HICI can contribute to three areas of research in image informatics: (1) database design and metadata structures for multidimensional images, (2) automated object tracking and event recognition in videos, and (3) interactive analysis and query by content of videos.

Keywords: lenslet arrays, molecular computing, multiplex bandpass filter

In the past, optics has served mainly to render the world more easily visible to humans. Now, computers are increasingly employed to make sense of the visual world in ways that people cannot. With a new generation of optics, scientists and engineers are recasting visual scenes for interpretation exclusively by computers. To the human eye, these pictures appear distorted at best, or at worst look like visual noise, without discernable meaning. But to computers, such data are worth more than a thousand words. Optimizing complete vision-and-action systems for computers lies at the core of integrated computational imaging. Computers are well established manipulators of digitized images, and image-processing programs do it routinely on desktop machines. However, what is new is the strategy of modifying image information as it is sensed to make it better suited for the "computer mind" [1].

For example, rather than the customary concave and convex disks, optical engineers are fabricating strangely shaped, fundamentally different lenses adapted to the strong points of computers. These optics diverge from the traditional approach in which lenses form something humans recognize as an image. In nature, some beetles navigate by detecting certain colors or the polarization of light in air without forming an image from the data. Scientists have been slow to explore such alternatives, however, because they have modeled optical instruments such as cameras after our own image-rendering eyes.

The revolution in integrated computational imaging extends beyond just lenses, however. A new trend in hyperspectral imaging is to speed the visual data processing and reduce data storage requirements by downloading some of the computation to the sensing detector itself. In many cases the detector array can perform both feature extraction (of both physical and spectral features) and encoding of these features. The codes are transmitted by the array to the computer, integrating the computation and imaging (ICI) to reduce the huge data load and speed the processing. Similarly, molecular computing in a multiplex image bandpass spectrometer can accomplish hyperspectral imaging as spatial integrated computational imaging performs feature extraction [2].

A simple dueling analogy suggests the advantage of doing as much of the processing as possible in the sensor. Imagine two swordsmen in a in a fight. The first swordsman's hand and sword are controlled by his brain using image information transmitted from the retinas of his eyes. Impulses must travel from his eye to his brain, and then from his brain to his hand. The second swordsman's hand and sword are controlled directly by the retinas of his eyes using nerve impulses that travel only one path instead of two. The second swordsman's weapon is likely to always be a bit ahead of the first's. Moreover, the second swordsman's brain is free to consider other strategies.

Both *spatial* and *spectral* features of samples can be encoded in ICI. When spectral images are simultaneously obtained and encoded at many different wavelengths, the process is termed hyperspectral integrated computational imaging (HICI). Molecular absorption filters can be used as mathematical factors in spectral encoding to create a factor-analytic optical calibration in a high-throughput spectrometer. In this system of molecular computing, the molecules in the filter effectively compute the calibration function by weighting the signals received at each wavelength over a broad wavelength range. Lenslet arrays and masks can also be employed to encode spatial features of a hyperspectral image. Spectrometer designs are possible that use molecular-computing to replace traditional principal component analysis in a computer with molecular filters tailored to produce factor scores at the detector. Spectrometer designs that use lenslet arrays to extract and encode selected image features are also being produced.

Given a set of training spectra collected at all available wavelengths (see Fig. 1, left side), it is possible to rationally select molecular filter (MF) materials to perform PCA (see Fig. 1, right side). PCA is designed to maximize the signals from the spectral regions with the most variability by most heavily weighting them (with the loadings line in left graph in Fig. 1) in calibration. However, PC loadings heavily weight signals in the positive and negative direction, which cannot be done with MFs without offsetting signal gained at one wavelength with signal lost at another wavelength. Because only absolute values can be represented in MFs, two filters are needed for a PC, one for the positive loadings (MF1) and one for the negative loadings (MF2). The filter materials can be selected by examining the sample spectra. The transmission spectrum (%T) of the filter material should be as similar as possible to the absolute value of the loadings spectrum being targeted.

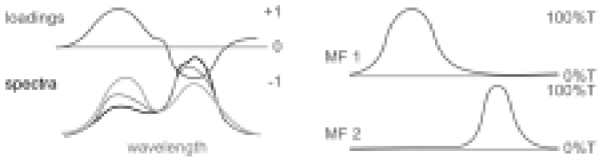


Fig. 1 Principal components (PCs) of spectral data are formed from loadings vectors (left). The highest loadings correspond to wavelengths where the variation of interest in the sample spectra is greatest. The variations can be captured optically by selecting molecular filter (MF) substances with transmission spectra similar to the loadings. If there are positive and negative loadings in the MFC bandpass, two molecular filters must be employed for that PC to avoid ambiguity.

Bandpass filters should be selected to ignore regions of the spectrum where there is no difference between the training spectra, as extra photons in those regions simply saturate the detector or add noise without providing any additional signal. The MF filters do not have to be featureless in the areas away from their peaks in the pictures above as long as bandpass filters (or prisms or gratings) are used to wipe out the %T peaks in undesired areas. In the infrared region, radiation sources like the synchrotron are ideal for near-field microspectrometry with molecular computing because the collimated beam has uniform intensity across the spectrum

.

Current sensor system architectures detect signals from a stimulus, convert them to electrical signals, convert the electrical signals to digital form for processing by computers, and, finally, extract critical information from the processed signals for utilization. Integrated Sensing and Processing (ISP), an initiative launched in the Defense Advanced Research Projects Agency (DARPA), seeks to exchange this chain of processes, each optimized individually, with new methods for crafting sensor systems that treat the total structure as a single end-to-end process that can be optimized globally [3].

The military rationale for ICI parallels the scientific one. In the 21st century global information dominance is necessary to protect U.S. air, space and ground assets. Sensor systems like synthetic aperture radar (SAR) and IR video collect unprecedented amounts of data, greater than 10^{12} pixels/day that require more than 10^{16} flops/day to process. At the same time, the "downsizing" personnel trend persists and the ratio of "pixels to pupils" is heading toward infinity. These trends combine to make training data collection, processing, downlink and distribution all problematic as the U.S. military seeks ways to rapidly reduce data from physical fields to high-level information. At the same time, computing resources are limited in size, weight, power, and cost. Application Specific Integrated Circuits (ASICs) do not really help because they solve a fixed problem in a changing sensor/target environment. ASIC design time and cost tend to be prohibitive. More flexible detection schemes like the digital micromirror array (DMA) measure features, not pixels, under computer control. This holistic approach boosts signal-to-noise ratio (SNR) and concentrates information.

Algorithms for both design and operation of sensor systems are being constructed that permit back-end exploitation processes, such as target identification and tracking, to automatically organize and establish the operating modes of sensor elements to guarantee the most relevant data are always being gathered as circumstances and settings evolve. The ISP program approach is enabling an order-of-magnitude performance enhancement in detection sensitivity and target classification accuracy, with no change in computational cost, across a broad assortment of DoD sensor systems and networks - from surveillance to radar, sonar, optical, and other weapon guidance systems. ISP has produced statistical methods to apportion the sensing channels in a configurable chemical sensor and developed feedback tactics to supervise the elements of an adaptive optical sensing system. ISP has invented new mathematical frameworks for global optimization of design and operation of a number of different types of sensor systems. It is also implementing its software prototypes of the new methodology in test-bed hardware systems, such as missile guidance and automatic ground target recognition modules. ICI will bring these same benefits to chemical analysis.

ICI researchers are conferring extensive depth of field on microscopes and other optical instruments [4]. Optical engineers are developing novel optics to assist computers in sensing motion and the physical and chemical properties of distant objects. Engineers are designing similar lenses that can manage other segments of the electromagnetic spectrum, enlarging the broad transformation in progress in the way scientists look at sensing. Standard cameras, microscopes, and other optical instruments use collections of convex and concave lenses to focus light onto planar sections of film or electronic detectors. For example, an autofocus camera classically moves the positions of certain optical elements forward and backward until a sensor that scrutinizes contrast variations in the field of view perceives satisfactory detail. Eliminating autofocusing and reducing component count begins by considering any scene observed through a lens as a montage of small points of illumination. Paradoxically, abolishing autofocusing systems depends on a defocusing lens. Rather than using a movable convex lens to focus light, a saddle-shaped lens is held stationary. This fixed lens contributes an apparently blurred image to a computer, which then runs a program that rebuilds the image point by point. The product of this procedure, which is termed wavefront coding, is an image with large depth of field (i.e., an image in sharp focus in both the foreground and background) [5].

The extended depth of field, which is at least an order of magnitude larger than it is for regular lenses, does involve compromises. As the computer eliminates the general blurring initiated by the wavefront coding lens, the computer adds a bit of random error in the form of noise. The noise appears as a slight coarsening of shiny and smooth surfaces. Nevertheless, the enhancement of total focus more than compensates for the effect of that noise. Also, supplementary computer processing can filter that noise. New industrial and medical devices that feature the wavefront coding technology include components for microscopes and extended depth-of-field endoscopes.

Wavefront-coding presents a means to reduce the number of aberration-correcting optical elements used in standard cameras and similar instruments because computers can also rectify some lens aberrations as images are de-blurred. Large space telescopes capable of spectrometry of distant planetary atmospheres [6] might be fabricated with relatively lenient construction tolerances by means of wavefront coding technology. The saddle-like lens and other wavefront-coding lenses produced up to now correspond to only a few of the myriad potential forms for computer-adapted optical elements.

Insect eyes also suggest sensing using arrays of miniature traditional lenses, known as lenslets [7,8]. Every lenslet focuses a small, low-resolution image onto a section of an electronic detector array. A computer can determine a single large scene at approximately twice the resolution than would be achievable if one traditional lens had been employed by manipulation of all of the lenslets' different viewpoints. A specific benefit of this method is that the thin lenslet array can focus light onto a detector less than a millimeter away. This extreme contraction of focal length has been used to establish a model camera as slim as a microscope slide. A number of other exceptionally thin cameras must employ tricks such as reflecting light off internal mirrors to achieve the necessary focal length inside a miniature container.

Other lenslet arrays are less pretentious, using merely apertures in place of lenses. For instance, a small polymer block packed with correctly angled holes allows photodetectors behind the block to collect light from a scene simultaneously from different viewpoints. The outcome is a tool that can reconstruct the movement of an enemy asset like an armored personnel carrier (APC) without acquiring or analyzing any images of the APC. A similar technique could be applied to cells under a microscope. Most contemporary motion-tracking mechanisms acquire images of a 2-D field and then analyze pixel patterns in pursuit of changes representing movement. This search is a protracted, computer-intensive process predisposed to errors. Using innovative aperture array devices, light from a selected target strikes detectors and forms a unique optical code from which a computer can quickly recreate movement with negligible computation.

Other optical elements are intended for concurrently recording spectra across the pixels of a full field of view. Such hyperspectral data may expose camouflaged missiles in a satellite image. Hyperspectral data can also reveal biological activities [9,10], often with the aid of fluorescent labels that bind to special cellular structures. A spectra-capturing lens, or filter like a linear gradient filter, yields a pattern in which a multicolor spectrum connected with every point in a field of view is mapped onto a detector. The pattern is not an image at that juncture, only a confusion of colors and pixels. However, sorting the data in a computer transforms this apparent disharmony into an image of the field of view at any selected wavelength. Hyperspectral data have become one of the principal methods by which scientists analyze the physical and chemical properties of sample targets ranging from atoms to Martian landscapes. ICI cameras that perform at infrared wavelengths for military surveillance and biological studies are now in development, as well as ICI cameras that use ultraviolet frequencies for studying fluorescently tagged biological samples.

Our lab formerly employed an ordinary InSb focal plane array infrared camera to monitor metabolism in rats [11]. In acute dosing experiments, rats are given drugs and their physical activity and thermogenesis are monitored. Six students with video recorders and computer workstations were needed to go through the video frame-by-frame to measure the movement of the rats (integrating the motion of center of mass) and heat emission (thermogenesis). In spite of all of these people and equipment, it took weeks to analyze the data collected from a one-hour experiment. Now, lenslet array cameras can measure motion and heat emission with greater accuracy (e.g., better correlation to oxygen consumption measurements, considered the “gold standard” for metabolic measurement) than can be obtained by circling image areas and measuring photon emission. The reason for the increased accuracy is that there is more variation in how different students interpret rat motion and thermal emission than there is variation in the lenslet arrays. Furthermore, there is more variation in how the same student makes such measurements over days and weeks, than in how an ICI array encodes rat motion and thermal emission.

Differentiating between basal metabolism and work-related thermogenesis is important in studies of anti-obesity drugs because different drugs can increase metabolism in different ways. Some drugs increase metabolic rate and body temperature by agitating the subject and producing hyperactive behavior. Other drugs are pyrogenic and increase body temperature, but make the subject feel feverish and sick. The two types of reaction lead to very different movement behaviors.

Fig. 2 shows how lenslet array cameras can capture accurate temperature and motion data from a freely moving rat in a cage. A simple ICI camera with three apertures and only one IR detector behind the aperture mask easily permits rat motion and thermal emission to be measured by positioning four cameras around the cage: two on the x-axis (1.0 and 1.1, with lines of sight shown with solid lines), and two on the y-axis (0.0 and 0.1, dotted lines of sight). Each camera detector has three lines of sight that together cover a triangular zone in the cage. Two additional ICI cameras on the z-axis (2.0 and 2.1, with cylindrical lenses yielding planar fields of view represented by solid and dotted lines) monitor the height of standing and climbing movements.

When a rat crosses in front of one of these cameras (see Fig. 2b), it can cross the field of view either close to the camera, or farther away. When the rat crosses close to the camera, it can illuminate all three apertures simultaneously. If the rat is very close to the camera, the signal increases in a stepwise fashion in relatively large steps. If the rat is not as close to the camera, the signal (with intensity *I*) increases in a stepwise fashion in relatively small steps.

When the rat crosses the field of view from very far away, it illuminates only one aperture in the mask at a time (see Fig. 2c, top row). Baseline detector drift can be a problem with thermal detectors. One way to eliminate excess drift is to modulate the signal, which in this case requires either active excitation or chopping the light at the aperture mask. Another way to correct detector baseline signal-drift is to use a detector amplifier operated in a high-pass filter configuration. In this configuration, the detector amplifier is differentiating, and the signal passed from the camera to the computer is the derivative of the stepwise signal (*dI/dx*). A simple low-speed A/D is all that is needed to interpret the camera data and calculate energy expenditure through basal metabolism and through work (work is done by the rat as it moves it body around in the cage).

Using the derivative of the signal, there is no confusion between a rat crossing the field of view close to the camera or far away because, when the rat crosses close to the camera, there are no negative signal excursions until the rat begins to walk away. In contrast, when a rat crosses the field of view from far away, every positive signal excursion is followed immediately by a negative one.

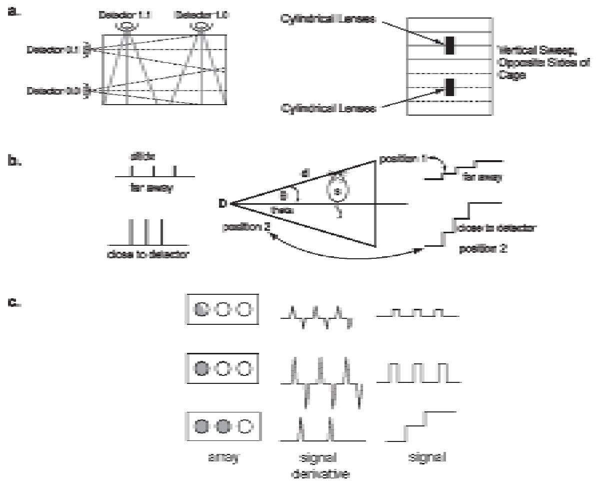


Fig. 2. **a.** Positions of six ICI lenslet cameras monitoring a rat metabolic cage. **b** spatial codes generated by a rat moving in front of a single ICI camera. **c.** (top) codes generated as a rat moves across the array very far away, (middle) as a rat moves across the array closer, (bottom) and as a rat moves across the array very close to the camera.

ICI requires only simple computing because the feature extraction and encoding done by the detector array. In contrast to calculating Longbow classifiers, Fourier-Mellin transforms, SVDs, and PCAs with powerful digital computers, in ICI simple digital manipulations and analog detection permit low-cost, rapid identification of targets. To provide an example, let us imagine that a target of known size is shaped like a rectangle. A lens projects an image of the target on a digital micromirror array. When DMA elements are 'on', they reflect light into a detector. When the elements are 'off', they reflect light into a beam dump. Starting at the center of the DMA, a small rectangle is turned on first, then its elements are turned off and the next largest size rectangle’s (the target shape) elements are turned on. This process is continued progressively until the target shape completely fills the DMA. Somewhere in this sequence the rectangular image matches the rectangular mirror elements that are turned on. When the DMA target shape and size exactly match the target image projected on the array by the lens, a signal spike appears at the detector. At this point the ICI system has identified the target through selective feature extraction. In addition, the range to the target has been calculated because the scale of the actual target is known, and the range is identified by size of the rectangle creating the signal spike at the detector.

The same DMA technology can be employed to compute principal components optically. For example, using a lens to collimate light for a transmission grating, a spectrum can be projected across a DMA. In this example, the columns of the DMA represent wavelength, and the rows reflect a fraction of light (between zero and one) collected at each wavelength. Principal components are simply a weighted sum of light intensities across a range of wavelengths, and it is easy to turn on a fraction of the mirrors in each column to represent the weighting at each wavelength. By employing this scheme, the signal observed at a single analog detector can be easily related to a principal component score, and even to an analyte concentration. Thus, a relatively complex computational function is reduced to measuring the voltage on a single detector, taking ICI full circle from spatial encoding back to spectral encoding.

Analytical and bioanalytical research is often data rich but information poor. Data threaten to cause a bottleneck in genetic research, drug discovery and development, and other areas. Research depends increasingly upon multidimensional images like hyperspectral images, 3-D multiwavelength confocal images to expose sites of gene expression, or time lapse videos to investigate cell behavior. At least four areas of research are important in image informatics: (1) database design and metadata structures for multidimensional images, (2) automated object tracking and event recognition in videos, (3) interactive analysis and query by content of videos, and (4) web-based video editing and customization. ICI can help with the first three areas of research. In databases, specific intrinsic metadata are needed, relating to the locations and timing of individual items and events within images and videos. ICI spectral and spatial encoding can serve as intrinsic metadata for databases. For automated object tracking and event recognition the amount of image data being produced is increasing exponentially, while our ability to absorb and process this information remains nearly static. Vast quantities of valuable information contained within image data may thus be lost because of time constraints and lack of objectivity in the human interpretation of content. The process of automated video analysis involves three stages. First, ICI techniques are employed to identify objects in the digitized videos, to determine their sizes and orientations, and to track the changes in their positions over time. Next, events are identified by specific image understanding procedures that read the spatial and spectral codes. Finally, the metadata are automatically stored, and subsequently may be used for query by content and video retrieval. Where the image content is too complex for fully automated analysis, content analysis may be handled in a computer assisted interactive manner, in which the user makes the value judgments but software assists the processes of metadata definition and storage.

ICI has many applications, such as in coronary catheters, where physicians do not have much time to locate vulnerable atherosclerotic plaques. ICI has many military applications for the same reason – the available time to identify correctly many different targets is short. The data-analysis capabilities of computers continue to grow in accordance with Moore’s Law. Moreover, significant developments in mathematical tools and innovations in optics manufacture permit more complex components to be made, such as molecular computing fiber optics and lenslet arrays. With the union of the latest computers and innovative optics, ICI is ready to reveal a universe of possibilities that have been concealed from the human eye.

ACKNOWLEDGEMENTS

This research was funded in part by KSEF-333-RDE-003, the NIH through HL58927, the NSF through DGE-9870691, and NIH and SESI through SC-NIAAA-93-01. The research is being carried out in part at the National Synchrotron Light Source, Brookhaven National Laboratory, which is supported by the U.S. Department of Energy, Division of Materials Sciences and Division of Chemical Sciences, under Contract No. DE-AC02-98CH10886.

REFERENCES

[1]. Weiss, P. New lenses create distorted images for digital enhancement. (2003) Science News, 163(13):200.

[2]. Cassis, LA, Dai B, Urbas, A, Lodder, RA. In vivo applications of a molecular computing-based high throughput NIR spectrometer. (2004) Progress in Biomedical Optics and Imaging, in press.

[3]. DARPA (2002) http://www.darpa.mil/

[4]. CDM Optics (2004) http://www.cdm-optics.com/site

[5]. Bains, S. Wavefront coding finds increasing use. (2004) Laser Focus World, 40(1). http://lfw.pennnet.com/

[6]. Terrestrial Planet Finder Book, NASA, 1999. http://planetquest.jpl.nasa.gov/TPF/tpf_book/index.html

[7]. Lewis, MF., Wilson, RA. The use of lenslet arrays in spatial light modulators. (1994) Pure Appl. Opt. 3:143-150.

[8]. Barge, M., Hamam, H, Defosse, Y. Chevallier, R., deBougrenetdelaTocnaye, J L . Array illuminators based on diffractive optical elements. (1996) J. Opt. 27:151-170.

[9]. Liu Y, Windham WR., Lawrence KC, Park B. Simple algorithms for the classification of visible/near-infrared and hyperspectral imaging spectra of chicken skins, feces, and fecal contaminated skins. (2003) *Appl Spectrosc.* 57(12):1609-1612.

[10]. Gillies R, Freeman JE, Cancio LC, Brand D, Hopmeier M, Mansfield JR. Systemic effects of shock and resuscitation monitored by visible hyperspectral imaging. (2003) Diabetes Technol Ther. 5(5):847-855.

[11]. Buice RG Jr, Cassis LA, Lodder RA. Near-IR and IR imaging in lipid metabolism and obesity. Cellular and Molecular Biology (1998) 44(1): 53-64

Robert A. Lodder received his B.S. degree cum laude in Natural Science from Xavier University (Cincinnati, Ohio) in 1981. After deciding to pursue a career in chemistry he worked under Professor Richard T. O'Neill at Xavier, and received his M.S. in Chemistry in 1983. He received his Ph.D. in Analytical Chemistry in 1988 after working with Professor Gary M. Hieftje at Indiana University, and is currently an Associate Professor of Pharmaceutical Sciences at the College of Pharmacy, University of Kentucky Medical Center. Dr. Lodder holds a joint appointment in the Division of Analytical Chemistry of the Department of Chemistry at Kentucky. He serves as Editor of the astroanalytical chemistry and astrobiology journal *Contact in Context*, and as a member of the U.S. Food and Drug Administration Advisory Committee on Pharmaceutical Science, Process Analytical Technologies subcommittee. Dr. Lodder is a first-prize winner in the 1990 international *IBM Supercomputing Competition*, as well as a winner of a National Science Foundation *Young Investigator Award*, the *American Society of Agricultural Engineers Paper Award*, a *Buchi NIR Award*, the *Tomas Hirschfeld Award in Near-IR Spectroscopy* and a *Research and Development 100 Award*.

Bin Dai is a graduate student at the University of Kentucky, working in Dr. Lodder’s lab as part of the Analytical Spectroscopy Research Group.

A Binary Partitioning Approach to Image Compression Using Weighted Finite Automata for Large Images

Ghim Hwee Ong, Kai Yang
Department of Computer Science, School of Computing
National University of Singapore, Singapore

ABSTRACT

Fractal-based image compression techniques give efficient decoding time with primitive hardware requirements, which favors real-time communication purposes. One such technique, the Weighted Finite Automata (WFA) is studied on grayscale images. An improved image partitioning technique — the binary or bin-tree partitioning — is tested on the WFA encoding method. Experimental results show that binary partitioning consistently gives higher compression ratios than the conventional quad-tree partitioning method for large images. Moreover, the ability to decode images progressively rendering finer and finer details can be used to display the image over a congested and loss-prone network such as the Image Transport Protocol (ITP) for the Internet, as well as to pave way for multi-layered error protection over an often unreliable networking environment such as the UDP.

Key Words: binary partitioning approach; image compression; weighted finite automata

1. INTRODUCTION

Data compression has been an important issue in relation to transmission and storage of information for a long time. In particular, digital images require a large amount of space in their internal representation often as arrays of pixels, incurring too much storage space and transmission time over a communication channel. Techniques for finding more compact image representations are desired. Also, general-purpose and selectively-reliable transport protocols have been developed to transmit compressed images over the Internet [1].

Many image compression techniques have been developed to suit different applications. Lossy image compression techniques can provide high compression ratios but introduce some annoying artifacts. Fractal-based image compression is one of the lossy data compression techniques that have been developed in the last decade. It makes use of self-similarities existing in an image at different resolutions under a set of affine transformations and exploits this kind of redundancy in order to alleviate the high cost of storage, thus achieving compression.

The WFA method attempts to represent a sub-image as a weighted sum of other sub-images to accomplish compression. WFA starts with an image to be processed. It locates sub-images that are identical or very similar to the entire image or to other sub-images, and subsequently constructs a graph that reflects the relationship between these sub-images vis-à-vis the entire image. The various components of the graph are compressed and the final product is a much more compactly represented domain.

An important issue involved in fractal-based image compression is the partitioning of an image into sub-images or blocks for encoding. There are various approaches to the use of different block shapes. One such commonly used approach is based on quad-tree partitioning. It has also been shown that a partitioning scheme based on rectangular blocks is simple yet offers high quality of decomposed images [2]. In this paper, we investigate a simple, new approach to WFA-based image compression with binary partitioning. In particular, we seek to improve the compression ratio by introducing intermediate steps into quad-tree partitioning. We will show that binary partitioning gives higher compression than quad-tree partitioning for images of 512×512 and 1024×1024 pixels.

2. WEIGHTED FINITE AUTOMATA FOR IMAGE COMPRESSION

First we introduce WFA as a modelling tool for specifying greyscale images, and then we describe inference algorithms in the next sections for their construction.

The WFA accepts grayscale pictures containing $2^m \times 2^m$ ($m \in Z^+$) pixels where pixel intensity ranges from 0 to 255. Furthermore, a WFA specifying an image does not merely search for one of the known states that best matches the sub-state under investigation; but instead, it uses a linear combination of (possibly all) known states to arrive at a better approximation [3, 4].

An *m*-state WFA *A*, over alphabet Σ is specified by:
(1). A row vector $I^A \in R^{1\times m}$ (initial distribution)
(2). A column vector $F^A \in R^{m\times 1}$ (final distribution)
(3). Weight matrices $W_a^A \in R^{m\times m}$ for all $a \in \Sigma$.

To display WFA using diagrams, represent the *m* states by circles; the initial (first value) and final (second value) distribution being shown inside each state. If $(W_a)_{i,j} \neq 0$, place an edge from state *i* to state *j*, labeled by $a((W_a)_{i,j})$. For example, suppose that we have the initial distribution $I^A = (1 \quad 0)$, final distribution $F^A = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$; weight matrices:

$$W_0 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}, W_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ 0 & 1 \end{pmatrix}, W_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ 0 & 1 \end{pmatrix}, W_3 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}.$$

Each state represents a portion of the image and the edges attached to a state represent the linear combination of states which, when calculated, produce an approximation of the image portion. The diagram of such a WFA is given in Figure 1 below for illustration.

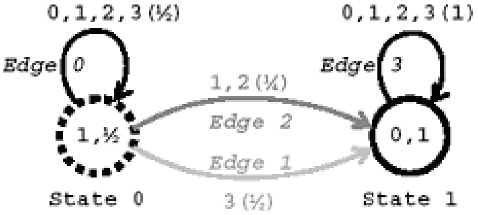


Figure 1. Diagram representation of the WFA

Given the WFA A over alphabet Σ shown in Figure 1, the initial distribution $I^A = (1 \quad 0)$ is obtained from the first value inside each state, ordered by the state number; similarly the final distribution $F^A = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$ is obtained from the second value inside each state.

When an edge exists from state *i* to state *j* (*i* and *j* may be the same state), place the value enclosed in parenthesis in position (*i, j*) of matrix W_a^A where *a* is the label associated with the edge. For example, the bottom edge joining state 0 to state 1 has label "3"and a value of ½; in position (0,1) of weight matrix W_3^A we place the value ½. To visualize the actual image that results from this WFA, refer to Figure 2 shown in the following column.

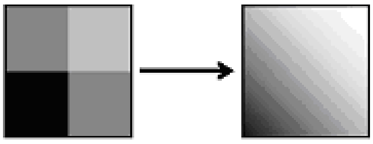


Figure 2. Actual images being described by this WFA

3. THE QUAD-TREE APPROACH TO WFA IMAGE ENCODING

A multi-resolution image is specified by assigning the grayscale value 0 to 255 to every node of the infinite quad-tree. If the outgoing edges of each node of the quad-tree are labeled 0, 1, 2, 3 we get a uniquely labeled path to every node; its label is called the address of the node. The address of a node at depth k is a string of length k over the alphabet Σ = {0,1,2,3}. Hence, a grayscale multi-resolution image can be specified as a subset of strings over the alphabet Σ. Regular sets of strings are specified by finite automata; therefore, finite automata can be used to specify regular multi-resolution images.

If a finite automaton *A* represents an image *I*, then each state of *A* must correspond to a sub-square of *I*, with the initial state corresponding to the entire *I*. Moreover, if there is a transition from state *i* to state *j* labeled by 0 (or 1 or 2 or 3), then the image corresponding to state *j* is the SW (or NW or SE or NE respectively) quadrant of the image corresponding to state *i*. An example below in Figure 3 shows the recursive zooming of the quad-tree.

11	**13**	**31**	**33**
10	**12**	**30**	**32**
01	**03**	**21**	**23**
00	**02**	**20**	**22**

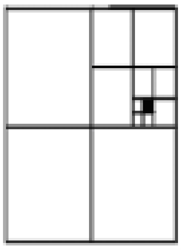


Figure 3. The addresses of the quadrants of the 4×4 square, and the sub-square specified by 3203

From the previous section we know that in order to arrive at a good approximation of any state image, a linear combination of some known states is employed. In mathematical terms, the crucial part of the WFA construction algorithm lies in representing a given vector by linear combinations of other known vectors. Since our goal is data compression, we have to approximate a vector with as few others as possible and as accurately as possible. Obviously these two requirements have to be balanced. The approximation with matching pursuit (MP) was introduced by Mallet and Zhang [5]. In this technique, an approximation is constructed step by step using a greedy strategy.

The MP approximation of a vector $\vec{v}$ is constructed by selecting the best matching vector from a given codebook. The component of this best matching vector is then subtracted from $\vec{v}$. The residue is now encoded in the same way as $\vec{v}$ and the process continues until a given abortion criterion is fulfilled.

Let $p, n \in N$ and $D = \{g_0, \ldots, g_{p-1}\}$ with $g_i \in R^n$ ($i \in \{0, \ldots, p-1\}$) be a set of $p \geq n$ non-zero vectors. The set D is often called the dictionary, codebook or domain pool. With span (D) we denote the set of all linear combinations of vectors in D. In order to make the following calculations a bit easier we assume that, without loss of generality, each vector in the dictionary has unit Euclidean norm $\|\vec{v}\|$. If span $(D) = R^n$ then D contains a set of n linearly independent vectors. The matching pursuit algorithm begins by projecting $\vec{v}$ on a dictionary vector g_{i_0} and computing the residue Rv by:

$$\vec{v} = \langle \vec{v}, g_{i_0} \rangle g_{i_0} + Rv$$

Since Rv is orthogonal to g_{i_0} (by virtue of the vector inner product), the following equation holds:

$$\|\vec{v}\|^2 = \left| \langle \vec{v}, g_{i_0} \rangle \right|^2 + \|Rv\|^2$$

Hence, we have to choose g_{i_0} such that $\left| \langle \vec{v}, g_{i_0} \rangle \right|$ is maximized, since residue $\|Rv\|$ has to be minimized. The next iteration continues with Rv instead of $\vec{v}$, and we compute the following iteration:

(1) Set $R^0 v = \vec{v}$

(2) Subsequent residues are computed by solving simultaneously:

$$\begin{cases} R^m v = \left\langle R^m v, g_{i_m} \right\rangle g_{i_m} + R^{m+1} v \\ \left\| R^m v \right\|^2 = \left| \left\langle R^m v, g_{i_m} \right\rangle \right|^2 + \left\| R^{m+1} v \right\|^2 \end{cases}$$

By choosing g_{i_m} to maximize each of the $\left| \left\langle R^m v, g_{i_m} \right\rangle \right|$.

(3) Finally, summing up the last equation in m from 0 to a stopping index $M-1$ yields:

$$\vec{v} = \sum_{m=0}^{M-1} \left\langle R^m v, g_{i_m} \right\rangle g_{i_m} + R^M v$$

4. THE BIN-TREE APPROACH TO WFA IMAGE ENCODING

The discussion so far has been under the assumption that the WFA is approximated by recursively dividing into 4 state image quadrants, each zooming to a higher level of detail. However, each of the quad-tree partition looks at only 25% effective area of the original state, much of the spatial correlations are lost in the division process.

Seeing in this light, we propose a modification of WFA image approximation using binary partitioning. The idea is to first divide the current (square) state q horizontally into 2 (rectangular) sub-states q'_0 and q'_1, and later divide q'_0 and q'_1 again vertically to produce 4 (square) sub-states q''_{00}, q''_{01}, q''_{10} and q''_{11} just as what a single division of the quad-tree partitioning would have done. This effectively separates the previous quad-tree partitioning into two steps: a horizontal cut followed by a vertical cut. The rational behind this separation is to better utilize the spatial correlations among image segments, by looking at 50% effective area. An example is given below in Figure 4. Instead of mechanically following the horizontal and then vertical cut to every (square) state image, a further enhancement tests both alternatives and picks the better choice. However, this involves some back-tracking.

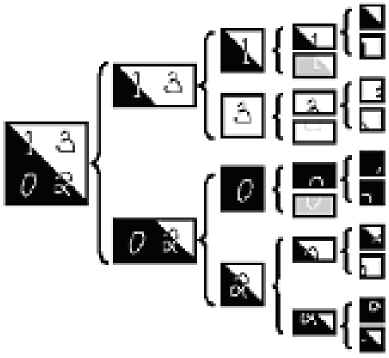


Figure 4. This image is cut twice: horizontally followed by vertically (redundant sub-images are in pale shades).

An outline of the WFA construction algorithm is presented below: Suppose that an image has been partitioned into a set of non-overlapping regions. Starting with the complete image, the current range image is recursively inspected if it can be approximated with a linear combination of arbitrary known sub-images. If a linear combination yields only a poor approximation, this range image will be bin-tree divided again and the recursion continues, while a new state image is appended to the WFA under construction. On the other hand, if this approximation already satisfies the given quality threshold, the corresponding transitions are appended to the WFA and the recursion terminates.

The main task of the image regeneration involves calculation of all the range images. This only requires some simple manipulation for each range image:

$$\psi^A = \sum_{i=1}^{M} w_i \cdot W_i^A$$

Where w_i are the (quantized) weights and W_i^A are the weight matrices for automaton A in the linear combination of M domain images. Every range image must not only be computed at its fixed size in the original image but also at any desired size in the linear combinations. Moreover, domain images that are referred to more than once are cached (stored in an array) to avoid multiple computations of the same image segment.

5. RESULTS AND DISCUSSIONS

Four gray scale images (shown in Figure 5), which are commonly used in the community of image processing and compression, were used to provide the image sample data with which to experiment in this paper. Tests were then conducted on these images of two sizes, 512 × 512 and 1024 × 1024 pixels, using binary partitioning and the quadtree partitioning methods. The evaluation of the proposed partitioning method is made on the basis of the results of compression, image quality, and CPU processing times on the four images.

The compression ratio is defined as the ratio of the number of bits in the original image file to the number of bits in the compressed file. The quality of images is determined by the Peak Signal-to-Noise Ratio (PSNR), which determines the difference between two images. It is defined as

$$\text{PSNR} = 20 \log_{10} \left(\frac{\text{b}}{\text{rms}} \right) \text{ dB.}$$

where b is the largest possible value of the signal or pixel value, and rms is the root-mean-square difference between two images. The PSNR is given in units of decibel (dB), which measure the ratio of the peak signal and the difference between two images. That is to say, the higher the value, the better the image quality.

Figure 5. Four images used in the test

Experimentally, it is found that the compression ratios are nearly independent of the initial vertical or horizontal rectangles chosen for binary partitioning. For each image size, the compression results are measured with both the binary and quadtree partitioning methods for PSNR values ranging from 25 to 34. Figure 6 depicts the average compression ratio versus PSNR results for the four images.

It is observed that the binary partitioning method consistently gives better compression than the quadtree method. This confirms that there are good matches for the rectangular blocks and the encoding of these blocks (instead of splitting into squares) gives better compression. The results show that the improvement in the compression ratio increases with the size of the image. On the average, the compression ratio improves from about 7 to 22%.

Experimentally, it is observed that the binary partitioning method takes a little longer to encode than the quadtree method. This is due to the additional steps taken by binary partitioning to search for some best matches. The decoding times for both the methods are found to be comparable.

6. CONCLUSION

In this paper, we have presented a binary partitioning method for the WFA-based image compression. A square image is split into rectangular and square blocks with binary partitioning of the image. Experimental results show that binary partitioning gives higher compression results than quad-tree partitioning for large images. The improvement can be as high as about 22% for 1024×1024 images. From the transmission perspective, the proposed binary partitioning approach to WFA-based image compression can be designed to decode partially received, out-of-order image data. This scheme can be customized to couple.

with the Image Transport Protocol for image transmission over loss-prone congested or wireless networks

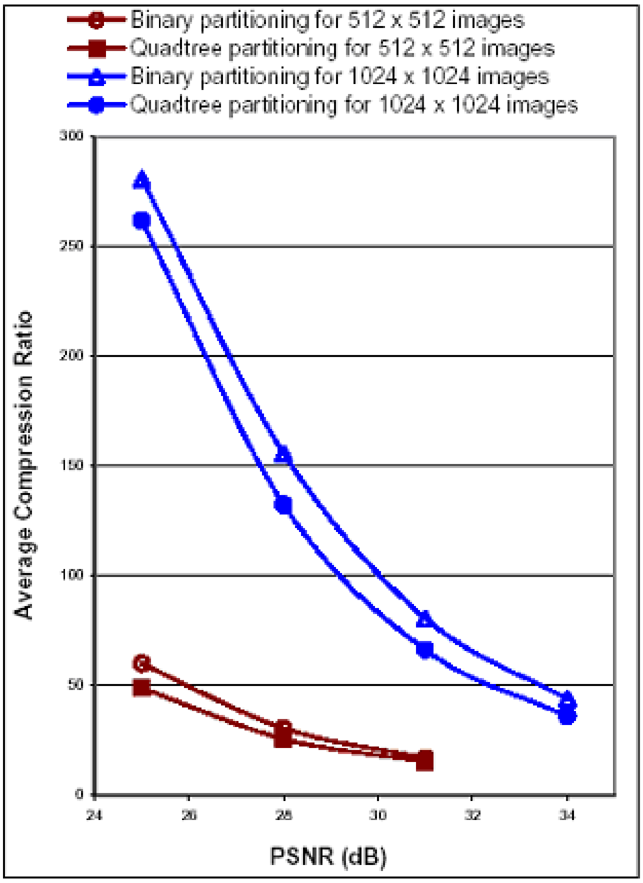


Figure 6. Average compression ration versus PSNR results for the four images

7. REFERENCES

[1] Raman, S., Balakrishnan, H., and Srinivasan, M. *An Image Transport Protocol for the Internet*. 2000 International Conference on Network Protocols, Japan, November 2000.

[2] Ponomarenko, N., et al. *Modified Horizontal Vertical Partition Scheme for Fractal Image Compression*. 5th Nordic Signal Processing Symposium, Norway, October 2002.

[3] Karel Culik II and Jarkko Kari. *Image Compression Using Weighted Finite Automata*. Computer and Graphics 17, 305-313, 1993.

[4] Karel Culik II and Jarkko Kari. *Efficient Inference Algorithm for Weighted Finite Automata*. Fractal Image Compression, ed. Y.Fisher, Springer-Verlag, 1994.

[5] S. G. Mallet and Z. Zhang. *Matching Pursuits with Time-Frequency Dictionaries*. IEEE Signal Processing, Vol. 41, No. 12, December 1993.

A Fractal Rotating Vector Algorithm of Radar Echo Image Plotting *

Dan Liu, Dayong Zhang
Institute of Nautical Science and Technology, Dalian Maritime University
Dalian, Liaoning 116026, China
Email: dliudlmu@newmail.dlmu.edu.cn **Tel.:** (0411)84729651

ABSTRUCT

Natural coastline is an intricate curve, which is in and out, discontinuous and has fractional dimension. In the radar echo image-plotting test, in order to generate complicated coastline with arbitrary details through the sampling points gained from digital instrument, fractal algorithm is the best choice. A fractal algorithm called fractal rotating vector algorithm is given, it can generate the coastline data for radar echo image with arbitrary level detail structure.

Keywords: Fractal, radar echo image, radar plotting

1. INTRODUCTION

At present,the original data of generating Virtual Radar image is all captured from chart by digital instrument,this method collects a series points on the coastline,then approach the coastline using the broken lines generated by joining the two adjacent points. Using straight line to approach curve to generate coastline by exists at least two disadvantages [1]:

(1)The shape of the coastline is too regular, which means that it is too different from the truth.
(2)The quantity of the data needed to be stored is much bigger. When use the straight line is used instead of curve to generate radar image, in order to gain good 3D effect, the only way can be adopted is to increase the sampling frequency of the curve, that is to say, to gain the 3D effect by adding sampling points. Since it will make the amount of the data needed to be stored increase, the cost is the memory.

If we use fractal algorithm to fine the coastline, we will get coastline data with arbitrary details structure and not increase the number of the sampling point.

1. THE CHARACTERS OF FRACTAL GRAPHICS

Basic characters of Fractal graphics [2]:

1. The quantity of original data is small, which is that we can get very fine and complicated pictures using less original data;
2. Have higher level complexity,that is having arbitrary details structure;
3. The ways generated by computer is simple. as we all known, some complicated Fractal picture is produced by a simple model after several iterations;
4. Has self-similarity, which can be approximate or statistic similarity.

2. THE DESCRIPTION OF FRACTAL ALGORITHM

Most fractal algorithms make use of the statistic similarity of the coastline, such as fractional Brownian motion model; the difference of these fractal algorithms is different probability distributions obeyed by the random numbers introduced. The fractional Brownian motion model pays more attention to the statistic similarity of the coastline on large scale rather than on small scale.

Although the coastline doesn't have strict self-similarity and recursion order on the whole interval (large scale) just like Cantor set,it has quite good self-similarity on small scale. The self-similarity on small scale can be thought as the relativity on small scale (not always linear).

Take into account; the information that can describe the trend and the shape of the coastline best should be among the data surrounding the coastline, not are the random ones based on the "statistic rule" of the whole coastline [3].

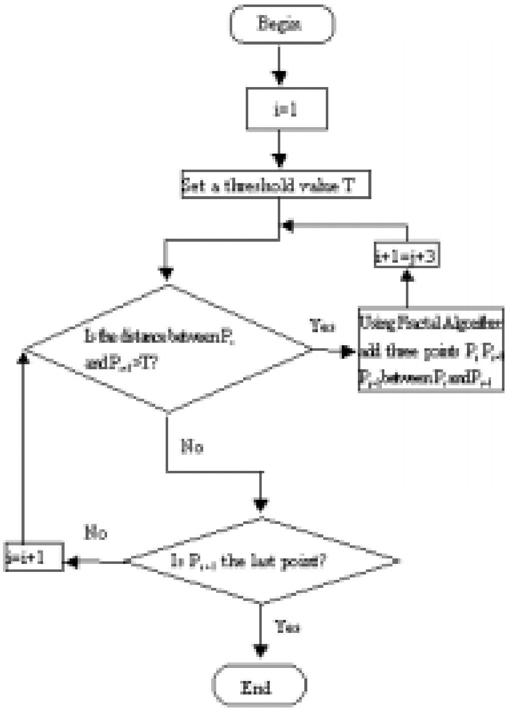
First of all, set a threshold value,which is the maximum of the lengths of all the lines. Secondly, search the whole chain table recording the coordinates of the coastline,if there is a point, the length between it and the next point is bigger than the given threshold value, add another three points between the two points by fractal algorithm. Thirdly, continue to search the next data node of the chain table.

Suppose the original data points are P_1,P_2,P_3,$P_4$and P_5,the unit of the horizontal Axis is 100,the given threshold value is 200, V_{ij} is the vector from the ith point to the jth point,D_{ij} is the distance between the ith point and the jth point. See the figure 1,only the distance value of P_3 is bigger than 200,which is D_{34}.

Deal with the data of the previous and next points of P_3. Define the dashed vector V_{15} be standard vector. Compute the matrixes of affine transforms between standard vector V_{15} and vector V_{12},V_{23},V_{34}. The standard vector V_{15} counter clock wisely rotates angle φ to vector V_{12},so the rotatory matrix between vector V_{15} and vector V_{12} is:

* This paper is supported by Fok Ying Tung Education Foundation(91077).

Table 1. The data structure of PINFO


Begin
i=1
Set a threshold value T
i+1=i+3
Is the distance between P_i and P_{i+1} >T?
Yes
Using Fractal Algorithm add three points P_i P_{i+1} P_{i+2} between P_i and P_{i+1}
No
No
Is P_{i+1} the last point?
i=i+1
Yes
End

$$\begin{vmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{vmatrix} \qquad (1)$$

the rotatory zoom matrix between vector V_{15} and vector $V_{12}$is:

$$\begin{vmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{vmatrix} * \frac{D_{12}}{D_{15}} \qquad (2)$$

inner-product of vector V_{15} and vector V_{12} is:

$$V_{15} \cdot V_{12} = D_{12} * D_{15} * \cos\varphi \qquad (3)$$

outer-product of vector V_{15} and vector V_{12} is:

$$V_{15} \times \ V_{12} = D_{12} * D_{15} * (-\sin\varphi) \qquad (4)$$

from formula (3),(4),we have:

$$\cos\varphi = \frac{\mathrm{V}_{15} \cdot \mathrm{V}_{12}}{D_{12} * D_{15}} \qquad (5)$$

$$\sin\varphi = \frac{-\mathrm{V}_{15} \times \mathrm{V}_{12}}{D_{12} * D_{15}} \qquad (6)$$

so the rotatory zoom matrix between vector V_{15} and vector V_{12} is:

$$\begin{vmatrix} V_{15} \cdot V_{12} & -V_{15} \times V_{12} \\ V_{15} \times V_{12} & V_{15} \cdot V_{12} \end{vmatrix} * \frac{1}{D_{15}^2} \qquad (7)$$

in the same way, the rotatory zoom matrix between vector V_{15} and vector V_{23} is:

$$\begin{vmatrix} V_{15} \cdot V_{23} & -V_{15} \times V_{23} \\ V_{15} \times V_{23} & V_{15} \cdot V_{23} \end{vmatrix} * \frac{1}{D_{15}^2} \qquad (8)$$

the rotatory zoom matrix between vector V_{15} and vector $V_{34}$is:

$$\begin{vmatrix} V_{15} \cdot V_{34} & -V_{15} \times V_{34} \\ V_{15} \times V_{34} & V_{15} \cdot V_{34} \end{vmatrix} * \frac{1}{D_{15}^2} \qquad (9)$$

From formula (7), (8) and (9) on vector V_{34} respectively,gain three new vectors,shown as the thin real line in the figure 1:

$$V_{P_3 A_1} = \begin{vmatrix} V_{15} \cdot V_{12} & -V_{15} \times V_{12} \\ V_{15} \times V_{12} & V_{15} \cdot V_{12} \end{vmatrix} * \frac{1}{D_{15}^2} * V_{34} \qquad (10)$$

$$V_{A_1 A_2} = \begin{vmatrix} V_{15} \cdot V_{23} & -V_{15} \times V_{23} \\ V_{15} \times V_{23} & V_{15} \cdot V_{23} \end{vmatrix} * \frac{1}{D_{15}^2} * V_{34} \qquad (11)$$

$$V_{A_2 A_3} = \begin{vmatrix} V_{15} \cdot V_{34} & -V_{15} \times V_{34} \\ V_{15} \times V_{34} & V_{15} \cdot V_{34} \end{vmatrix} * \frac{1}{D_{15}^2} * V_{34} \qquad (12)$$

So the coordinates of the three new added points are:

$$\begin{cases} x_{A_1} = x_{P_3} + x_{V_{P_3A_1}} \\ y_{A_1} = y_{P_3} + y_{V_{P_3A_1}} \end{cases} \qquad (13)$$

$$\begin{cases} x_{A_2} = x_{A_1} + x_{V_{A_1A_2}} \\ y_{A_2} = y_{A_1} + y_{V_{A_1A_2}} \end{cases} \qquad (14)$$

$$\begin{cases} x_{A_3} = x_{A_2} + x_{V_{A_2A_3}} \\ y_{A_3} = y_{A_2} + y_{V_{A_2A_3}} \end{cases} \qquad (15)$$

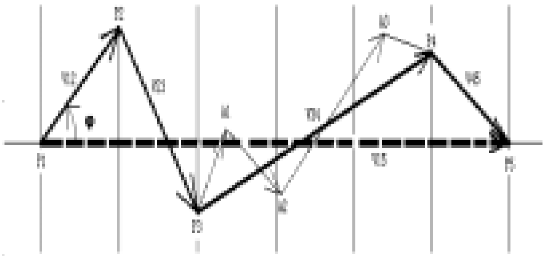


Figure 1. Sketch map of the Fractal algorithm

Remodify and save the chain table of the coordinates of the coastline, insert point A_1,A_2,A_3 between point P_3 and point P_4,remodify the forward and backward pointer of each node, and compute the vector and distance value of each node,the serial number value No of each point behind P_3 adds 3. Search each node behind point P_4, if there is a point, whose distance is bigger than the given threshold value distance, repeat the upwards steps.

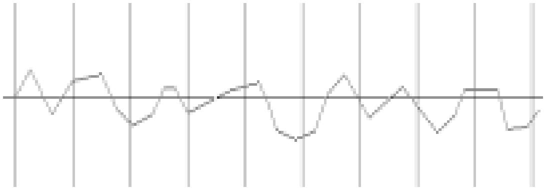


Figure 2. Original data chart

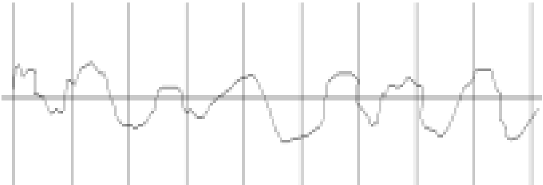


Figure 3. The fined graph when the given threshold value is 30

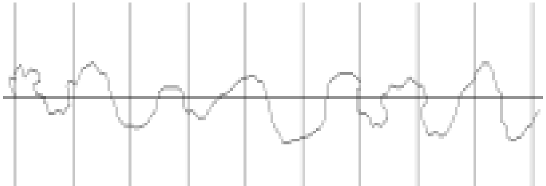


Figure 4. The fined graph when the given threshold value is 20

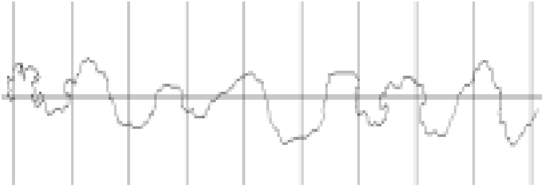


Figure 5. The fined graph when the given threshold value is 2

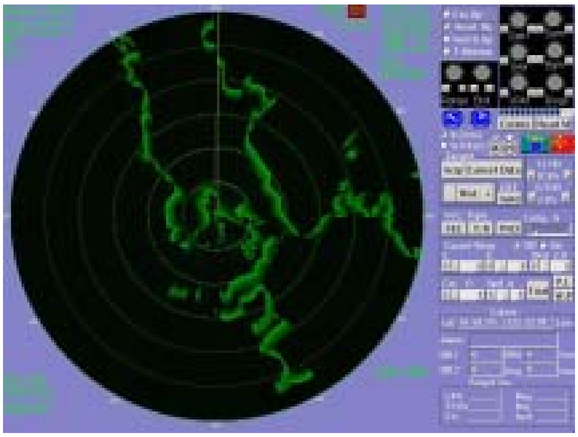


Figure 6. A simulating radar echo image of coastline

3. CONCLUSION

According to the front algorithm, make adaptive partition of the points when add them to the screen, prosecute it recursively, until all the lengths of the generated sublines are less than the given threshold value. The detail of the radar coastline generated by this way will not be affected by the screen resolution, displaying proportion and sampling resolution. Rotating vector algorithm keeps the shape and trend of the original sampling points very well, can generate complicated and irregular coastline with details, and the computing rule of it is simple. The reliable result is *available* (see Figure 6).

4. REFERENCES

[1] Philippe Husni. Visual Simulation White Paper, 1990.

[2] Barn sly, M.F. Fractal functions and interpolation [J]. Constr, Approx. 1986(2): 303-329.

[3] Dan Liu,Practical Fractal Graphics,Dalian Maritime University Press, Dalian, China,2001,176-181.

Dr. Dan Liu is an Associate Professor in Institute of Nautical Science and Technology, Dalian Maritime University. She graduated from Dalian University of Technology in 1998. She is a holder of Fok Ying Tung Education Foundation (1998); was a visiting scholar of University of Technology, Sydney (2003). She is one of the co-chairs of VIP international conference. She has published three books, over 30 Journal papers. Her research interests are in digital image processing, Fractals, computational geometry and virtual reality.

An Iterated Algorithm for Implicit Surfaces Rendering

Ni Tongguang, Gu Yaolin
The School of Information Technology, Southern Youngtze University
Wuxi, 214036 P.R.China
Email: hbxtntg-12@163.com **Tel:** 0510-5886217

ABSTRACT

In computer graphics, implicit surfaces are easy to generate geometric form, but difficult to render. We present a point-based rendering algorithm for high-quality rendering of implicit surface. Its implementation is simple. It can render arbitrary implicit surfaces, but the speed is slow. Towards the implicit surfaces, one variable can be expressed in term of the other two, the method can be used to render the surfaces efficiently. Furthermore, it can also be used to removal the hidden surface using a z-buffer and to create shadows using a shadow buffer.

Keywords: Implicit Surfaces, Iterated Function System, z-Buffer, Shadow buffer.

1. INTRODUCTION

There are two common ways for representing surfaces mathematically in R^3. One of the two ways is to use parameter equations, such as the form $x=x\ (u, v), y=y\ (u, v), z=z\ (u, v)$ with parameters u, v, where (x, y, z) is a point in R^3. The other is implicit surfaces that are represented by mathematical functions of the form $f\ (x, y, z)=0$, where for arbitrary functions $f\ (x, y, z)$, it is not possible to express x, y, and z in terms of parameters.

Implicit surfaces have many uses in mathematics and science because they are the isovalued surfaces of scalar fields. Rendering implicit surfaces is therefore useful in investigations which involve the visualization of scalar fields. Although Computer Aided Geometric Design uses predominantly parametric surfaces, implicit surfaces also have uses in this area. For example, the equations of implicit surfaces, which are of the general form $f\ (x, y, z)=0$, can be used to test the proximity of a given point to a surface by testing the magnitude and sign of $f\ (x, y, z)$. Parametric representations cannot be used for this task. Implicit surfaces are also used for blending and offset surfaces.

Implicit surfaces find widely use in engineering and scientific problem solving in the real world. It can also be used in CAD modelling, both for surface definition and use as blending surfaces.

2. IMPLICIT SURFACE RENDERING ALGORITHM

Basically, the main techniques using for rendering parametric and implicit surface are polygonization, scan line, ray tracing, and point based systems.

Regular implicit surface can also be polygonized for rendering, but the process is different form parametric surfaces. For example, Suffern’s algorithms are based on octree based recursive space subdivision techniques within a viewing cube specified by the user [1]. But some implicit surfaces, such as self-intersecting implicit surfaces are difficult to polygonize [2]. Scan line technique can render implicit surfaces, but its implementation is complex. Ray traced images produce some of the most realistic images of model scenes. Ray tracing allows a wide variety of surfaces that are not suitable for rendering using polygons to be drawn. An example of such a surface is Steiners’s surface [3] which self intersects. Missing polygons form along the line of self-intersection using polygonal methods as the curvature cannot be accurately estimated. but the ray-surface intersection equation has to be solved for each surface rendered.
Along with the abroad use of light mask display, point based rendering technique for implicit surfaces develop fast. Point based rendering techniques use points, or objects such as discs, as rendering primitive.

In recent years, De Figueiredo and Gomes [4] used physical-based particle that obey Newtonian equations of motion to render differentiable implicit surfaces. They used points for rendering, but did not produce shaded images. Tanaka et al. [5] used particles that obey stochastic differential equations to render twice differentiable implicit surfaces. The points are evenly distributed and the surfaces intersections can be rendered are the nice feature of their works, however, the method is complex. The method can produced shaded image with discs.

Our method for rendering implicit surfaces is a point-based algorithm that is based on iterated function systems [6,7]. Compare to the above point-based method, the main advantage of our method is its simplicity. As discussed in the following section, the technique is very simple to program. So it can be used as a rapid visualization tool for surfaces.

For those implicit surfaces where one of the variables can be expressed in terms of the other two, it distributes points over the surfaces. For these surfaces it is faster than the techniques of de Figueiredo and Gomes [4] and Tanaka et al. [5] because points are distributed directly onto the surfaces, instead of being the result of the iterative solutions of differential equations. Only the implicit function $f(x, y, z)$ and its gradient are used to render the surfaces.

We create enough points to render smoothly shaded images of the surfaces with hidden surface removal using a z-buffer, and with shadows using a shadow buffer. The points themselves are used for the rendering. We do not guarantee that all pixels on the surfaces have been filled in, but instead stop the rendering interactively when the images look stop the rendering interactively when the images look satisfactory.

3. ITERATED FUNCTION SYSTEM

Iterated function systems (IFSs) were first developed in the 1980's. Consider a finite set of contraction mappings $w_m : X \rightarrow X$ on the complete metric space (X, d), then $X, w_m : m=1,2,\ldots,M$ is an iterated function system (IFS). It can generate image of fractal objects in 2D and 3D, and are used in image compression [8].

```
Points $V_0, V_1, \mathrm{K}\ V_n$ are given, staring point is $P_0 = V_0$ .
For (i=1;i<Max-i;i++)
{
        $V$ = random ($V_0, V_1, \mathrm{K}\ V_n$);
        //select a random point for $V_0, V_1, \mathrm{K}\ V_n$
        $P = (V + P_{i-1})/2$ ; //generate a point
        pixel $P_i$ ; // render the $P$
}
```

Algorithm 1 A IFS to Generate Points in a Unite Square

General IFSs use a number of contractive transformation functions. In algorithm 1, the functions are scaled by the factor half, which are found by calculating a point mid-way between a current point and one of fixed points that define the system. When *n=3*, because V_1, V_2 and V_3 are not collinear, points P_1, P_2, …finally fill a fractal Sierpinski triangle (Figure 1, left). When *n=4*, V_1, V_2, V_3 and V_4 are the vertices of a square, which can be filled by a random scatter of points (Figure 1, right)[9].

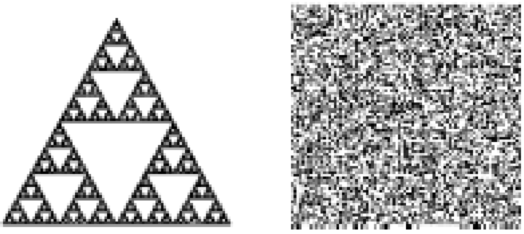


Figure 1 Results of Algorithm 1 Applied to Vertices of a Triangle and a Square

Because only one integer random number and divisions by 2 are used in finding each point, the above algorithm is an efficient way of obtaining a random scatter of points in a 2D space. Directly applying a random number generator to find *(u, v)* coordinates uses two real valued random numbers. For example, for a unit square in [0, 1] whose vertices V_0, V_1, V_2 and V_3, run counter-clockwise from V_0 to V_3, the four selection functions can be represented as given in Jones [10] as follow:

$$\begin{aligned} V_0 : u_i &= u_{i-1}/2, & v_i &= v_{i-1}/2 \\ V_1 : u_i &= (1+u_{i-1})/2, & v_i &= v_{i-1}/2 \\ V_2 : u_i &= (1+u_{i-1})/2, & v_i &= (1+v_{i-1})/2 \\ V_0 : u_i &= u_{i-1}/2, & v_i &= (1+v_{i-1})/2 \end{aligned} \tag{1}$$

Initially with $\begin{cases} u_0 = 0.5 \\ v_0 = 0.5 \end{cases}$

A random integer in [0…3] is used to choose the value of V_0 to V_3 that is then used to calculate the next u_i, v_i plot point, P_i, as given in the above Algorithm. Surface normals for rendering are evaluated functionally from surface formulae. If the algorithm loops enough times, the scatter points generated can render a smooth, beautiful surface.

With the algorithm loop allowed to run long enough, exact depictions to pixel precision of the image are generated on the display. Using Phong shading, a *z*-buffer or a shadow-buffer when rendering, is particularly easy, because at a time, only one point is processed, which reduces many of the problems of computer graphics to simple decisions. The process is slower than standard scan line methods, but exact depictions of curved surfaces are generated.

4. REDERING IMPLICIT SURFACE

In common, implicit surfaces can be rendered by a brute force algorithm, but it consumes much time. We use the algorithm 1 with n=8. This is used to evaluate scatted points inside a box containing the surface. If a point lies inside or on the implicit surface, it is rendered and shaded using the gradient function *▽f(x, y, z)*, otherwise it is not rendered.

Algorithm 1 can render arbitrary implicit surfaces, but it consume much time, because is basically a brute force technique that randomly distributes points in space instead of over the surfaces. In order to reduce the time of rendering implicit surfaces, we need distribute point over the surfaces rather than in the space. We can do this for implicit surfaces that allow us to express one of the variables in terms of the other two, for example, towards the variable *z*, we can reduce *f (x, y, z)=0*, to one or more equations of form *z=g (x, y)*. So the surface has one or two parameters. Therefore we can use (1) to generate *(x, y)* pairs for substitution of *g (x, y)* now. Surfaces are generated in this way by 'lofting' *z* with scattered points in a *(x, y)* region.

The Cross-cap surface (2)(figure 2) is a type of the Steiner surface [6]. The Steiner surface is a quartic nonorientable surface. The Cross-cap surface is one of the three possible surfaces obtained by sewing a Moebius strip to the edge of a disk. The other two are the Boy surface and Steiner's Roman (3) (figure 3), all of which are homeomorphic to the real projective plane (Pinkall 1986).[6]. It has a segment of double points which terminates at two pinch points known as Whitney singularities. A cross-handle is homeomorphic to two cross-caps (Francis and Weeks 1999)[6].

$$f(x,y,z) = 4x^2(x^2+y^2+z^2+z)+y^2(y^2+z^2-1)=0 \tag{2}$$

$$x^2y^2 + y^2z^2 + x^2z^2 + xyz = 0 \tag{3}$$

By arranging (2) as a quadratic equation in z ,

$$(4x^2+y^2)z^2+4x^2z+4x^2y^2+4x^4+y^4-y^2=0 \tag{4}$$

$$z = \frac{-2x^2 \pm \sqrt{(y^2+2x^2)(1-4x^2-y^2)}}{4x^2+y^2} \tag{5}$$

gives two solutions for *z*. When $(4x^2 + y^2) = 0$, the equation has only one solution,

$$z = \frac{y^2 - y^4}{4x^2} - x^2 - y^2 \tag{6}$$

found from (4). Otherwise, it yields two surface points *(x, y, z)* by selecting the + or - sign from (5). When the discriminant from (5), $(y^2 + 2x^2)(1\text{-}4x^2\text{-}y^2)$, is negative, there is no real solution for *z*; there are no surface points over these values of *(x, y)*. If the discriminant is zero, two coincident solutions are found, but this does not need special treatment. Figure 2 renders Cross-cap surface by ' lofting in *z* '. Error traps for the linear equation (4) and for values of *(x, y)* for which no surface exists enable all surface points to be rendered. Normals for shading are found by converting $\nabla f(x,y,z) = (\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \quad \frac{\partial f}{\partial z})$, the gradient of the implicit surface function (2)[6], to unit length. The symmetry of (2) allows efficiency savings, since if *(x, y, z)* is a solution, then (cycling from *z* to y to z and on to x again), *(y, z, x)* and *(z, x, y)* are also solutions. For every point from (4) or (5), two other points can be found by cycling the coefficients of the point and normal.

Generating surfaces with 'point as primitive' allows relatively easy mapping of surface or coordinate based characteristics onto the surface. For example, Figure 4 shows the same Steiner's Roman surface with points color coded to represent the magnitude of the gradient of *z*.

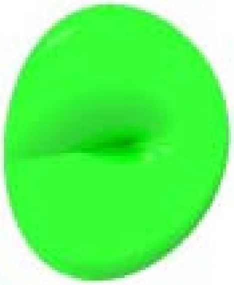


Figure 2 A Cross-cap Surface Using Point Scatting

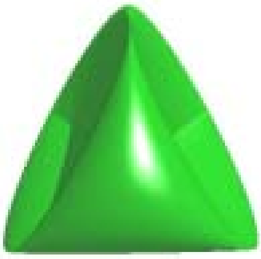


Figure 3 A Steiner's Roman Surface Using Point Scatting

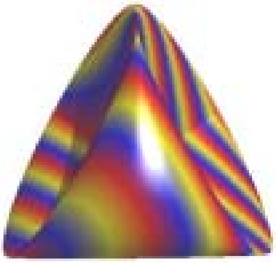


Figure 4 A Steiner's Roman Surface Using Point Scattering with Coded Color

5. CONCLUSIONS AND FUTURE WORK

The point-based algorithm renders implicit surfaces to the pixel precision successfully. Our method is simple, distributing points on surfaces and rendering them very fast. The technique runs automatically with no user intervention.

It only takes a few seconds on a 733Mhz machine after the points are distributed over the surfaces. For visualization purposes, it does not matter if a few surface pixels are missing. Phong shaded images with *z*- and shadow buffers take from 1 to 3 minutes to render on the machine using un-optimised code. Surface shapes become visually clear within a few seconds, the extra time is needed to ensure full surface cover. The z-buffer is 800 pixels square, the shadow buffer 1500 pixels square; surfaces shown are 500 to 600 pixels in horizontal extent.

While most of the implicit surfaces used in this paper have used the ' lofting in *z*' method, which essentially changes the problem to a parameterized system, we render implicit surfaces by generating points in a rectangular volume using Algorithm 1 with n=8. This does not require casting the problem into a parametric form.

A number of useful techniques, such as adding surface color, can be incorporated into our method. We also do not do any antialiasing, so rendering the images at higher resolution is not perfect. In order to render smooth surfaces, our method use a large number of points, so the rendering time is great. We intent to do further work in this area, using interval methods and octree subdivision, to produce smaller rectangular regions that better fit the surface.

6. REFERENCES

[1]BALSYS, R.J., and SUFFERN, K.G. 2001. Visualization of implicit surfaces, *Computers and Graphics*, 25:89－107.

[2]SCHMIDT, M.F.W. 1990. Cutting cubes – visualizing implicit surfaces by adaptive polygonization. The Visual *Computer*, 10:101－115.

[3]WEISSTEIN,E.W.2001.www.mathworld.com,CRC Press, LLC.

[4] De Figueiredo L. H. and Gomes J. 1996. Sampling implicit surfaces with physically-based particle systems, *Computers and Graphics*, 20:3:365－375.

[5]TANAKA, S., SHIBATA, A., YAMAMOTO, H., KOTSURU,H. 2001. Generalized Stochastic Sampling Method for Visualization and Investigating of Implicit Surfaces, *Proc. Eurographics 2001, Computer Graphics Forum*,20:3:359－367.

[6]JONES, H. and MOAR, M. 2000. Rendering through iterated function systems. *In Paradigms of complexity:fractals and structures in the sciences*. M. Novak (ed.).World Scienti.c, Singapore, pp. 167－177.

[7]JONES, H. 2001. Exact object rendering through iteratedfunction systems. *Eurographics UK Chapter AnnualConference, University College, London*. pp. 27－36.

[8]Hart, J. C. Fractal image compression and the recurrent iterated function systems. *IEEE Computer Graphics and Applications*, 1996.

[9] JONES, H. 1990. Dϋrer, gaskets and Barnsley ' s chaos game. *Computer Graphics Forum*, 9:327－332.

[10]JONES, H. 2001. Exact object rendering through iterated function systems. *Eurographics UK Chapter Annual Conference, University College, London*. pp. 27－36.

Ni Tongguang is a postgraduate, the School of Information Technology, Southern Youngtze University. He graduated from Wuxi University of light industry in 2000. His research interests are in visual reality and compute graphics.

An Implementation of Quarter Pixel Block Motion Estimation Using SIMD*

Xinchen Zhang Ruimin Hu Deren Li Zhongyuan Wang
The Key Laboratory of Multimedia and Network Communications Engineering,
Hubei Province, Wuhan University
Wuhan, Hubei 430072, China
Email: xc.zhang@263.net **Tel.:** +86-27- 62360104(THS Mobile), 87886634 (lab)

ABSTRACT

For improving video coding efficiency, sub-pixel motion estimation (ME) is used extensively in the existing video coding standards. The quarter pixel ME is one of the high complexity tools in H.264/AVC. In this paper, a parallel quarter block motion estimation algorithm that not only accelerates the process of sub-pixel motion estimation but also maintains accuracy as that of the original algorithm is proposed. In Intel P4 CPU, the SIMD (single instruction multiple data) technique is commonly used to provide an execution speedup. The implementation of this algorithm using parallel processing on P4 platform is discussed. The proposed algorithm satisfies in particular the requirements of low-rate real-timed video communication. Experimental results show that the optimized video encoder is more than 13.5 times faster than the original reference software while keeping the accuracy of the latter approximately.

Keywords: Sub-pixel Block Motion Estimation, Computation Complexity, Parallel Algorithm, SIMD, H.264/AVC

1. INTRODUCTION

The development of video communication is strongly supported by the advances in the research on video compression algorithms and release of video coding standards, especially H.264 [1]-[2] which can provide equivalent objective quality at a data rate about 50% lower than that required by H.263 [3] (Baseline) and MPEG-4 [4] (SP). It is the current video standardization project of JVT (formed by ITU-T VCEG and ISO/IEC MPEG). The primary goals of H.264 are improved coding efficiency and improved network adaptation, so it must be made an attractive candidate for all video communication applications in future.

However, the algorithm complexity of H.264 dramatically increases due to the adoption of a number of video coding tools. The variable block motion estimation is the greatest calculation in H.264 coding process. Many fast block motion estimation algorithms have been proposed in the literature, such as the three-step search algorithm [5], the four-step search algorithm [6], the diamond search algorithm [7] and other novel algorithms. Those algorithms improve the speed of integer-pixel motion search processing, but the complexity of sub-pixel motion search also holds H.264 back to be applied in real-time video communication. SIMD technique provides a possible solution to accelerate execution. Using SIMD and the fast ME algorithm we had accelerated the performance of the integer-pixel ME process, integer translation and intra perdition process in H.264. Because the video data in sub-pixel block motion search process is inconsecutive, the present algorithm is unsuitable to utilize SIMD code. In this paper, a fast quarter block motion estimation algorithm is proposed and implemented.

This paper is structured as follows. Section Ⅱ introduces SIMD technique and the quarter block motion estimation algorithm of H.264. Section Ⅲ proposes a fast sub-pixel block motion estimation algorithm and gives an implementation using SIMD and parallel processing on P4. Section Ⅳ provides experimental results to be analyzed. Section Ⅴ makes a conclusion.

2. SIMD AND SUB-PIXEL ME

Single Instruction Multiple Data

The SIMD is a widely used technology for parallel computation. A SIMD operation is n-parallel b/n-bit sub-operations executed by a modified b-bit functional unit. An instruction corresponding to a SIMD unit executing SIMD operations is called a SIMD functional unit. With the SIMD technology, the processor is able to handle multiple data in a single instruction.

On P4 platform, there are eight MMX registers used to perform operations on 64-bit packed integer data and eight XMM registers on 128-bit packed integer or double-precision floating-point data [8]. For example, a pixel is represented by an 8-bit data. A 128-bit SIMD integer instruction can operate on 16*8-bit packed integer operands located in the XMM registers.

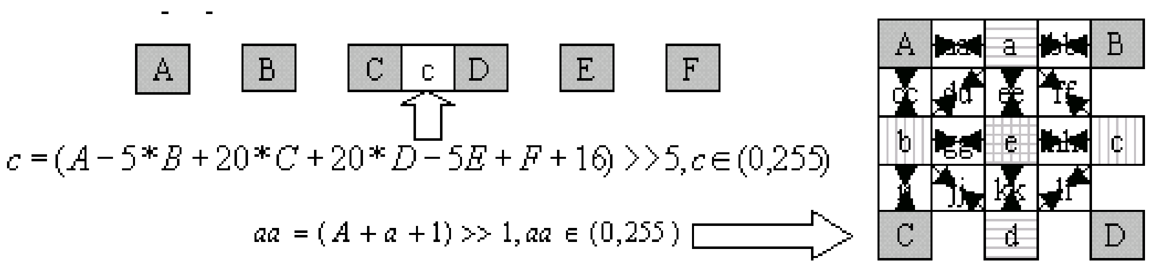
Quarter pixel block ME of H.264/AVC

Sub-pixel motion compensation can provide significantly better compression performance than integer-pixel compensation, at the expense of increased complexity. Quarter-pixel accuracy outperforms half-pixel accuracy. In the quarter pixel motion estimation/compensation process of h.264 video standard, there are two steps that are the interpolated sub-pixel samples and sub-pixel block motion search.

The simplified illustration about the interpolated sub-pixel samples in the reference frame is provided in Fig. 1. The luma prediction values at half sample positions shall be derived by applying a 6-tap filter with tap values (1, -5, 20, 20, -5, 1) to the nearest integer position samples in the horizontal/vertical direction. The luma prediction values at quarter sample positions shall be derived by averaging with upward rounding of the two nearest samples at integer and half sample positions.

In H.264, the motion search uses the block match criterion with bit cost to select candidate motion vector, according to Fig.2.

*Support by the National Natural Science Foundation of China (60272097).



$$c = (A - 5*B + 20*C + 20*D - 5E + F + 16) >> 5, c \in (0,255)$$

$$aa = (A + a + 1) >> 1, aa \in (0,255)$$

Fig. 1 Interpolation of luma half-pixel and quarter-pixel positions

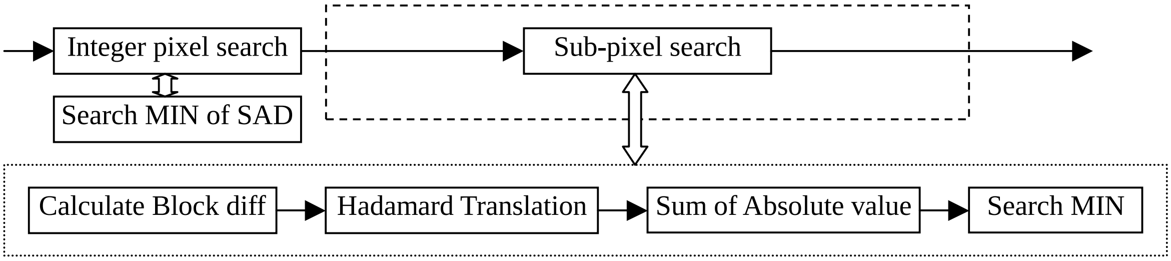
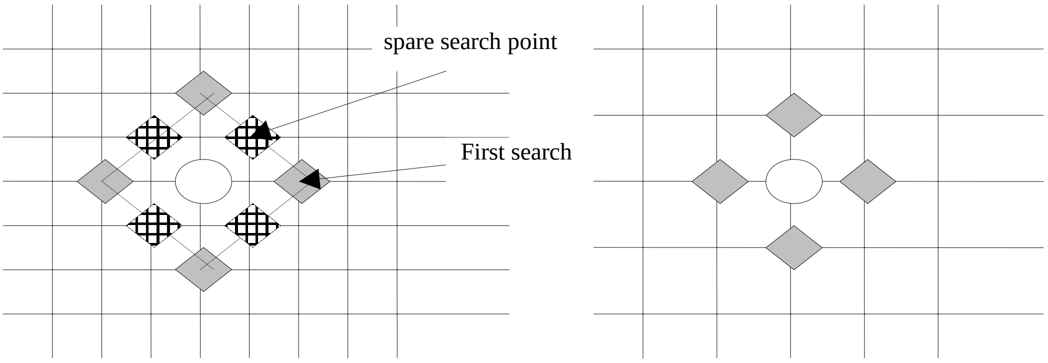


Fig. 2 The block motion search in H.264



A. The variable shape search pattern

B. The small diamond-shaped search pattern

Fig. 3 Two Search patterns

After implemented the integer translation, intra prediction and the fast integer-pixel ME algorithm [11] by SIMD codes, we measure that the interpolated samples process (61.6%) and sub-pixel motion searching (29.1%) spend over 90% run-time of all. In order to speed up this process, the appropriate algorithm should be able to decrease the amount of operations and implement it easily with SIMD.

3. PROPOSED FAST ALGORITHM AND IMPLEMENT

The proposed algorithm of sub-pixel block motion search

The complexity of the sub-pixel motion search is also from the more search points. The number of quarter pixels around integer samples is 48. This means it needs 49 times to calculate the process of block matching. Many fast integer pixel block motion estimation algorithms [9-11] attempt to locate optional motion vectors step by step by evaluating as few points as possible based on a gradient scheme. To develop the DS algorithm with an adaptive threshold and use it to sub-pixel block search, the variable diamond-shaped search pattern and the small diamond-shaped search pattern are employed, as Fig. 3. We suppose that the best matching point is around the best integer sample and the distance of them is closer than the distance of the best matching point to the other integer sample. We can simply define the distance of two points as $|x_1-x_2| + |y_1-y_2|$. So the candidate points we searched are those points whose distance to the best integer sample less than 4. The value 4 is equal to the distance of the closest integer sample to the best integer sample.

The proposed method consists of tree steps:

1. Starting form the suited integer pixel, use the coarse diamond painted with gray background in the picture A of Fig.3 to compute the block distortion measure (BDM).

2. If the minimum block distortion does not occur at the center point and the min-cost is less than the threshold (min_cost is BDM of the best-matching integer pixel), go to step 3; Else, Compute the residuary four points painted with tartan background.

3. Use the small diamond-shaped pattern to compute BDM around, select the least BDM point as the best-matching point.

As the proposed search scheme, the possible searching points are all close around the best integer sample and the distance is also less than 4. It is remarkable that the number of the search point is form 8 to 12. If the threshold is small, the number is high to 12; if large, the number is small to 8.

Use SIMD to implement the algorithm

The goal is to implement the algorithm efficiently using an SIMD technique to further accelerate the process of sub-pixel motion estimation.

Usually, the luma of a video frame is saved in the memory in a row-by-row raster scan manner. We can load 16 integer-pixel samples successively to a XMM register and perform addition, subtraction, multiplication and multiply/add operations on 128-bit packed byte integers by one arithmetic instruction. To generate interpolated sub-pixel samples, partition data as 16 points in length and store data by row and by column. Storage by row implies that parallel access and computation is by column, vice versa.

But, storage data in row-by-row raster scan manner after luma sample interpolation process is not suitable for sub-pixel block motion search. In order to calculate the BDM between the interpolated reference block and the current block by SIMD more efficiently, we transform the position of quarter pixel to 4*4*Width*Height matrix, expressed as Eq. (1) . Then we can fetch pixels successively at a time to calculation the BDM, shown in Fig.4.

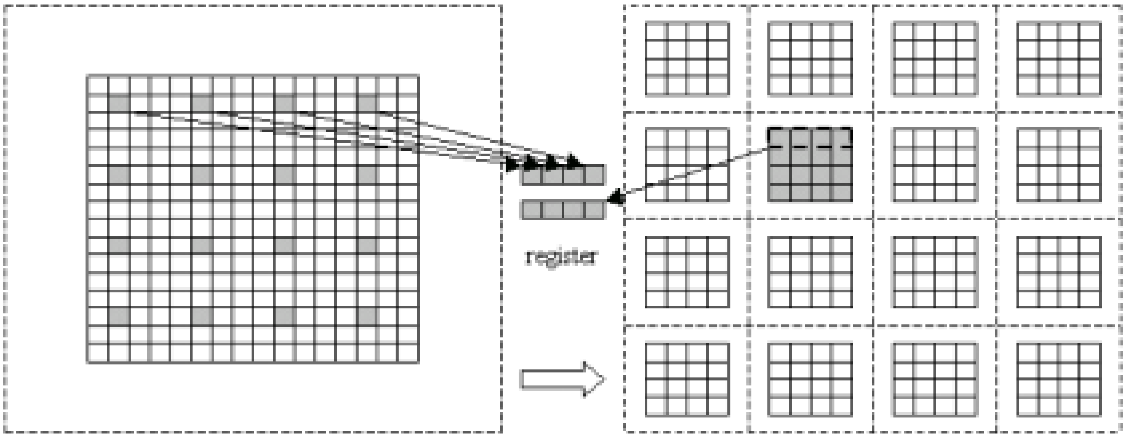


Fig. 4. The Transform of Data Storage

$$\begin{aligned} &B_{nrq}[p_y][p_x][y1][x1] \\ &= B_{nrq}[y\%4][x\%4][y/4][x/4] = B_{rq}[x][y] \end{aligned} \tag{1}$$

Now, the mean absolute difference for candidate vector at quarter pixel is defined as

$$\begin{aligned} MAD_q(V_x, V_y) &= \sum_{m=0}^{B_h}\sum_{n=0}^{B_w} r_{v(x,y)q}(m,n) \\ &= \sum_{m=0}^{B_h}\sum_{n=0}^{B_w} \left| B_{nrq}(p_{v(x,y)}, m, n) - B_c(m,n) \right| \end{aligned} \tag{2}$$

If we regard the block as a matrix and save it in a contiguous memory space, the subtractions in Eq. (2) can be carried out simultaneously using an SIMD instruction. This storage policy has an additional advantage for saving the integer pixel memory space. When both of py, px equal zero, the transformed matrix is the integer sample of the reference frame. To perform the three steps of the proposed algorithm, the search process based on order is declared as follows:

The value x and y is the position of the optimal block matching integer pixel.

Step 1: calculate the MAD at
(0, 2, y, x-1), (2, 0, y-1, x), (0, 2, y, x), (2, 0, y, x)

Step 2: judge condition; and if true, then search other four points at
(3, 3, y-1, x-1), (3, 1, y-1, x), (1, 1, y, x), (1, 3, y, x-1)

Step 3: search the left, up, right, down point around the center point.

To reduce memory R/W time and avoid cache misses, the advices are mentioned as the following:

1) Single long type data replaces many single byte data, and the memory address must be aligned to a 16-bit boundary or 8-bit boundary.
2) Increase memory/video fills bandwidth using 64/128 -bit move instruction.

In addition, because of variable-size block used in block ME, the code expense of the excess motion vector must be considered in block matching searching process. So when encoding a macro-block, we choose the best mode by searching the minimum sum of the sub-block MAD and the cost of DMV.

4. EXPERIMENTAL RESULTS

For our simulation, the first 100 frames of three sequences (NEWS, GARDEN, TEMPETE) with different types of motion were used. The size of the frame is 352×288 pixels (CIF format) quantised to 8 bits. The experimental results are carried out on P4 computer using the reference software JM6.1 of H.264. The 100 frames have been considered as that only first frame is I frame and others is P frame. The optimized encoder had been performed the integer-pixel ME algorithm in [11], the proposed quarter-pixel block ME algorithm, the intra-prediction process and the integer transform using SIMD. The encoder use 16×16, 16×8, 8×16 and 8×8 block sizes, only one reference frame used in the inter-prediction process and not use rate-distortion option.

Tab. 1 Comparison of performance for different video sequences

“NEWS”			
	PSNR (dB)	Bit Rate (kbit/s)	Run time (sec/frame)
Original	32.00	123.01	1.192
Optimized	31.86	130.38	0.092
Difference or Speedup ratio	-0.14	+6.00%	12.96
“TEMPETE”			
	PSNR (dB)	Bit Rate (kbit/s)	Sec/Frm
Original	28.34	380.07	2.144
Optimized	28.22	402.49	0.154
Difference or Speedup ratio	-0.12	+5.90%	13.92
“GARDEN”			
	PSNR (dB)	Bit Rate (kbit/s)	Sec/Frm
Original	27.67	497.60	1.699
Optimized	27.62	531.51	0.124
Difference or Speedup ratio	-0.05	+6.81%	13.70
Average			
	PSNR (dB)	Bit Rate (kbit/s)	Speedup ratio
Difference or Speedup ratio	-0.10	+6.24%	13.53

In Tab 1 we can see the performance in terms of peak signal-to-noise ratio (PSNR dB), bit rate (kbit/s), the increase ratio of bit rate to the original (Add-bit %), the average run time of encode one frame (Run time sec/frame) of the optimized encoder and speedup ratio of coding time.

According to the results, the optimized encoder is more than 13.5 times faster than the original reference software while keeping the accuracy approximately. We measure the average drop is 0.1dB and about the bit rate increases about 6.2%. The drop of PSNR and the increase of bit are come from both the fast integer sample motion estimation algorithm and the quarter pixel motion estimation algorithm. The optimized H.264 encoder has the capability of coding approximately 10 frames per one second. Although adding a few bits, the encoder basically satisfies the requirement of real-time video communication, especially low-bit application.

5. CONCLUSIONS

In this paper, the proposed sub-pixel block motion estimation algorithm improves the search speed while keeping the same accuracy. Using SIMD, we implement the algorithm in order to process parallel video data. Experimental results show that it performs the H.264 encoder faster efficiently to satisfy the request of the real-time video application.

6. REFERENCES

[1] “Final committee draft: Editor’s proposed revisions,” in *Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG*, T.wiegand, Ed., Feb.2003, JVT-F100.

[2] T.Wiegand, G.J.Biontegaard, and A.Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 560-576, July 2003.

[3] “Video Coding for Low Bitrate Communication, Version 1,” ITU-T, ITU-T Recommendation H.263, 1995.

[4] “Coding of Audio-Visual Objects-Part 2: Visual,” ISO/IEC JTC1, ISO/IEC 14496-2 (MPEG-4 visual version 1), 1999.

[5] T.Koya, K.Linuma, A.Hirano, and so on, “Motion-compensated inter-frame coding for video conferenceing,” in *Proc. NTC81*, new Orleans, LA, pp.961-965, Nov. 1981.

[6] L.M.Po and W.C.Ma, “A novel four-step algorithm for fast block motion estimation,” *IEEE Trans. Circuits Syst. Video Technol.*, vol.6, pp. 313-317, June 1996.

[7] S.Zhu, KK, Ma, “A new diamond search algorithm for fast block motion estimation,” *Proc. Of Int. Conf.* Information, Communicaiton and Signal Processing, vol. 1, pp. 292-296, 1997.

[8] “IA-32 Intel Architecture Software Developer’s Manual Volume 1:Basic Architecture,” Intel Corporation, Order Number 245470-012, 2001

[9] Shan Zhu, Kai-Kuang Ma, “A new diamond search algorithm for fast block-matching motion estimation,” IEEE Transactions on Image Processing, vol. 9, no. 2, pp. 287-290, Feb. 2000.

[10] Jo Yew Tham, Ranganath, Maitreya Ranganath and Surendra Ali Kassim, “A novel unrestricted center-biased diamond search algorithm for block motion estimation,” IEEE Transactions on Circuits & Systems for Video Technology, vol. 8, no. 4, pp. 369-377, Aug. 1998.

[11] A.M.Tourapis, O.C.Au and M.L.Liou, “Fast block-matching Motion Estimation using Predictive Motion Vector Field Adaptive Search Technique,” in ISO/IEC JTC1/SC29/WG11 MPEG200/M5866,

Noordwikerhout, NL, Mar. 2000.

Zhang Xinchen is a Ph.D. candidate in the key lab of multimedia network communication engineering, Wuhan University. He received the Master Degree and Bachelor Degree in circuit and System from Wuhan University s in 2001 and 1999. His research interests are in multimedia network communication and video processing.

Hu Ruimin is a Full Professor, doctoral supervisor and vice-president in the Institute of Computer Science and Technology, 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. In the end of 1994, he joined the Postdoctoral Station for Information Engineering in Surveying, Mapping and Remote Sensing, his adviser was Dr. Li Deren, an academician of Chinese Academy of Sciences, and Chinese Academy of Engineering. 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, etc.

Rendering CG Objects into Photographs with Light Probe Images

Wang Jun, Gu Yaolin
The School of Information Technology, Southern Yangtze University, Wuxi, 214036, China
Email: julyhorse@etang.com **Tel.:** 0510-5864233

ABSTRACT

According to one of the existing shortcomings of the conventional image-based technique, we propose a new conceptual model of image-based graphics in this paper. Then a general method of its implementation is introduced. Meanwhile, we present a rendering algorithm based on ray-tracing technique. Our method has the application in the fields of virtual reality, architectural visualization, computer aided design and so on.

Keywords: Light Probe Image, Image-Based Lighting, Virtual Reality, Image-Based Rendering and Modeling, Ray-Tracing

1. INTRODUCTION

Rendering computer-generated (CG) objects into the photographs of the real world has been an essential part of virtual reality. For a successful integration it is crucial to render the computer-generated objects into photographs seamlessly. In general this work is extremely hard and requires recovery of real world information, such as camera positions and parameters, and shapes, material properties, and motion of real objects.

Lighting simulation is one of the key issues often being involved. Current techniques of lighting simulation are labor intensive and not always successful. A common technique is to manually survey the positions of the light sources in real world, and to instantiate a virtual light of equal color and intensity for each real light to illuminate the CG objects. Unfortunately obtaining the accurate parameters of the light sources is not always successful. Another technique is to photograph a reference object in the scene where the new object is to be rendered, and use its appearance as a qualitative guide in manually configuring the lighting environment. Lastly, the technique of reflection mapping is useful for mirror-like reflections. These methods require considerable hand-refinement and none of them could easily simulate the effects of indirect illumination from the environment.

Debevec recovered the illumination information in real world by a collection of photographs taken with different exposures and stored the true illumination in light probe image. In this way, he constructed the scene with a light-based model.[1,2] This is the premise of our work.

2. BACKGROUND

The current image-based approach, which includes image-based modeling and rendering, differs from the traditional geometry-based graphics in that both the geometry and appearance of the scene are derived from real photographs. The techniques often allow for shorter modeling times, faster rendering speeds, and unprecedented levels of photorealism. However, it ignores to recover the illumination information from the photographs and is not successful in lighting simulation.

Image-based lighting (IBL) is the process of illuminating the scenes and objects (real or synthetic) with images of light from the real world. In this paper, we introduce it as a new approach to illuminate the scene.

Today, we could construct the virtual environment in three ways: geometry-based model, image-based model and light based model. All of them could be recovered from the original photographs. Here we use the term light-based model to refer to a representation of a scene that consists of radiance information, possibly with specific reference to light leaving surfaces, but not necessarily containing material property information. [2]

3. A CONCEPTUAL MODEL OF IMAGE-BASED GRAPHICS

In this paper we propose a conceptual model of image-based graphics (Fig. 1).

According to our conceptual model, both the appearance of the scene and the illumination information can be obtained from the photographs. The appearance of the scene can be represented by geometry model, image-based model, light-based model or the arbitrary combination of them. You could recover the geometry-model based on computer vision approach, or create the image-based model by image mosaic [5,6,7]. The light-based model can also be used to represent the 3D scene. Meanwhile, the light-based model records the true illumination, so we could illuminate the CG objects with it. At last we could get the photorealistic scene under the proper illumination with rendering algorithms.

The implementation of the above conceptual model includes the following steps:
(1) Capture the real illumination of the scene and record in light probe image.
(2) Construct the scene with light-based model, geometry model and image-based model.
(3) Render the CG objects into the scene.

4. IMPLEMENTATION ISSUES

4.1 Make the Light Probe Image

Light probe images are photographically acquired images of the real world and have two important properties [11]. Firstly, they're omnidirectional. It means that for each direction in the real world, and there's a pixel corresponding to that direction. Secondly, their pixel values are linearly proportional to the amount of light.

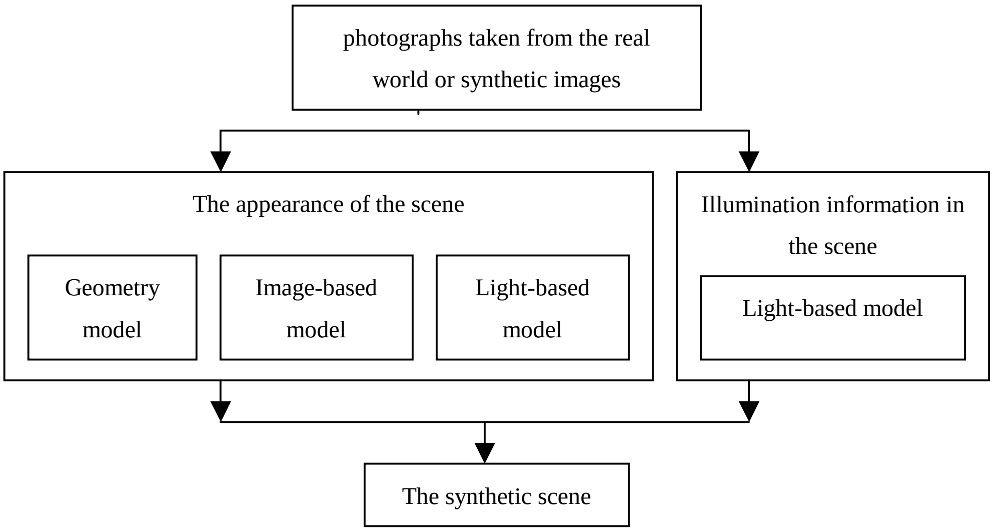

photographs taken from the real world or synthetic images
The appearance of the scene
Geometry model
Image-based model
Light-based model
Illumination information in the scene
Light-based model
The synthetic scene

Figure 1 A new model of the image based graphics

There are several ways to make light probe images. The simplest one is to take a collection of photographs of a polished steel ball reflecting the surrounding environment with the different shutter speed and the fixed viewpoint [2]. But this method will make the background of our result obscured because it samples the scene nearby and back poorly and introduces something unexpected. So it is necessary to edit the photographs in image editing programs manually.

Here we used a different method to acquire the whole scene. Using unchanged shutter speed, we took six photographs of surroundings in six directions which is parallel to the coordinate axes of the space coordinate frame, i.e. forward, backward, right, left, up and down. The six photographs constitute the cubic environment maps and contain the whole illumination information of surroundings. If we consider the light probe image to be normalized to have coordinates u=[0,1], v=[0,1], we have:

$$\begin{cases} \theta = \arctan\dfrac{v-0.5}{u-0.5} \\ \varphi = \pi\sqrt{(u-0.5)^2+(v-0.5)^2} \end{cases} \tag{1}$$

The unit vector pointing to the corresponding direction could be obtained by rotating the vector (0,0,-1) by φ degrees around the y (up) axis and then θ degrees around the –z (forward) axis. Note that the mapping used here is different from the conventional spherical environment mapping [10] in that the radical dimension is mapped linearly with angle and it avoids the problems of poor sampling around the edges.

Then we could compute the components of the corresponding unit vector as follows:

$$\begin{cases} x = -\sin\varphi\cos\theta \\ y = \sin\varphi\sin\theta \\ z = -\cos\varphi \end{cases} \tag{2}$$

The component with largest absolute value determines the cubic face and the other two components are the index of the texture coordinates.

When we make light probe images in this way, the response curves [1] of our digital camera and the shutter speed as we took photographs are also should be known. Fig.2 shows the panorama image represented in cubic environment maps and Fig.3 shows the light probe image that will be used in our experiment.

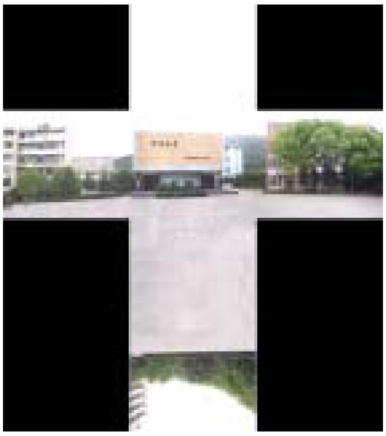


Figure 2 The Panorama Image Used in Our Experiment

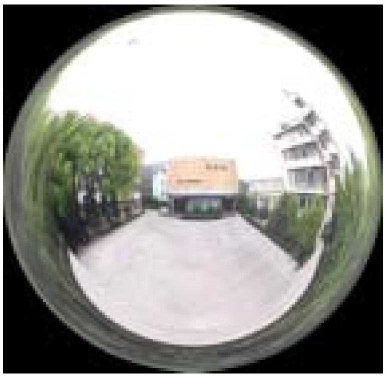


Figure 3 The Light Probe Image that Wwill be Used in Our Experiment

4.2 Scene Construction

The second stage of implementation is to reconstruct the scene by recovering the illumination information. The interplay of the light between the CG objects and their surroundings must be taken into account when the virtual scene is being rendered. However, a new object always has no significant effect on the appearance of the scene far from it. Here we divided the scene into three components: the distant scene, the local scene, and the CG objects.

The distant scene is assumed to be that radiates light toward the local scene and the CG objects, ignoring light reflected back to it. We created a light-based model of the distant scene with a light probe image and the technique of environment mapping. A light-based model is a representation of a scene that consists of the real radiance information. We can get it by mapping the light probe image onto a representation of the scene, usually a box or a sphere. Here we mapped the light probe image to a sphere. The sphere not only stores the illumination information from all the directions but also can be used as the background image when our rendering algorithm is applied because it shows the true appearance of the scene when we took photographs of it.

The local scene includes the surfaces which will interact with the CG objects, such as those which can catch shadows and receive reflected light from the new object. It is modeled as geometry because it needs to fully participate in the illumination solution.

Fig.4 shows our scene, which includes the table and CG objects, five balls.

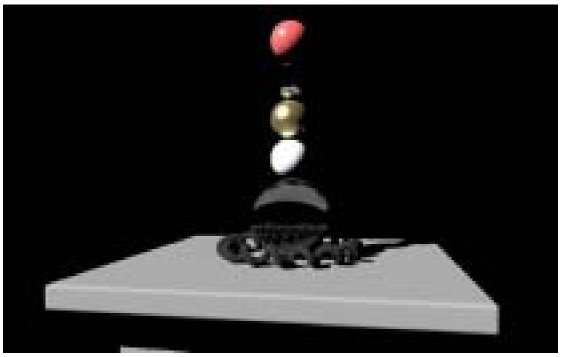


Figure 4 The Geometry Models of the Local Scene and the CG Objects

5. RENDERING ALGORITHM

We've developed a rendering algorithm in which for each pixel we trace a sample ray from the viewpoint to the scene and calculate the pixel value using the ray-tracing algorithm. When the sample ray hits the background, the algorithm looks up in the light probe image and takes the corresponding color. Otherwise, three components contribute to the pixel value: the diffuse reflection component, the specular reflection component and the transmission component.

5.1 Stochastic Sampling Strategy

Here we only mention the key issues of computing diffuse reflection component. At each surface point, the diffuse reflection component can be computed as the following integral of luminance over the hemisphere:

$$I_{diff} = \rho_{diff} \int_0^{2\pi}\int_0^{\frac{\pi}{2}} L(\theta,\phi)\cos\theta\sin\theta d\theta d\phi \qquad (3)$$

where θ is the polar angle and ϕ is the azimuthal angle. $L(\theta,\phi)$ is the luminance from direction (θ , ϕ). ρ_{diff} is the diffuse reflectance.

Usually the integral is approximated as follows:

$$I_{diff} = \rho_{diff} \frac{\pi}{2n^2} \sum_{j=1}^{n} \sum_{k=1}^{2n} L(\theta_j, \phi_k) \qquad (4)$$

where: $\theta_j = \sin^{-1}\left(\sqrt{\frac{j - X_j}{n}}\right)$, $\phi_k = \pi \frac{k - Y_k}{n}$, X_j and Y_k are uniform random numbers between 0 and 1. $2n^2$ is the total number of samples.

We implemented it by dividing the hemisphere into $2n^2$ parts and sending out a random sample ray for each one. This method could obtain better estimate than the pure sampling strategy on the whole hemisphere. [8]

5.2 Incident Radiance Cache [3,8,9]

The sampling strategy mentioned above will introduce heavy computation. Instead, we perform the sampling process only as needed for the sake of efficiency. If one or more values are stored near the point being sampled, we use the stored values. Otherwise, we invoke the process mentioned above to compute the new value and store the result in an octree-tree structure.

To determine whether or not there are computed values nearby, we estimate the smoothness of the local incident radiance on the surface around the sample location. We looked at the relative change $\varepsilon_i(\vec{P})$ in radiance as the surface location and orientation changes:

$$\varepsilon_i(\vec{P}) = \frac{\left\|\vec{P} - \vec{P}_i\right\|}{R_i} + \sqrt{1 - \vec{n} \cdot \vec{n}_i} \qquad (5)$$

where $\vec{P}$ is the newly sampled point location (for which we want to compute the change) and $\vec{P}_i$ is the location of restored sample i . R_i is the harmonic mean distance to objects visible from $\vec{P}_i$. $\vec{n}$ and $\vec{n}_i$ are the normal of the newly sampled point and the restored sample i respectively.

Given this estimate of the local change in incident radiance, we developed an octree structure where previously stored samples could be reused whenever possible. All samples are stored in the following tree, which makes it possible to quickly locate previous samples:

```
typedef struct ambtree {
        AMBVAL  *alist;
                        /* value list for the points nearby*/
        struct ambtree     *kid;          /* 8 child nodes */
}    AMBTREE;
```

When a new sample is requested, the octree is queried first for previous samples near the new location. For these nearby samples the change in radiance ε_i is computed. If samples with a sufficiently low ε_i are found then these samples are blended as the follows:

$$E(\overset{\varpi}{P}) = \frac{\sum_{i:w_i(\overset{\varpi}{P})>1/\alpha} w_i(\overset{\omega}{P})\left[E_i + \left(\overset{\varpi}{n_i} \times \overset{\varpi}{n}\right)\cdot\nabla_r E_i + \left(\overset{\omega}{P} - \overset{\omega}{P_i}\right)\nabla_t E_i\right]}{\sum_{i:w_i(\overset{\varpi}{P})>1/\alpha} w_i(\overset{\varpi}{P})} \quad (6)$$

where: $w_i(\overset{\varpi}{P}) = \frac{1}{\varepsilon_i(\overset{\varpi}{P})}$ is the weight of sample i. α is a user-selected accuracy goal. E_i is the incident radiance value at sample i. $\nabla_r E_i$ and $\nabla_t E_i$ are the rotational gradient and translational gradient respectively.

The whole procedural of the diffuse reflection calculation works as follows:

```
Query the octree for new values;
if (one or more values is stored near the point)
    Compute the new value with the stored values;
else{
    Send out a multitude of sampling rays R1,R2,…, Rn in random directions;
    for each sample ray Ri
        raytrace(&Ri, Depth-1);
    $I_a$ = ( R1->color + R2->color +…+ Rn->color)/n;
        /*n is the total number of the sampling rays*/
    Store the new value in octree;
}
```

Fig.5 shows the result of our algorithm.

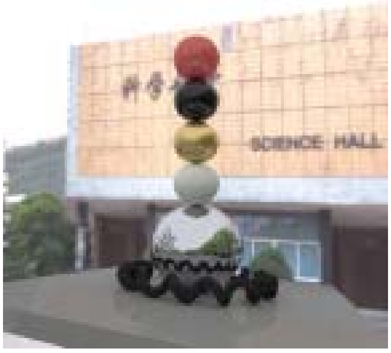


Figure 5 The Result of Our Algorithm

6. CONCLUSIONS

Now we have presented a general method that could integrate the CG object into the virtual scene seamlessly. We record the true illumination in light-probe images, which can be used as both the illumination information and the representation of the scene when the rendering algorithm is applied. Some additional work could be done to accelerate the rendering process. For example, we could dynamically send out sample rays according to some criterion to make the rendering process faster.

7. REFERENCES

[1] P.E. Debevec and J. Malik, “Recovering High Dynamic Range Radiance Maps from Photographs”, Computer Graphics (Proc. Siggraph 97), ACM Press, New York, 1997, pp369-378

[2] Paul Debevec, “Rendering Synthetic Objects into Real Scenes: Bridging Traditional and Image-Based Graphics with Global Illumination and High Dynamic Range Photography”, Computer Graphics (Proc. Siggraph 98), ACM Press, New York, 1998, pp189-198

[3] Gregory J. Ward and Paul Heckbert. “Irradiance gradients”, Third Eurographics Workshop on Rendering, pp 85-98, May 1992

[4] N. Green, “Environment Mapping and Other Applications of World Projections”, IEEE Computer Graphics & Applications, 6(11):21-29, November 1986

[5] Richard Szeliski, “Video Mosaics for Virtual Environments”, IEEE Computer Graphics and Applications, 1996.3:22-30

[6] Zhang Tao, Chen Yiming, “Using OpenGL to Navigate the Full View of a Panoramic Image on PC”, Computer Engineering and Applications, 2001.7

[7] Shi Yihui, Lu Dongming, Pan Yunhe, “Technical Research on Virtual Color Shading of Dunhuang Fresco”, Application Research of Computers, 2002.6

[8] Henrik Wann Jensen, et al, “Monte Carlo Ray tracing”, Siggraph 2003 Course 44

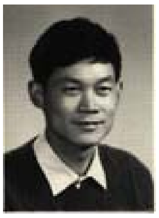
[9] Ward, Gregory and Maryann Simmons, “The Holodeck Ray Cache: An Interactive Rendering System for Global Illumiantion in Nondiffuse environments”, ACM Transactions on Graphics, 18(4):361-98, October 1999

[10] J.F. Blinn and M. E. Newell, “Texture and Reflection in Computer Generated Images”, Communiacations of the ACM, vol. 19, no. 10, Oct. 1976, pp542-547

[11] Paul Debevec, “Image-based Lighting”, IEEE Computer Graphics and Applications, March/April 2002, pp26-34

[12] Jan Kautz, Katja Daubert and Hans-Peter Sedel, “Advanced Environment Mapping in VR Applications”, OpenSG symposium 2003

Wang Jun: A postgraduate student of the School of Information Technology, Southern Yangtze University, P. R. China. He obtained his bachelor degree from Wuxi University of Light Industry in 2001. Research interests include Computer Graphics and Virtual Reality.



Gu Yaolin : A Full time Professor and Deputy Dean of the School of Information Technology, Southern Yangtze University, P. R. China. He graduated from the Computer Science Department of Shanghai Jiaotong University in 1982. As a Visiting Scholar of the University of Stony Brook in USA from 1994 to 1995, his main research interests are Computer Graphics and Parallel Computing. Now he is a Senior Member of Computer Society, IEEE.

Network Architecture for Real-Time Distributed Visualization and 3D Rendering

Lamei Yan[1], Xiaohong Zeng[1], M.Mat Deris[2]
[1]Zhuzhou Institute of technology, Hunan, China,412008
E-mail : y.lm@163.com **Tel :** 0733-2622808, 13017120018
[2]Department of Computer Science, University College of Science & Technology Malaysia
E-mail: mustafa@uct.edu.my

ABSTRACT

Online visualization enables developers to test, debug, and monitor the behavior of distributed systems, while they are running. In this paper ,we present a new network architecture for 3D real-time visualization that utilizes a cluster of conventional PCs to generate high-quality interactive graphics. A distributed real-time system for visualization and 3D rendering is presented which uses distributed interaction for control. Our work is unique in that it supports both an online and off-line visualization and rendering of a distributed system.

Key words: real-time rendering; distributed visualization system, visualization

1. INTRODUCTION

Scientific visualization is the process of using computer graphics to view or explore 3D objects, natural phenomenon, or complex data. Many real-time systems, particularly hard real-time systems, are not correct unless they can meet their deadlines. Rate monotonic analysis and other scheduling strategies have been developed to help ensure that real-time systems achieve this goal.

Software for visualizing real-time system behavior must address the following significant challenges that are not faced when visualizing the behavior of non-real-time systems. First, the visualization framework must not interfere with the correct timing behavior of the real-time system. Second, the framework must be flexible to address diverse system behaviors, particularly when sources of non-determinism appear. Finally, the framework must support both independent and correlated visualizations of distinct event streams.

In the last decade the visualization of virtual environments and interaction within was possible only with specialized hardware. This hardware was very expensive, had a lack of scalability and used specific protocols, busses, networks for communication and specialized graphics hardware for visualization.

Despite large advances in computer technology, some types of visualization and rendering techniques (such as ray tracing [1]) are still too slow to be used for real-time applications on desktop computers. Fortunately, visualization / rendering is a problem that lends itself very well to parallel processing [2]. By using large multiprocessor computers, techniques such as real-time ray tracing are feasible. Parker et al. have developed a real-time ray tracing system that was demonstrated on a 60-CPU Silicon Graphics Origin 2000 [3]. Their system performs well enough to be useful for many applications-one of their examples is a visualization tool for CT Scan data sets [4], [5].

In this paper, we describe the Distributed Visualization System (DVS), a Java-based software architecture for real-time distributed or cluster-based visualization. The distributed interaction and visualization system has two major parts:

1. Processing of the interaction data and fusion of information.
2. Updating of all affected data and displaying the new state of the application.

The next section will give a brief overview of the related work**.** Section 3 introduces a System Architecture and Control of visualization network by interaction distributed visualization and distributed interaction processing. Section 4 we present the performance numbers and show the amount of 3D texture data moving between the application PEs and the rendering servers. Finally, we draw some conclusions in section 5.

2. RELATED WORK

A lot of research has been performed in the fields of distributed distribution and rendering, network architectures for distributed rendering, distributed multimedia systems, interactive scene representation and cooperative work in the last ten years. Our approach concerns these fields and we will review the most significant work in these fields.

For visualization of large scientific simulations it’s useful to render on a remote or distributed renderer and only visualize results on local displays. The most efficient approach for remote and distributed rendering is WireGL [8, 9]. It provides the OpenGL API to each node in a cluster, virtualizing multiple graphics accelerators into a sort-first [10] parallel renderer with a parallel interface.

Considerable work has been done in the visualization and analysis of the execution of Java programs (see, e.g. [11]). Jinsight [12] does visualization of trace information produced by a special instrumented version of the Java Virtual Machine. Similarly, Walker et al. [13] use program event traces to visualize program execution patterns and event-based object relationships like method invocations. All these systems support only an offline mode where the trace of the program execution gets visualized. Also they are not designed for visualizing a distributed system. Jive [14] does on-line visualization of Java programs. Jive also supports only non-distributed systems.

Our work is unique in that it supports both an online and off-line visualization and rendering of a distributed system. It also allows an easy mapping of the high-level abstractions through a specification language.

3. SYSTEM ARCHITECTURE

3.1 System Architecture

DVS is an object-oriented design that implements the basic functionality needed for real-time distributed rendering, but not the actual rendering algorithms. The popular Parallel Virtual Machine (PVM) [6] or Message Passing Interface (MPI) [7] software packages were available to us. DVS' architecture is relatively straightforward (Fig.1). Each PC in the cluster is responsible for producing pixels for a small part of the display.

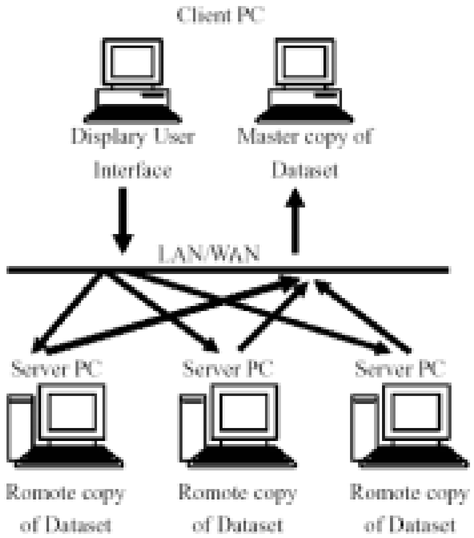


Fig.1 the Structure of Distributed Visualization System

3.2 Control of Visualization Network by Interaction

In many virtual environments user interaction has to be processed. We assume that every type of interaction processing has one process (interaction server) which is able to transmit interaction processing information to our visualization network (see fig.2). For distributed interaction processing one process has to collect the relevant interaction information, fuse the information and transmit it to the visualization network.

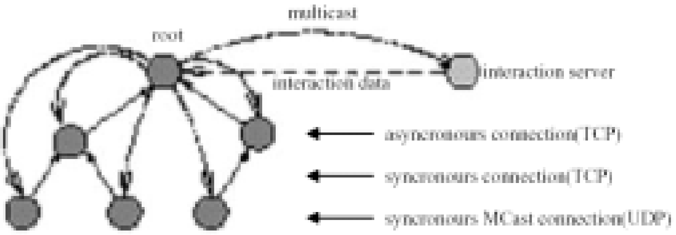

multicast
root
interaction server
interaction data
asyncronours connection(TCP)
syncronours connection(TCP)
syncronours MCast connection(UDP)

Figure 2. Architecture for visualization and interaction processing interaction

This protocol is optimal with respect to the number of messages sent and therefore produces few collisions on the network. The protocol is also able to distribute payload information on each synchronization step to the graphics nodes.

4. PERFORMANCE METRICS

The performance numbers we present below show the amount of 3D texture data moving between the application PEs and the rendering servers. For the testing, 20 PCs were used as rendering servers, not including the client PC. These PCs ran a mix of Windows 98/NT/2000 and Red Hat Linux version 6.2 Operating systems. They ranged from a Pentium-II 266MHz to a dual-processor Pentium-IV 2.4GHz. The average PC tested was a single-processor 800MHz Pentium-III**.**

Figure 3 shows that the amount of 3D texture traffic inbound at all parallel servers for a given dataset is a function of the number of the displays, not the number of application PEs.

This represents the fact that some blocks of 3D texture data are sent to more than one rendering server, since one 3D volume block may project onto more than one display tile. Figure 4 shows the amount of 3D texture traffic inbound to a varying number of graphics servers from five application nodes. Increasing the number of rendering servers increases the amount of total traffic. Smaller block sizes will result in fewer overlaps, and decreased duplication. In Figure 5, we see the amount of 3D texture data inbound to each of the five parallel rendering servers in a six-display configuration. Figure 6 shows the maximum bandwidth requirement for each of the configurations of parallel rendering servers we tested in our runs.

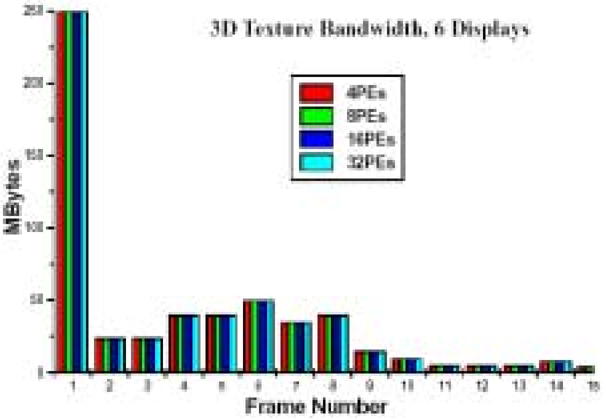


Figure.3. Amount of 3D texture data inbound to all rendering servers on each frame, varying the number of application PEs

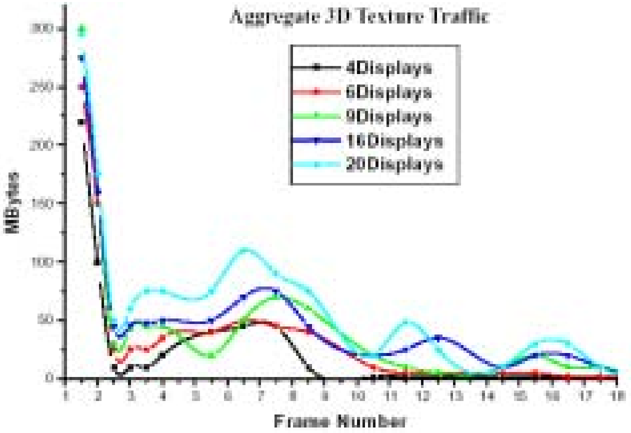


Figure.4 Per-frame 3D texture data inbound to Chromium rendering servers during a parallel sort-first volume rendering transformation sequence

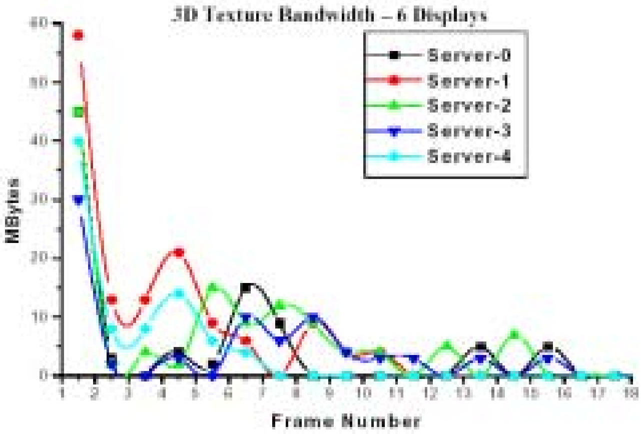


Figure.5 Per-frame 3D texture bandwidth for six parallel rendering servers

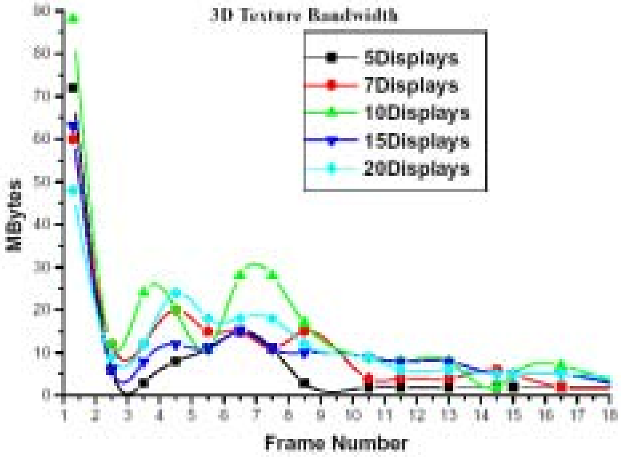


Figure.6 Per-frame 3D texture bandwidth for several configurations of parallel rendering servers

5. CONCLUSIONS

In this paper, we describe the Distributed Visualization System (DVS), a Java-based software architecture for real-time distributed or cluster-based visualization. DVS is an object-oriented design that implements the basic functionality needed for real-time distributed rendering, but not the actual rendering algorithms. A distributed real-time system for visualization and 3D rendering is presented which uses distributed interaction for control. Our work is unique in that it supports both an online and off-line visualization and rendering of a distributed system. It also allows an easy mapping of the high-level abstractions through a specification language. The performance numbers we present show the amount of 3D texture data moving between the application PEs and the rendering servers.

6. REFERENCES

[1] An Introduction to Ray Tracing. A.S. Glassner, ed., Academic Press, 1989.

[2] A. Chalmers and E. Reinhard, “Parallel and Distributed Photo-Realistic Rendering,” Proc. Course#3, SIGGRAPH98Conf., July 1998.

[3] S. Parker, W. Martin, P.-P.J.Sloan, P. Shirley, B. Smits, and C. Hansen, “Interactive Ray Tracing,” Interactive 3D, Apr. 1999.

[4] S. Parker, M. Parker, Y. Livnat, P.-P. Sloan, C. Hansen, and P. Shirley, “Interactive Ray Tracing for Volume Visualization,” IEEE Trans. Computer Graphics and Visualization, pp. 238-250, July-Sept. 1999.

[5] S. Parker, P. Shirley, Y. Livnat, C. Hansen, and P.-P. Sloan, “Interactive Ray Tracing for Isosurface Rendering.” Proc. IEEE Visualization ’98, 1998.

[6] A. Beguelin, A. Geist, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam, PVM: Parallel Virtual Machine, A Users’ Guide and Tutorial for Networked Parallel Computing. MIT Press, 1994.

[7] P. Pacheco, Parallel Programming with MPI. Morgan Kaufmann, 1996. Many real-time systems, particularly hard real-time systems, are not correct unless they can meet

[8] I. Buck, G. Humphreys, and P. Hanrahan. Tracking Graphics State For Networked Rendering. In Proceedings of SIGGRAPH 2000, August 2000.

[9] G. Humphreys, M. Eldrigde, I. Buck, G. Stoll, M. Everett, and P. Hanrahan. WireGL:A Scalable Graphics System for Clusters. In Proceeding of SIGGRAPH 2001, pages 129– 140, August 2001.

[10] S. Molnar, M. Cox, D. Ellsworth, and H. Fuchs. A sorting classification of parallel rendering. In IEEE Computer Graphics and Algorithms, pages 23–32, July 1994.

[11] W. D. Pauw, E. Jensen, N. Mitchell, G. Sevitsky, J. Vlissides, and J. Yang. Software visualization, state-of-the-art survey. LNCS 2269, 2002.

[12] G. Sevitsky, W. D. Pauw, and R. Konuru. An information exploration tool for performance analysis of java programs. March 2001.

[13] R. J. Walker, G. C. Murphy, B. N. Freeman- Benson, D. Wright, D. Swanson, and J. Isaak. Visualizing dynamic software system information through high-level models. In Conference on Object-Oriented, pages 271–283, 1998.

[14] S. P. Reiss. Jive: Visualizing java in action demonstration description. 2003.

[15] J. Mahovsky, L. Benedicenti. An Architecture for Java-Based Real-Time Distributed Visualization. pages 570-579, October-December 2003.

Image Segmentation Based on Fuzzy Maximum Entropy and Simulated Annealing Algorithm

LIU Huikang, LI Juan, WU Jin
College of Information Science and Engineering, Wuhan University of Science & Technology,
Wuhan, Hubei, 430081, China,
Email: hust_wu@163.com, lilian0122@163.com **Tel:** +86(0)27 62000502

ABSTRACT

Thresholding is an important topic for image processing, pattern recognition and computer vision. Selecting thresholds is a critical issue for many applications. It is generally believed that many image properties, such as brightness, boundary, region, etc., are fuzzy in nature, and can be described as a fuzzy set. In this paper, we propose a method for image segmentation, basing on the maximum fuzzy entropy. The brightness for gray levels of an image is used as a fuzzy set, the Maximum Entropy Principle is used as the criterion to find the optimal membership function which will best represent the membership of the brightness for each gray level of an image. The simulated annealing algorithm is used to obtain the optimal threshold value for the image segmentation. Experiments on many images have been conducted. The experimental results show that the proposed method can segment an image effectively and rapidly. The main features of the original images can be preserved very well.

Keywords: Image Processing, Thresholding, Fuzzy Set, Maximum Entropy, Simulated Annealing Algorithm

1. INTRODUCTION

Image thresholding, which extracts the object from the background in an input image, is one of the most common applications in image analysis. Among the image thresholding methods, bilevel thresholding separates the pixels of an image into two regions: the object and the background. One region contains pixels with gray values smaller than the threshold value, and the other contains pixels with gray values larger than the threshold value. Many research studies have been devoted to the problem of selecting the appropriate threshold value. The survey of these papers can be seen in the [1, 2].

Fuzzy set theory has been successfully applied to many image processing and pattern recognition tasks. Because many image properties, such as brightness, boundary, region, etc., are fuzzy in nature. Fuzzy set can describe the properties of the image felicitously. In this paper, fuzzy set theory is applied to partition the image into two regions by maximizing the entropy of fuzziness of the image [3].

The proposed method starts with the concept of the fuzzy event introduced by Zadeh [4]. Based on his definition, a fuzzy event is a fuzzy set in the sample space. The “brightness of gray levels” is a fuzzy event, and can be described as a fuzzy set. According to the Maximum Entropy Principle, a fuzzy event contains most information when its associated entropy is maximum. Therefore, our purpose is to find a brightness membership function. We adopt membership function with rectangular distribution which is determined by the parameters a and c. The original problem of how to determine the membership function becomes the problem of how to find a best combination of parameters (a_{opt}, c_{opt}). The simulated annealing algorithm is used to solve this problem.

2. FUZZY MAXIMUM ENTROPY

2.1 Brightness of Gray Levels As a Fuzzy Event

We consider an image having L gray levels， ranging from r_0 to r_{L-1}, and a histogram of $h(r_k)$,k=0,1,…,L-1.In fuzzy set theory, a membership function is used to denote the degree of an element in the sample space belonging to a fuzzy set. For the fuzzy set “brightness of gray levels”, the membership function $\mu_{bright}(r_k)$ can be used to denote the degree of brightness possessed by the gray level r_k, k=0,1,…,L-1.It can be written as:

$$bright = \mu_{bright}(r_0)/r_0 + \mu_{bright}(r_1)/r_1 + ... + \mu_{bright}(r_{L-1})/r_{L-1} \tag{1}$$

Where “+” means union.

The probability of this fuzzy event can be computed by

$$p(bright) = \sum_{k=0}^{L-1} \mu_{bright}(r_k)\, p(r_k) \tag{2}$$

Where $P(r_k) = h(r_k)$, k=0,1,…,L-1.

2.2 Entropy of a Fuzzy Event

The entropy that is used as a measure of fuzziness is in analogy with the entropy in information theory [5]. The entropy for the occurrence of a fuzzy event A can be defined as:

$$H(A) = -P(A)\log(P(A)) - (1-P(A))\log(1-P(A)) \tag{3}$$

According to information theory, the larger the entropy for an occurrence of the fuzzy event is, the more information the fuzzy event has.

2.3 Membership Function

The membership function we adopt is defined as follows:

$$\mu_{bright}(x) = \begin{cases} 0 & x \le a \\ \dfrac{x-a}{c-a} & a < x < c \\ 1 & x \ge c \end{cases} \tag{4}$$

Where x is the independent variable and a and c are parameters determining the shape of the above membership function as shown in Fig. 1.

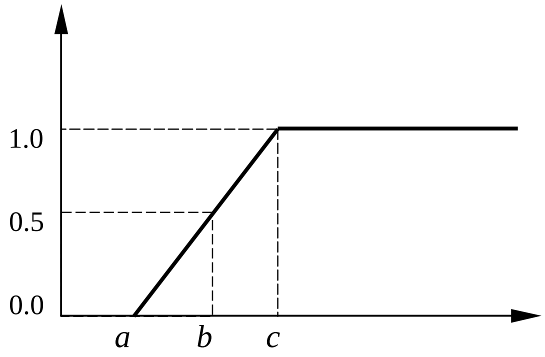


Fig. 1 The membership function $\mu_{bright}(x)$

The gray levels smaller than a have the membership of 0 for the fuzzy set brightness, which means that pixels with these gray levels do not belong to the white group at all. As the gray level increases, the membership becomes greater and greater. When the gray level is greater than c, the membership for the white group becomes 1.0, which means that pixels with these gray levels definitely belong to the white group.

When a pixel's membership is 0.5 degrees of belonging to the white group, it also has 0.5 degrees belonging to the black group. In this case, it could be classified either into white or black group. Thus, in this paper, we select this point as the threshold value. By our design, it is the mid-point of the interval [a , c].

So, we want to find a best combination of parameters (a_{opt} , c_{opt}) such that

$$H_{\max}(bright; a_{opt}, c_{opt}) = \max\{H(bright; a, c) \mid r_0 \le a < c \le r_{L-1}\} \quad (5)$$

From the best parameter combination (a_{opt} , c_{opt}), we can obtain the optimal threshold value th :

$$th = b_{opt} = (a_{opt} + c_{opt})/2 \quad (6)$$

3. SIMULATED ANNEALING ALGORITHM

For an image with 256 gray levels ,in the case mentioned in the previous section, the search space is about 256*255/2≈ 3.3*10^4.The search is exhaustive. Hence, another searching algorithm is necessarily needed. In this paper, we use the simulated annealing algorithm to solve this problem. The simulated annealing algorithm was first introduced by Kirkpatrick et al [6]. There are four ingredients of the simulated annealing algorithm, they are described as follows:

(1) Configuration representation.
The configuration of this task is a list of two parameters,

$$X = (a, c) \qquad 0 \le a < c \le 255 \quad (7)$$

(2) Move set
There are four possible moves:

$$\begin{aligned} (a,c) &\to (a-1,c) \\ (a,c) &\to (a+1,c) \\ (a,c) &\to (a,c-1) \\ (a,c) &\to (a,c+1) \end{aligned} \quad (8)$$

(3) Cost function
The simulated annealing algorithm is an algorithm to search the minimum, the cost function can be defined as:

$$Cost(X) = 1.0 - H(bright; X) \quad (9)$$

(4) Cooling schedule

$$T_{n+1} = \alpha T_n \quad (10)$$

Where α is the cooling rate, 0< α <1.

4. PROPOSED METHOD

We assume that the image has 256 gray levels. The detailed procedures are described below:

(1) Read the image and compute its histogram $h(r_k)$,k=0,1,…,255.

(2) Generate an initial state X_{init} =(a , c), randomly select a's value from [0,127], randomly select c's value from [128,255]. We set X_{init} =(100,150) and let $X_{curr} = X_{init}$.

(3) We set the initial temperature $T_0 = 1.0$ and the cooling rate α =0.996.

(4) This is a loop for finding the solution in search space.
Loop {

① According to many experiments by other people ,we can find a solution in about 3000 iterations. If the iterations are met, leave the loop.

② Randomly select a move from the move set, and apply it to X_{curr}, which produces X_{new}. If a <0 or $a \ge c$ or c >255, X_{new} is an illegal state, try another move.

③ Compute the cost change:
$\Delta E = Cost(X_{new}) - Cost(X_{curr})$

④ If $\Delta E \le 0$, X_{new} is a better state:
$X_{curr} = X_{new}$.

⑤ If $\Delta E > 0$,set p= $e^{-\Delta E / T_{curr}}$. If p<y, where y is a random number between [0,1], we set y=0.5, then $X_{curr} = X_{new}$. Otherwise, do not change the current state.

⑥ Reset the current temperature according to the cooling schedule.
}

(5) Return X_{curr} =(a_{opt} , c_{opt}).

(6) We can obtain the optimal threshold value $th = (a_{opt} + c_{opt})/2$.

5. EXPERIMENTAL RESULTS

We have done experiments on many images. We just use a few of them to demonstrate the performance of the proposed approach comparing it with Otsu's method and maximum entropy thresholding.

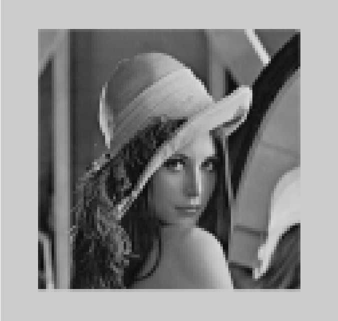


Fig.2 The original image "LENA".

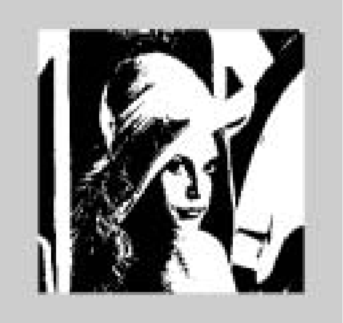


Fig.3 Thresholding image using Otsu's method. (th =101)

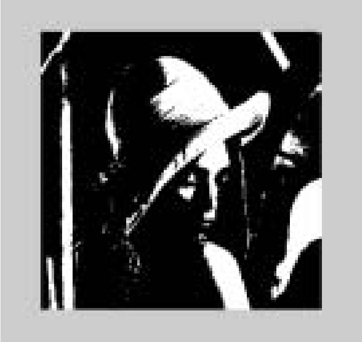


Fig.4 Thresholding image using maximum entropy principle. (th =141)

.

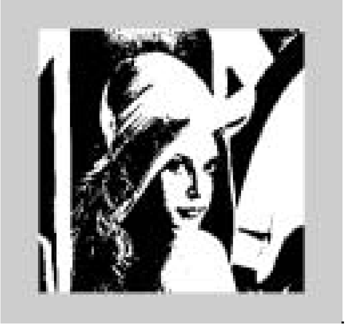


Fig.5 Thresholding image using the proposed method. (a_{opt}, c_{opt}) = (91， 102), (th =96)

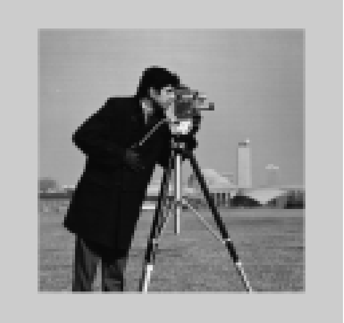


Fig.6 The original image "cameraman".

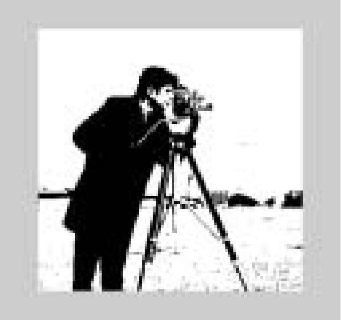


Fig.7 Thresholding image using Otsu's method. (th =89)

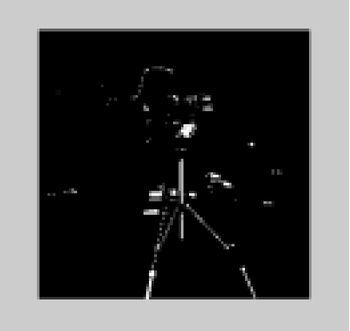


Fig.8 Thresholding image using maximum entropy principle. (th =193)

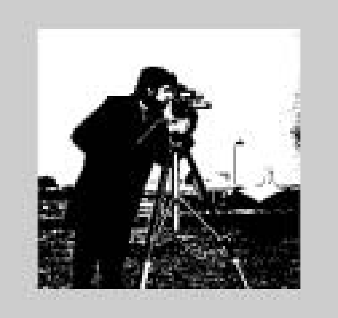


Fig.9 Thresholding image using the proposed method. $(a_{opt}, c_{opt}) = (140，147), (th = 143)$

Fig. 2 is the original image “LENA” with size 256*256. The results of thresholding of the original image using Otsu’s method, maximum entropy thresholding and the proposed method are given in Fig. 3－Fig. 5, respectively. In Fig. 3, the main features of the original image are well preserved, but the nose and the mouse of “LENA” are noisy. In Fig. 4, some contours of the objects in the image disappear, and the image is much noisier, especially the face and the hair of “LENA”. In Fig. 5, the main components of the image are well segmented, and the main features of the image are well preserved.

Fig. 6 is the original image “cameraman” with size 256*256. The results of thresholding of the original image using Otsu’s method, maximum entropy thresholding and the proposed method are given in Fig. 7－Fig. 9, respectively. In Fig. 7, some buildings in the image have vanished. In Fig. 8, most buildings in the image have vanished, and the man in the image nearly merges with the background. In Fig. 9, the main features of the original image are well preserved, the main components are segmented better than those in Fig. 7 and Fig. 8.

According to experimental results, the proposed method can automatically and effectively segment the image.

6. CONCLUSION

In many image-processing tasks, the brightness of the gray levels in an image is one of the most useful pieces of information. In this paper, we use “brightness of gray levels” of an image as a fuzzy set and partition the image space into two regions by maximizing the entropy of fuzziness of the image. And the use of the annealing algorithm reduces the computational time.

The experimental results have shown the effectiveness and usefulness of the proposed algorithm for image thresholding.

7. REFERENCES

[1] P.K.Sahoo, S.Soltani and A.K.C.Wong. A survey of threshold techniques. Comput. Vision Graphics Image Process, Vol.41, 1988, pp.233~260.

[2] S.U.Lee and S.Y.Chung. A comparative performance study of several global thresholding techniques for segmentation. Comput. Vision Graphics Image Process. Vol.52, 1990, pp.171~190.

[3] P.K. Saha, J.K. Udupa, and D.Odhner. Scale-Based Fuzzy Connected Image Segmentation: Theory, Algorithms, and Validation. Computer Vision and Image Understanding, Vol.77, 2000, pp.145~174.

[4] L.A.Zadeh. Probability measures of fuzzy events. J. Math. Anal. And Appl. Vol.23, 1968, pp. 421~427.

[5] Cheng H.D., Chen J.R., Li J.G.. Threshold Selection Based on Fuzzy C-Partition Entropy Approach [J]. Pattern Recognition, Vol.31, No.7, 1998, pp.857~870.

[6] S.Kirkpatrick, C.D.Gelatt, Jr., and M.P.Vecchi. Optimization by simulated annealing. Science, Vol. 220, 1983, pp.671~680.

LIU Huikang received the Eng. Degree in Industry Automation in 1985, and the M.Eng. degree in Theory and Engineering of Control in 1988, both from Wuhan Iron and Steel University. He is a Vice Professor, also a Major Researcher of Department of Automation, Wuhan University of Science and Technology. He has published over 5 Journal papers, obtained Prenium of Science and Technology from Hubei Province. His main areas of interest are intelligent control.

LiJuan received the Eng. degree in Electronics and Information Engineering in 2001 from Huazhong University of Science and Technology, now is a master of Wuhan University of Science and Technology. Her main areas of interest are image processing and analysis, pattern recognition.

The Study of Image Modeling in the Digital City Based on Distributed Computation

Lufeng GeShun
Information Technology Department, WuHan University of Technology
WuHan, HuBei Province, China
Email: lufengwut@126.net Tel: 027-87870877

ABSTRACT

After introducing the Digital City and the distributed computation, the paper advances a new thought to use distributed computation in the work on large amount of data in the Digital City, especially to generate image. Then gives the computing model. Finally make the looking forward of this thought.

KeyWords: the Digital Earth; the Digital City; coordinates-counterchange; distributed computation; image modeling;

1. INTRODUCTION

Now, the study of the Digital City is very hot. The concept of the Digital City came from the concept of the Digital Earth. This “Digital Earth” we are talking about is a three-dimension expression which can contain a large amount of geographical coordinates data and has high resolution[1]. And because it is city that be the geographical syntheses of high dense of population、resources、environment and social economic factors, it is considered to be the most important component on the earth and further more, the city is also very complex，active and frequent communication[12]. That is to say, the study of the Digital City will be the most important part of the study in the Digital Earth.

When we talk about the concept of the Digital City, we can see it as a integrate system. That is to say, the whole or the most part of the city, including basic establishments, digital function establishments and database building, are connected with high speed network. It is made controlling more freely more automatically and more intelligently. On the other hand, if we consider the system from solving the city problems, we can compartmentalize it into three stages[3]: the first stage is to build the Digital City with the existing hardware, software and the database; the second stage is to generate three-dimension models with professional instruments and perfect software, and then render them to get city model which can change information with people; and the third stage is to devote more advancing technologies into the system and to build advanced Digital City system. Thus, our studies now are in the second stage, and the system users now are concerning more about the modeling technology of the city information in this stage.

2. THE THREE-DIMENSION IMAGE MODELING IN THE DIGITAL CITY

From that mentioned before, we know the Digital City system concerns very broad, including data integration, large amount of data storage, 3S technology, dynamic information exchange, network technology, multi-dimension virtual reality and so on[10]. This article talks about the technology of image modeling only. From the above, it’s known that the main modalities of generating images is virtual reality and three-dimension modeling, and the image quality affects the evaluation of the whole Digital City system, so let’s introduce the image generation first.

In the image generation of the Digital City, the whole process can be divided into two steps: the shape modeling and the rendering. The shape modeling is to sketch the outer shape of the scene with simple formers; and the rendering is to render the figures that the shape modeling generated just now to get the more vividly scenery. If the first step doesn’t exit, the second one can’t be done, and if the second step doesn’t exit, the simple figures that modeling only by the first step couldn’t be comprehended easily.

The Shape Modeling

The shape modeling is the former step of the image generation. In this step, there is a very important process called coordinates-counterchange, it will be introduced as follows[4]. Show as the figure 1.

The coordinates of the point A in the absolute coordinate space O-XYZ is (x,y,z), and the coordinates in the observer coordinate space O’-UVW is (u,v,w). Here, O’ is the observer point. So, through the coordinates-counterchange, the relations between the new coordinates (u,v,w) in observer coordinate space O’-UVW and the coordinates (x,y,z) in absolute coordinate space O-XYZ are:

$$\begin{cases} u = \dfrac{(x\cos\alpha - y\sin\alpha + l)D}{x\sin\alpha\sin p + y\cos\alpha\sin p + z\cos p + n} & (1) \\[2ex] v = \dfrac{(x\sin\alpha + \cos p + y\cos\alpha\cos p - z\sin p + m)D}{x\sin\alpha\sin p + y\cos\alpha\sin p + z\cos p + n} & (2) \\[2ex] w = \dfrac{D}{\cos\beta} & (3) \end{cases}$$

In the relationship above, p is the angle nipped by the horizontal line and axis *W*, α is the angle nipped by the axis X and the line that though the point O and the point which is projected into the surface XOY by O’, *m, n, l, D* are distances.

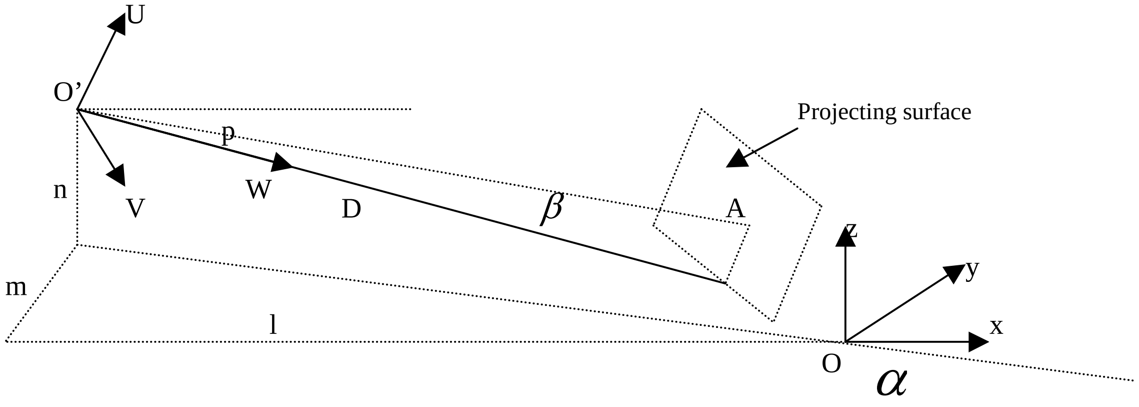


Figure 1: coordinates-counterchange

Use the same way, we can counterchange lines and surfaces from the absolute coordinate space O-XYZ into the observer coordinate space O'-UVW. From this, we know that the work of coordinates-counterchange will be very heaven even modeling one static image of the city only, which would consume a lot of memories and the CPU time of PC. Further more, if we want to "visit" the Digital City----that is to say, to make the image moves, it will be a heavier work to counterchange the coordinates because the observer point will move ceaselessly.

The Rendering

To render a generated figure will be a much more burden. Generally, we compartmentalize surfaces differently such as the frontispiece faces, the side faces, the distinguishingly terrain and so on to simplify the heavy work[8].

In this step, to grab the tiny texture character is a very elaborate work. Texture data is always grabbed from on the spot photograph outdoors, which related the three-dimension scenery tightly. And its quality will decide the final effect of the scenery and even the Digital City system. So to render faces to get vivid sceneries is also a very heavy work[14].

From the above, we know that the virtual reality and the three-dimension modeling of the Digital City are so heavy burdens. But on the other hand, the real time character of the Digital City is also a very important target. Obviously, under the conditions that the hardware and software of the system haven't been changed notably, it's very hard for only one computer to do the work to cater to users' needs. And now, with the development of the technology, the distributed computation which has been more and more concerned gives us new thought to solve this problem.

3. THE DISTRIBUTED COMPUTATION

It's been 50 years from the computers' appearance till now, while the competence of computer advanced remarkable, especially the speed of computation, has been about hundreds of thousands of times as before. But limited by some factors such as velocity of light, physics measure and so on, the speed of computation couldn't be too fast[2]. To solve this problem, so many new thoughts are raised, such as quantum computation. This method has so bright future, but it only could do simple computations now, in another words, it couldn't be put in use currently. At present, parallel computation with multi-CPUs quicken the speed of computation, but its high cost limits the application. On the other hand, the study of parallel computation with multi-PCs advances so fast, using multi-PCs to do the corresponding computation to achieve the computation speed as huge computer or multi-CPUs computer offers has been put in use and has achieved a lot[7].

The work process of distributed computation are as follows: when the server find the CPU of client PC is free, it will tell the manage server to send some work to this client. The client receives the work which the server sends to it, it will do the job at its free time, and send the result back to the server. When the job is done at client, the client computer will do it background or in other ways, but doesn't effect whatever being done at client at the moment. And if there is some new task will be done at this client, the CPU controlling will be turned back to the client in no time because the client can't bear its own work is delayed by other work.

At present, in some area of studies and expert applications, the distributed computation has showed its power contrasting to the technology before in solving problems. In conditions of not affecting the client job and not devoting more money, it can raise the performance of the whole system, and achieve the visible effect and more financial profit. Now, the distributed computation is concerned more and more in applications of electricity power[9], water conservancy and medical treatment and so on, and it will be applied more and more in the future[5].

4. THE DISTRIBUTED COMPUTATION IN THE IMAGE MODELING IN THE DIGITAL CITY

We know from the above that the three-dimension image modeling is a project that dealing with the great amount of data with complex method and professional processes, and the distributed computation can just solve this sort of problems very well than traditional methods. So, if we solve the image modeling in the Digital City with the method of distributed computation, we will obviously success.

Further more, when we talk about the hardware of the system, the Digital City is a gigantic system, whose future will be very well. Once the whole system was been set up, it will including so many fields such as government affairs solving, city layout[13], tour, information searching, urgent affairs management and so on, so the number of client PCs in the system will be very large, so the system will has the potential of dealing with large scales of computation. On the other hand, the advanced network technology ensures the information transmission speed also. Therefore, the hardware that the distributed computation runs at is satisfied, and it’s possible for us to use the distributed computation in the Digital City system.

The figure 2 shows how the distribute computation system runs[6]. The main pane in the middle is the system of the Digital City. By its sides, the two panes represent that the system has the ability to deal with the activity that the client PC joins or secedes the system to adapt the scale changing of the system so that the system can response the request of the client actively[11].

In this figure, let’s suppose that the client 2 has sent a request to the server (and let’s suppose that the request is about image modeling), the system’s response can be divided into four steps:

A、the client 2 sends the request to the server and waits for the response, at the same time, the client 2 joins to the system to contribute its own free CPU time for the distributed computation;

B、the server gets the corresponding coordinates data that will be dealing with from the database and divides the job into parts in some certain ways, then packs the work and distributes them to those who has been joined the system. Of course, let’s suppose one of the packages is sent to client 1 (as the figure shows and let’s suppose the client 1 had joined the system before);

C、the client 1 deals with the work and gets the result, and then, sends the result back to the server, and the other clients also send their results back to the server in this period of time because the work of every package will cost nearly the same long;

D、the server assembles the results into the final result and then sends it to the client 2 as the answer to the request it asked just now.

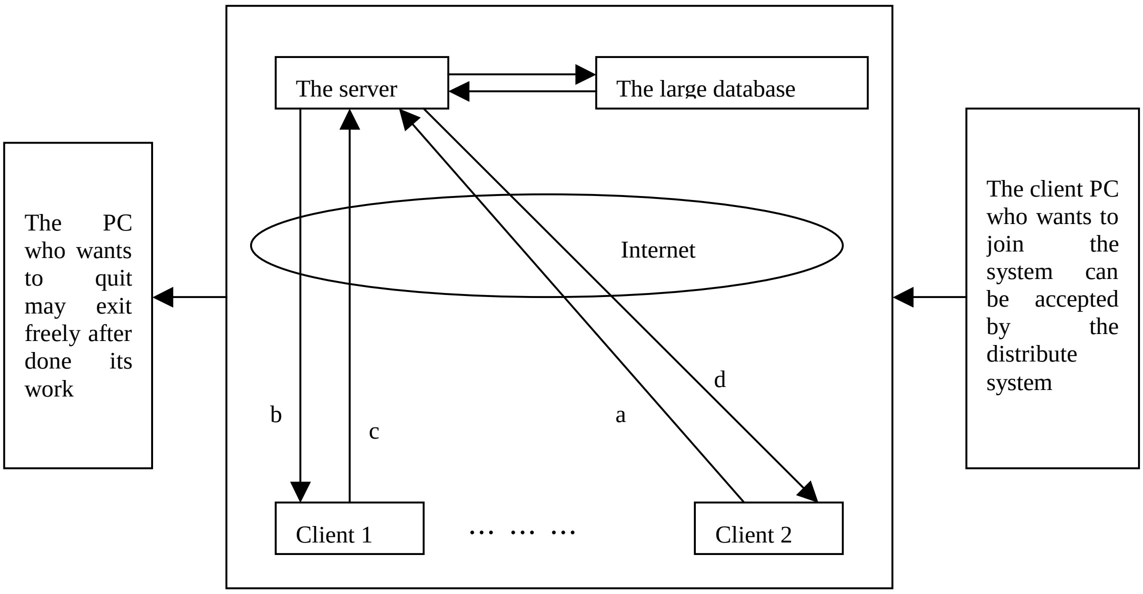

The server
The large database
Internet
The PC who wants to quit may exit freely after done its work
The client PC who wants to join the system can be accepted by the distribute system
b
c
a
d
Client 1
… … …
Client 2

Figure 2：Computation model of the distribution

In this way, the client 2 can be served of its request to modeling the image of the city, so as the other requests. The whole process seems very complex, but in fact, the speed of it is very fast in high speed network. This method may raise the effect of the modeled image and shorten the processing time. Especially in solving the problem of gigantic image modeling in the Digital City, it shows great advantages. And at the same time, this method release the server from the burden of the computation, it will deal with the maintaining of the system, answering to the request, getting the data from the database, packing data and communicates with client PCs, so its ability enhances, and the application fields of the system is widened.

5. CONCLUSION

The application of the distributed computation in the Digital City will be very bright. With the development of this technology, it can be more and more perfect and mature to advance the studies and applications of the Digital City. Of course, there are so many bottlenecks that limit the applications needed solving, such as the increasing the network speed to ensure the speed of the data transmission to shorten the working time, the improvement on the program of the distributed computation including the security and the redundancy and so on. But advantages of low cost, high speed and the better effect have determined the dominant status of the distributed computation in the Digital City.

6. REFERENCES

[1] LiLi Cui, Tao Huang “the digital city” and its key technology the study of the mapping software technology 2001 9 (in chinese)

[2] Chong-Wei Xu and Bo He PCB-A distributed computing system in CORBA Journal of Parallel and Distributed Computing 2000

[3] XueJun Duan, ChaoLin Gu, TaoFang Yu the pilot study of “the digital city” geography and the country study

2001 5 (in chinese)

[4] GuoJun Peng, TianHe Chi, LiYu Tang, ShuQin Wu three-dimension image building technology of the digital city the science of earth information 2003 5 (in chinese)

[5] XianTai Gou, WeiDong Jin the study of the distributed computation application in the next gap of network computer application 2003 8 (in chinese)

[6] HongTao Huang, Hui Wang the distributed computation based on Web computer application 2000 (in chinese)

[7] Chen Wang, Yong Meng Teo Supporting parallel computing on a distributed object architecture The Journal of Systems and Software 2001

[8] JianSi Yang, ZhiQiang Du, ZhengHong Peng, JingNan Huang, YongXi Chen the technology of the three-dimension image modeling in the digital city WuHan University Paper 2003 6 (in chinese)

[9] ChunFeng Yu, Gang Chen the application of the multi lays structure based on B/S computer and modernization 2002 11 (in chinese)

[10] JiangYa Gong the basic concepts and the realize strategies of the digital city 2001 annual paper collection of GIS 2001 (in chinese)

[11] M.S. Shephard, J.E.Flaherty, C.L.Bottasso, H.L.de Cougny, C.Ozturan, M.L.Simone Parallel automatic adaptive analysis Parallel Computing 1997

[12] AL Gore the digital earth—to know about our plant in the 21st century china science paper 1998 (in chinese)

[13] MingXing Hu the technology of virtual reality and its application in the city layout layout 2000 (in chinese)

[14] QingHui Sun, JunXi Zhao the three-dimension image modeling in the digital city earth information science 2003 9 (in chinese)

Lu Feng is a Full Professor and vice dean of Information Technology School, Wuhan University of Technology. He graduated from Wuhan University Technology in 1982; from Huazhong University of Science and Technology in 1989 with specialty of cybernation. He is a director of International Institute for General System Study, China sub-committees; director of Chinese Ceramic Society, Automatic sub-committees; standing director of Hubei Province Youth Science and Technology Society; director of Hubei Province electrician technology society; was a visiting scholar of Kanagawa University of Japan (1996). He has published two books, over 80 Journal papers. His research interests are in computer network communication, information system and information security technology, computer control and emulation; grey system theory and application.

Ge Shun is a graduate student of Information Technology of Department, Wuhan University of Technology. His research interests are in computer network communication, distributed computation and virtual reality.

Grey-Level Image Processing with a Parallel-Distributed System Model

Ge Hongwei XuWenbo
School of information engineering, Southern YangTze University
Wuxi, Jiangshu, People's Republic of China
Email: ghwjm@sina.com **Tel.:** +86 (0)510 2618967

ABSTRACT

In this paper, we propose a parallel-distributed system model that can be used to process grey-level image and present the description and analysis of the model in theory. One example of this system be applied to edge detection of grey-level image is also has been given.

Keywords: parallel-distributed system, pattern recognition, grey-level image, edge detection

1. INTRODUCTION

The traditional methods of image processing cannot process images in real-life environment rapidly and self-adaptively. One important implemental approach of image processing rapidly is using massively parallel distributed processing technology.

The RAM Weighted Adaptive Pattern Recognition System [1] can process image very rapidly. But the system can only process binary image due to its structure, and grey-level image must be inverted to binary image before it can be processed. This kind of image processing method loses many useful information and lower system processing ability. In this paper, we propose a new parallel-distributed system model based on the ram weighted adaptive pattern recognition system, which can be used to process grey-level image directly. The new system can process high-resolution video with the rate of 25 frames per second, and can be applied broadly in robot vision, character recognition, security, and surveillance etc. realms.

2. BRIEF INTRODUCTION OF THE RAM WEIGHTED ADAPTIVE PATTERN RECOGNITION SYSTEM

The notional structure of the system is shown in Fig.1. Operation may be summarized as follows. The incoming image is a m_1 by m_2 binary image, and each pixel is represented by a single bit. A number 'n' of such bits taken from the image using a random but fixed mapping forms an n-tuple. A m_1 by m_2 binary image contains $m_1 \times m_2$ bits, therefore,

$$(m_1 \times m_2)/n = k$$

n-tuples are taken.

Conceptually, each n-tuple addresses a RAM element, which contains 2^n bits, and k RAM elements are needed. Each array of such RAMs is called a discriminator. In order to classify M samples, M discriminators may be provided simultaneously. These discriminators will be trained one by one and classify simultaneously

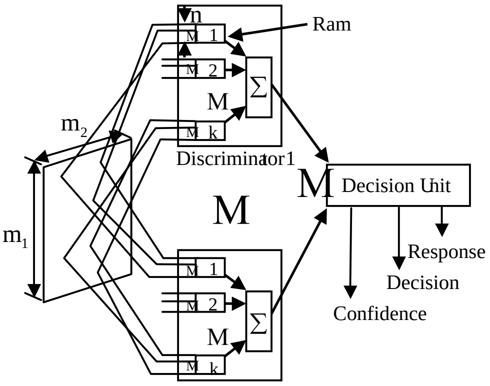


Fig.1 Notional structure of the system

The desired recognition activity is built up by a process of training on examples of the objects to be recognized. The system selects a particular discriminator to be trained and, assuming that all the stores in that discriminator is set to zero to start with, at the same time, the system uses the idle RAMs in the other discriminators as counter to record the occurrence frequency of each n-tuple. After all samples for this discriminator have been trained, ones are entered in all the locations addressed by that n-tuples whose occurrence frequency greater than the threshold we selected. Then, using the same method, we trained the other discriminators with else class samples. This has the effect of causing that particular discriminator to produce a logical one if the training image is presented again. If the input image is slightly changed, not all the RAMs within the selected discriminator will respond with a one. The ratio of the number of RAMs responding with a one to the total number of RAMs in a discriminator is called the response of that discriminator. These responses are fed into a simple calculator which first identifies which of the discriminators has the strongest response and outputs the class number associated with that discriminator. It also provides the actual response of the discriminator as well as the difference between the response of the strongest discriminator and the next one in line. This difference provides a measure of confidence with which the decision is made.

3. A PARALLEL DISTRIBUTED SYSTEM MODEL THAT CAN BE USED TO PROCESS GREY-LEVEL IMAGE

3.1 System Structure

The system model described in earlier parts of this paper can only recognize and process binary images, but we can improve it to process grey-level images. The new system model is shown in Fig.2.

For a 2^B -grey-level image, which has B bits per pixel, we can slice it easily to B bit-planes, each of which uses one window and one discriminator. B independent windows scan the B bit-planes simultaneously. Therefore, at any arbitrary position, we get B responses from B independent discriminators. According to the weighted sum of these responses, the decision will be made.

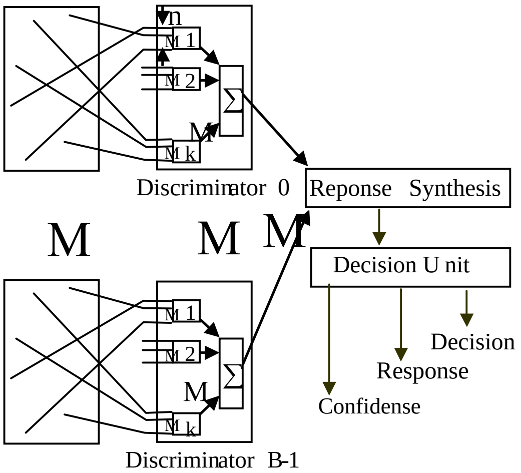


Fig.2 notional structure of new parallel distributed processing system

3.2 The Description and Analysis of the Model in Theory

It is assumed that B independent windows are used each covering R bits in a one-bit plane, and these windows scan B bit-planes simultaneously in pixel-wide jumps. The input pattern of an arbitrary window labeled W can be signified by:

$$(p_1, p_2, \Lambda\ , p_R)$$
$$where\ \ p_i = 0\ \ or\ \ 1 (1 \le i \le R)$$

The number of n-tuples is K=R/n

Assuming that the training set of the discriminator for the window W is

$$T = \{T_1, T_2, \Lambda\ , T_M\}$$

where M is number of training patterns.

T_i is an arbitrary training pattern of the set. It can be labeled

$$(p_1^i, p_2^i, \Lambda\ , p_R^i) \qquad (1 \le \mathrm{i} \le \mathrm{M})$$

Each T_i implies a state for each n-tuple, these states being labeled

$$\{S_1^i, S_2^i, \Lambda\ , S_K^i\}$$

where $S_{\mathrm{j}}^i (1 \le \mathrm{j} \le K)$ represents the state of n-tuple j.

Note that, for each n-tuple h, there exists a set of training states

$$S_h = \{S_h^1, S_h^2, \Lambda\ , S_h^\alpha\} \quad where\ \ 1 \le \mathrm{h} \le \mathrm{k},\ \alpha \le \mathrm{M}$$

The occurrence frequency of each state labeled

$$C_h = \left(\mathrm{C_h^1}, \mathrm{C_h^2}, \Lambda\ , \mathrm{C_h^\alpha}\right)$$

Let the threshold be σ, the valid set of training states be $S_h{}'$, and $S_h^\beta\ (1 \le \beta \le \alpha)$ be an arbitrary state in S_h. $S_h^\beta \in S_h{}'$ if $C_h^\beta \ge \sigma$. It is obviously that $S_h{}' \subseteq S_h$.

Let an arbitrary test pattern be U. The bit-plane j for the window W can be labeled

$$(p_{j1}, p_{j2}, \Lambda\ , p_{jR})$$

The set of n-tuple states is

$$\{S_{j1}, S_{j2}, \Lambda\ , S_{jk}\}$$

The responses set for the bit-plane is

$$(A_{j1}, A_{j2}, \Lambda\ , A_{jk})$$

$$\text{where } \mathrm{A_{jh}} = \begin{cases} 1 \text{ when } S_{jh} \in S_h{}' \\ 0 \text{ when } S_{jh} \notin S_h{}' \end{cases} \quad (1 \le \mathrm{h} \le \mathrm{k})$$

thus the terminal response for bit-plane is

$$A^j = \sum_{i=1}^{k} A_{ji}$$

Of course, the importance of each bit-plane is different. The responses of these bit-planes should be weighted in order to enlarge the difference of these responses for different input patterns. Using this method, the terminal response for the test pattern U is

$$A^U = \sum_{j=0}^{B-1} \omega_j \cdot (\sum_{i=1}^{k} A_{ji})$$

where ω_j is weight coefficient.

When U and training pattern is identical, the response of each discriminator will be k, and the terminal response is

$$A_{\max}^U = K \sum_{j=0}^{B-1} \omega_j$$

This response is the largest, and the terminal relative response is

$$R^U = \sum_{j=0}^{B-1} \omega_j \cdot (\sum_{i=1}^{k} A_{ji}) / K \sum_{j=0}^{B-1} \omega_j \tag{1}$$

R^u can be forecasted in theory. In general, each discriminator will be trained the same patterns. The calculator of R^u may be described as follows.

(1) Single training pattern

Consider first the case where there is only one pattern T_1 in the training set, and its states set of n-tuples is S.

Assuming that U is the test pattern, and the n-tuple states set of bit-plane j is S^j, then the response A^j is the number of the identical states of n-tuples for S and S^j indeed when $\sigma=1$.

If T_1 and bit-plate j have a proportion D_{j1} of the total number of pixels in the window in common, that is, the number of such pixels is $D_{j1}(0\leq D_{j1}\leq 1)$, we call D_{j1} the relative overlap similarity between T_1 and bit-plane j . For a given randomly sampled n-tuple, the probability of all n samples being in the overlap area is D_{j1}^n (assuming independent sampling).

So, $A^j = KD_{j1}{}^n$, and the relative response is

$$R^U = (\sum_{j=0}^{B-1}\omega_j \cdot D_{j1}{}^n)/\sum_{j=0}^{B-1}\omega_j \tag{2}$$

(2) Multiple training patterns
It is assumed that the set of training patterns is

$$T = \{T_1, T_2, \Lambda\ .T_M\},\ and\ \sigma = 1$$

We define the relative overlap area between $T_i(1\leq i\leq M)$ in T and the bit-plane $\mathrm{j}(0\leq \mathrm{j}\leq \mathrm{B}-1)$ of the unknown test pattern U as D_{ji}. As before, the probability of sitting the n-tuple entirely in the common area between T_i and bit-plane j is $D_{ji}{}^n$.

However, in this case we must avoid double counting of overlaps between various areas. To this end we define

$$\left(D_{jp}\cdot D_{jq}\cdot D_{jr}\Lambda\right)$$

as the common relative overlap area between bit-plane j and T_p and T_q and T_r Λ As before the probability of sitting an entire n-tuple in the common area between bit-plane j and T_p and T_q and T_r Λ is

$$\left(D_{jp}\cdot D_{jq}\cdot D_{jr}\Lambda\right)^N$$

and hence the most likely relative response to U is

$$R^U = \left(\sum_{j=0}^{B-1}\omega_j\cdot\sum_{i=1}^{M}(-1)^{i+1}\cdot P_{ji}\right)/\sum_{j=0}^{B-1}\omega_j$$

where

$$P_{ji} = \sum\left(\prod_{k=1}^{i}D_{ja_k}\right)^N \forall\{a_1, a_2, \Lambda\ a_i | 1\leq a_x\leq M, 1\leq x\leq i\}$$

3.3 Grey-level Image Processing Based on the System Mode

3.3.1 Analysis in Theory

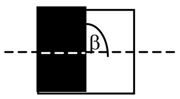
Here we take the edge detection of grey-level image as an example to introduce the application based on the system model.

The bit-planes containing more significant bits is more important than the others containing less significant bits, hence let the weight coefficient be

$$\omega_j = 2^j \quad (0\leq j\leq B-1)$$

According to the Eqn.1, we can get the nether formula

$$R^U = \left(\sum_{j=0}^{B-1}2^j\cdot\sum_{i=1}^{k}A_{ji}\right)/k\left(2^B-1\right)$$



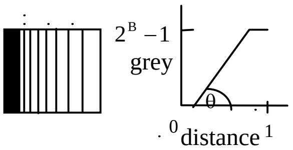


Fig 3 training pattern Fig 4 test pattern

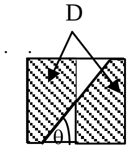


Fig.5 overlap area of training pattern and test pattern

To illustrate the effect, consider the simple training pattern shown in Fig.3 and the test pattern shown in Fig.4. Note that, all discriminators for each bit-plane are trained the same training pattern. Fig.5 indicates that the relative overlap area D_{j1} varies with the intensity gradient θ. When $\theta = 0°$ (all pixels in test pattern have the same grey-level, that is there is no edge in the pattern), D_{j1} may be shown to be $D_{j1} = 0.5\ (0\leq j\leq B-1)$. Hence, according to Eqn.2, the relative response $R^U_{\theta=0°}$ is

$$R^U_{\theta=0°} = (0.5)^N$$

When $\theta = 90°$ (the test pattern and the training pattern is the same one), D_{j1} may be shown to be $D_{j1} = 1\ (0\leq j\leq B-1)$. Hence $R^U_{\theta=90°} = 1$, and

$$P^N_{90°/0°} = R^U_{\theta=90°}/R^U_{\theta=0°} = 2^N$$

When $\theta = 45°$, the relative overlap D_{j1} is

$$D_{j1} = \begin{cases}0.5 & \text{when } 0\leq \mathrm{j}\leq \mathrm{B}-2\\ 1 & \text{when } j = B-1\end{cases}$$

Hence

$$R^U_{\theta=45°} = \left[\sum_{j=0}^{B-2}2^j\cdot(0.5)^N + 2^{B-1}\right]/(2^B-1)$$
$$= \left[(0.5)^N(2^{B-1}-1)+2^{B-1}\right]/(2^B-1)$$
$$> (0.5)^N(0.5-2^{-B})+0.5$$

and

$$P^{U}_{45°/0°} = R^{U}_{\theta=45°} / R^{U}_{\theta=0°}$$
$$> 0.5 + 2^{N-1} - 0.5^{B} > 2^{N-1}$$

Using the same calculation method, it is not difficult to draw a conclusion

$$P^{N}_{\theta/0°} > 1 \ (\theta > 0°)$$

which implicates that the response at $\theta > 0°$ (there is an edge in test pattern) is greater than the response at $\theta = 0°$ (no edge in test pattern). Hence, a threshold can be selected to detect the edges with different intensity gradient θ.

In actual application，in order to detect the edges with different direction, β shown in Fig.3 may be changed, and we may select eight training patterns

($\beta = 45° \times j \ (0 \le \mathrm{j} \le 7)$) or sixteen training patterns($\beta = 22.5° \times j \ (0 \le \mathrm{j} \le 15)$). Note that, if there are too many training patterns, let the Ram set threshold be $\sigma > 1$.

3.3.2 Experiment Results

The experiment has been carried out using emulation package for new system model. One input image is a four-grey-level image shown in Fig.6 (a), and the other is the same image with noise shown in Fig.6 (b).

The result image is shown in Fig.8. Note that, two discriminator have been trained the same sixteen patterns ($\beta = 22.5° \times j \ (0 \le \mathrm{j} \le 15)$), and the threshold value is 0.12. Fig.7 shows the result produced by the sobel operator in order to compare different methods. The result implicates that the system can detect the edges with different directions and different intensity gradient θ. In fact, the edges are marginally clearer and more precise, and the result image contains less noise.

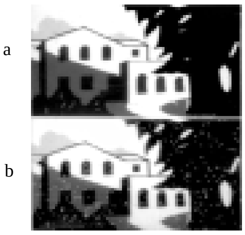


Fig 6 input image

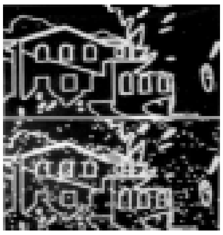


Fig 7 sobel operator on input image

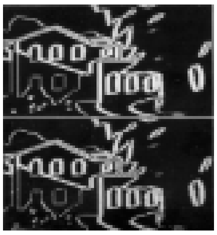


Fig 8 new system applied to input image

$R = 8 \times 8 \quad N = 4 \quad \sigma = 2$

4. CONCLUSION

The parallel-distributed system model proposed in this paper can be used for grey-level image processing, including edge detection, texture analysis and image smoothing. Of course，in this paper, we have only given the example of edge detection, and other related contents will not be introduced here. In fact, the system can be effected with conventional RAM chip, and with the RAM chip price declining continuously, this kind of system will even have the foreground in industry.

5. REFERENCES

[1] Ge Hongwei, RAM Weighted Adaptive Pattern Recognition System, Journal of Wuxi University of Light Industry, Vol.18, No.1，March 1999，pp82-86 (in Chinese)

[2] I.Aleksander, H.Morton, An Introduction to Neural Computing, Chapman and Hall, 1992, pp.70-90

[3] Dong YuNing, An Efficient Algorithm for Parallel Image Processing with Variant Templates, Chinese Journal of Computers, Vol.26, No.3, April 2003, pp.332-339 (in Chinese)

Hongwei Ge is a vice professor of School of Information Technology, Southern Yangtze University. He graduated from Nanjing University of Aeronautics and Astronautics in 1989. In 1992, he graduated from the same University and acquired Master's degree with specialty of compute science and technology.. He has brought to success more than 10 scientific research programs since 1992 and published one book, over 30 papers. His current research interests are in distributed and parallel computing, pattern recognition, and network security.

Wenbo Xu is a full professor and dean of School of Information Technology, Southern Yangtze University. He graduated from Tsinghua University in 1968. In 1981, he graduated from Tsinghua University and acquired Master's degree with specialty of theoretical electrotechnics. He was a visiting scholar and a visiting professor of University of Toronto in 1987 and in 2000, respectively. He is one of the DCABES international conference founder, was the chairman of DCABES 2002. He has brought to success more than 30 scientific research programs since 1990 and published over 100 papers. His current research interests are in distributed and parallel computing, intelligent computation, quantum computation and computational finance.

Approach of Feature Extracting Based on Single Sample

Xu Dongping, Chen Jingliang
Science and Technology of Computer, Wuhan University of Technology
Wuhan 430063, China
Email: DPXU@public.wh.hb.cn **Tel.:** +86(0) 27-86551167

ABSTRACT

It is still an open problem to extract feature from image, although there had been a lot efforts put in this area. This paper describes the general thinking of feature extraction from image, reviews the general approaches and the state-of-the-art of the method of feature extraction from image, presents a new method of manual feature extraction from image based on single sample image to deal with the same type of images, analyzes the model of the method, the particular processes of realizing and the status of application. Finally, exploratory research is recommended for future studies.

Keywords: Feature extraction; Image; Transform coding; Modeling of method.

1. THE WEIGHTTINESS AND THE GENERAL THINKING OF FEATURE EXTRACTION FROM IMAGE

In the pattern recognition system, feature extraction from image is a series of processes between source data and segregator that allow source data to be compressed from many dimensions to specific dimensions. Feature extraction can wipe off the information that we are not interested in, so it can optimize the impression of an image. For example, we can describe the feature extraction in the following form: *T : ER*－>*ED*. *ER* means higher dimensional space. *ED* means characteristic space. *T* is a specific mapping.

Feature extraction from image is a very important research field and it will exert a profound influence on many subjects, such as computer network, artificial intelligence[1], computer graphics[2], pattern recognition[3], and topography. It is especially important to computer network, because we can compress the image by extracting feature. The compressed image has few parameters and it is very useful to the limited bandwidth network. Many academic organizations have achieved successes in the area such as American McKeown laboratory, Bonn university of German, geographical institute of France.

2. THE STATE-OF-THE-ART OF THE IMAGE FEATURE EXTRACTION METHODS

Nowadays automatic methods of feature extraction from image based on nervous network have many difficulties that can’t be solved. What is more, there are a lot of shortcomings in the application of nervous network. For example, nervous network needs a great deal of experimental data, and it will take a long time to educate the network, and nervous network is short of interaction. All of these show that recognition is very difficult to computer, but to human recognition is so easy. So this paper presents a new method which is called manual feature extraction from image based on single sample[4]. In this method computer’s task is measure and recognition belongs to human.

3. RESEARCH OF METHOD OF MANUAL FEATURE EXTRACTION FROM IMAGE BASED ON SINGLE SAMPLE

3.1 FEATURE EXTRACTION

How to extract feature from an image? We can solve the problem by filtering waves.

The main idea is same as the reception of a radio. In geospace there are many radio waves. Every wave has it’s own frequency and swing. When we watch TV, we can get the interesting program from the compound waves by TV set. When we listen the radio, we also can get the program from the compound waves by radio. The compound waves can be explained by a visual example.

Another example of feature extraction from image is to observe a mountain. To a mountain, we can look it as a basic shape and a series of small shapes. All of these shapes compound the mountain, so to the different shapes we can sort them as figure 1.

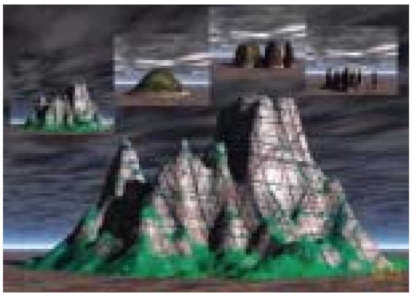


Figure 1 Decomposing Mountain

To an image, we can transform it to compound waves and sort them by frequency and swing. Choosing the waves we are interested in, we can get the feature we are interested in.

3.2 MODEL OF MANUAL FEATURE EXTRACTION FROM IIMAGE BASED ON SINGLE SAMPLE

The expression space of information is unorderly, so it’s difficult to separate the feature that we are interested in. In order to solve the problem, we have to transform the expression space to frequency space. Sorting and disposing the frequency, we can get the feature we are interested in. Figure 2 is the model of manual feature extraction from image based on single sample.

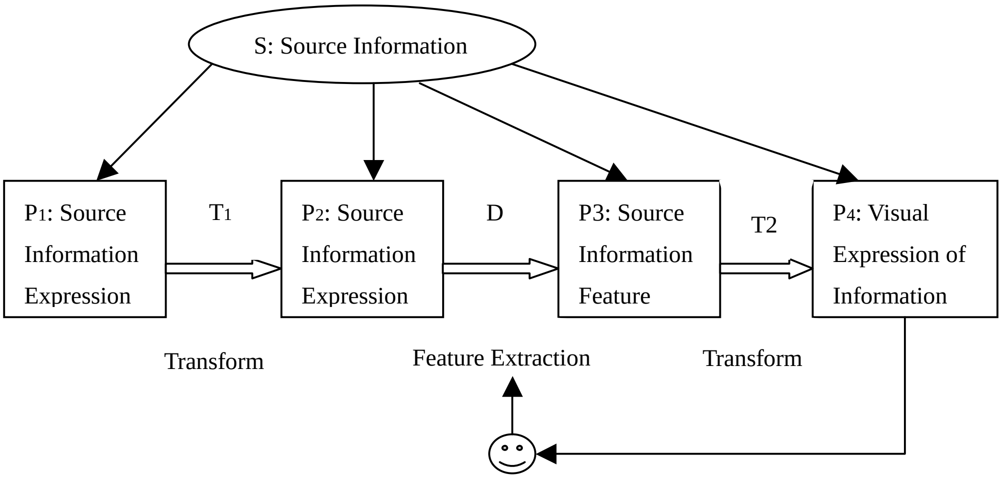

S: Source Information
P1: Source Information Expression
T1
P2: Source Information Expression
D
P3: Source Information Feature
T2
P4: Visual Expression of Information
Transform
Feature Extraction
Transform

Figure 2 Model of manual feature extraction from image based on single sample

We only need single sample in the processes of feature extraction. The processes of feature extraction are transformation. If we established the model of transformation, to the same kind of images, the same model can be used to extract feature.

For example, to the problem of vector graph, S is the content that we want to express; P1 is the expression space of the image; T1 is mathematical transformation; P2 is frequency space expression; D is feature extraction of frequency information; P3 is the vector expression of frequency information; T2 is transformation from frequency space to original space; P4 is raster expression of vector. In this model, we can change D continuously, until we are satisfied with the result and recode the parameters within all processes.

3.3 TRANSFORM CODING

The most important step is to establish the model of transformation.

The processes of transform coding are the processes of compression. Using linear transformation, we can map the source data to characteristic data. The characteristic data are the compressed code because we have wiped off the information, which we are not interested in.

There are many methods of transform coding, such as Karhunen_Loeve Transform(KLT), Discrete Fourier Transform(DFT), Walsh_Hadamard Transform(WHT), Discrete Cosine Transform(DCT), etc. All of these are orthogonal transformation.

Practically in order to establish the model of transformation we need to choose one kind of transform coding or compound some kinds of transform coding based on single sample. It's necessary to analyze these transform coding which are in common use.

KLT is the best method of transformation, and it's very important in data transformation area. The transform matrix of KLT is not fixed, and it changes with the source data, so it will take a long time to calculate.

DFT can't wipe off the relativity of information entirely, but its transform matrix is fixed. DFT is quicker than KLT.

WHT is orthogonal transformation with the parameter 1 and –1,so the calculation only includes addition and subtraction. WHT is the quickest.

DCT is the standard of JPEG(Joint Photographic Expert Group), MPEG(Motion Picture Expert Group) and the M.261 of CCITT, and it is in common use too.

The raster of image can be expressed with a space function, and it's general form is

$z=f(x, y)$, or

$f = [f_0, f_1, f_2, ... f_m]^t$, $f_i = [f_{i0}, f_{i1}, f_{i2}, ... f_{im}]$,

$i = 0,1,2,3, ... m$.

It's difficult to show the characteristic of an image, because *f* is orderless. We have to transform *f* to another form. If we call the transformation *T*, *T* should be:

(1) *T* is linear;
(2) *T* is orthogonal;
(3) *T* shown the transformation of frequency;
(4) *T* can be calculated easily.

We can transform *f* to frequency space, and it's general form is $F(u,v) = TfT^t$,

$F = [F_0, F_1, F_2, ... F_m]^t, F_i = [F_{i0}, F_{i1}, F_{i2}, ... F_{im}]$,

$i = 0,1,2,3, ... m$.

$T = [T_0, T_1, T_2, ... T_m]^t, T_i = [T_{i0}, T_{i1}, T_{i2}, ... T_{im}]$,

$i = 0,1,2,3, ... m$.

To index F with frequency, we can descript F as:
$F=[F_{q(k)}]$, $k=0, 1, \dots, q(k)$ is frequency.
For example, the transformation is $T=H_k$, and the general form of H_k is:

$$H_k = \begin{bmatrix} H_{k-1}, H_{k-1} \\ H_{k-1}, -H_{k-1} \end{bmatrix}$$

$[H_0]=1$, $k = 2^p$, where *p* is a positive integer. H_k is a

$2^p \times 2^p$ matrix, and the sign of each row is different. We can change the rows, in order to sort the transformation degree of each row. Subsequently we will get the matrix *T*. All of these are the processes of Walsh_Hadamard Transform. We can adopt the other methods of transform coding too, such as DCT, DFT, and KLT.

F is the expression of frequency space. We can choose some frequencies continuously, until we are satisfied with the result. Once we establish the model of transformation, the same kind of image can use the model, which is the general thinking of manual feature extraction from image based on single sample.

4. THE REALIZING OF MANUAL FEATURE EXTRACTION FROM IMAGE BASED ON SINGLE SAMPLE

There are three kinds of parts of feature extraction from image. The first is surface feature extraction; the second is line feature extraction; the third is spot feature extraction. For example, sometimes we need to extract the feature of people’s fingerprint, which is line feature extraction.

The high frequency signals express the line feature of image in the frequency space. In order to extract the line feature, we need to get high frequency signals and screen low frequency signals. The particular processes of realizing are the following four steps:

(1)For one of the same type of images, select the model of transformation.
(2)In the frequency space, select the frequency signals that should be screened or attenuated.
(3)Adjust the value of attenuation and scheme.
(4)Get the scope of pixel value that should be recognized.

In this example, the model of transformation is WHT; the frequency signals which should be attenuated are *Fq(0)* and *Fq(1)*; the value of attenuation is 50 and the scheme is linear; the scope of pixel value which should be recognized is above 100. $F_{f(k)}$ is frequency space and the *n* is the value of attenuation noting in Figure 3. We can extract the feature from other same type of image as above single one with the same parameters.

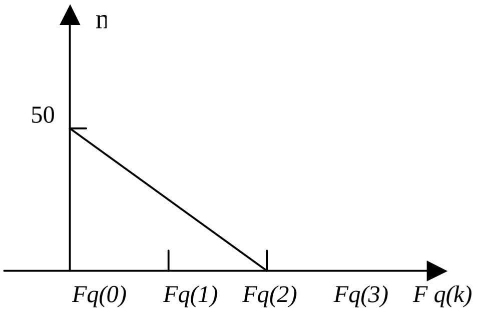


Figure 3 function chart of F and n

The result of realizing refers to Figure 4 and Figure 5

5. PROSPECTS OF FEATURE EXTRACTION

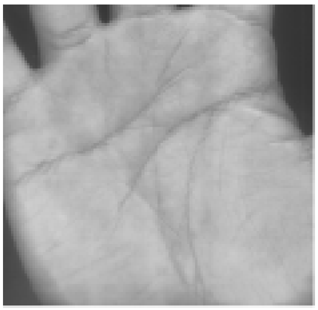


Figure 4 Image of handprint

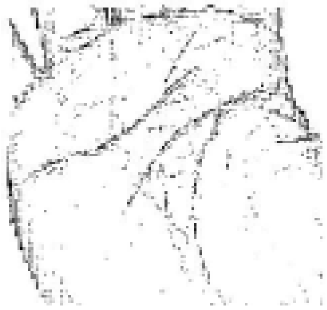


Figure 5 Feature of handprint image

FROM IMAGE

Nowadays it is still an open problem to extract feature from image, although there had been a lot efforts put in this area. The method which this paper presented is very suitable and exact to feature extraction. With the development of computer and the extend of research, extract feature will be wildly used in a good many areas.

6. REFERENCES

[1] TRINDER JC, WANG YD, SOWMYAA, etal. Artificial Intelligence in 3D Feature Extraction [A].Automatic Extraction of Man-made objects from Aerial and Space Images (2)[C].Basel: Birkhaeuser Verlag, 1997, 257-265.

[2] YASSIN M Y, KARAM L J.Morphological Reversible Contour RePresentation[J].IEEE Transform on Pattern Analysis and Machine Intelligence, 2000, 22(3):227-239.

[3] NEVATIAN，BABUKR.LinearFeatureExtraction and Description[J].Computer Graphics and image Processing, 1980, 13(2):257-269.

[4] Xu Dongping. Real Time Dynamic Interactive Visual Simulation: [Ph.D. thesis]. Wuhan University of Technology,2001.

[5] Samal A, Iyengar P A.Automatic Recognition and Analysis for Human faces and Facial Expressions:A SurVey[J].pattern Recognition，1992,25(1):65-77.

[6] Hong Ziquan . Algebraic Feature Extraction of Image for Recognition[J].Pattern Recognition, 1991, 24(3):211-219.

The Approach of Extracting the Graphs from Images of 3-D Objects

Tao Hongjiu Wen Yujuan Tong Xiaojun
Wuhan Polytechnic University, Wuhan 430023, P R China.
Wuhan University of Technology, Wuhan 430070, P.R.China.
E-mail: thjll@263.net

ABSTRACT

This paper describes a method of extracting such graphs from images of 3-D objects. There are few accounts of practical methods for extracting any sort of high level feature from images and this is an important gap in the computer vision and image processing literature. The use of the same feature extraction methods on real and synthetic images requires the image synthesis to be realistic. The use of structural features in matching seems to be extremely flexible from the point of view of software development and the ability to merge data from the multi-sensor sources, illustrated here by colour and range images.

Key words: 3-D, Rang images, Extracting, Computer verision.

1. INSTRUCTION

High level features carry information about an image in an abstracted or propositional form, [1]. The use of high level features is appropriate in model based computer vision, as illustrated in figure 1, where the matching is applied to attributed relational graphs which are widely used as a flexible high level feature.

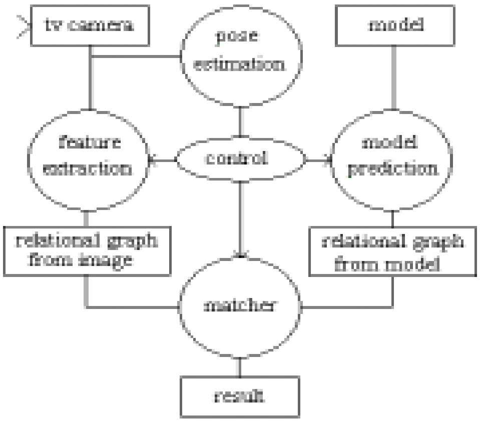


Figure 1 Computer vision using structural methods

This paper describes a method of extracting such graphs from images of 3-D objects. There are few accounts of practical methods for extracting any sort of high level feature from images and this is an important gap in the computer vision and image processing literature. Examining figure 1, real world data is processed so as to create an attributed relational graph describing the image of an object. The reference representation was computed off-line from a geometric or ray-tracing model. This predicts the appearance of a 2-D projection of the real world as a graph with the same conventions as that created from the real world data.

A graph matching procedure is used to associate fragments of the two graphs and obtain a cost. A control strategy selects candidate classes and presents fragments of reference graphs for matching. It also monitors system performance and allocates system resources to more promising candidates, [2]. This paper is limited to describing the extraction of the relational graphs.

2. CONSTRACTION OF ATTRIBUTED RELATIONEL GRAPFS

2.1 Overview

The following describes the method that is in use for constructing relational graphs from colour and range image data. The relational graphs are built from nodes, representing the contours of regions, connected by arcs representing the relative positions of the contours in the image's 2-D coordinate system. This basic graph structure has attributes attached to it creating a flexible and extensible propositional representation.

2.2 Steps In The Generation Of The Relational Graphs

(1) Separation of regions

The input image(s) are processed to identify and label connected regions. The standard definition of a connected region: '...set of pixels that are connected under some definition and share some common property...' is employed. 4-connectedness is required, and is taken as an indication of adequate sampling, [3]. The examples presented here employ coloured images registered (to within 1 pixel) with range images, forming a 4 component vector valued pixel image.

In the present implementation, the colour image has the dominant role in defining regions. The 'common property' is a uniformity of colour ratio, r:g:b, while the total intensity remains above a given threshold. Regions are extracted by subjecting the 3-component colour images to systematic averaging in small patches to reduce variance without crossing the 'walls' of high variance between regions.

The images have been segmented and figure 4 shows the regions identified in the images which are used as masks to control the extraction of data to complete the graph.

This work is supported by Natural Science Foundation under Grant 2002 AB041 of Hubei Province and Foundation of Hubei Provincial Department of Education under Grant 2004Q001, China.

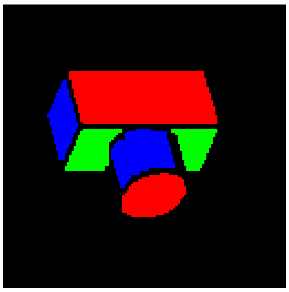


Figure 2 Shows a colour image of an object.

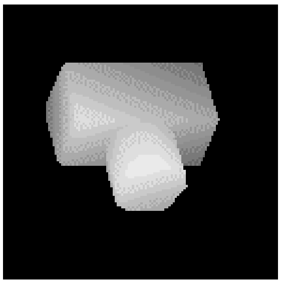


Figure 3 Shows a depth image that is registered with the image in figure 2.

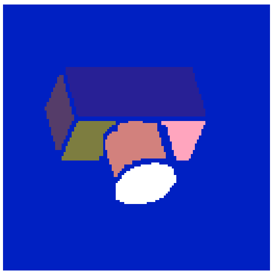


Figure 4 Shows the regions identified in the images

(2) Extracting the nodes
The next stage is the tracing of the contours of the regions such as those of figure 4.

The output is a list of coordinates of the 8-connected contour pixels. The contour trace defines pixels in the image which are used to calculate a number of numerical and logical attributes for each contour.

The attributes extracted in the present implementation are:

(a) Centroid, cg(x,y);
The centroid of the contour is calculated. This is expressed by the proposition cg(xbar,ybar). The centroid is used as the 'handle' by which to describe the position of the contour and region.

(b) Area, aa(area);
The area enclosed by each external contour, excluding areas enclosed by any internal contours, is computed. The area is the pixel count in the image. The proposition is aa(area).

(c) Moments, mm(m1,m2,m3,m4);
The phenomenon of rotation and scale invariant combinations of body centred moments is well known. The first 4 such combinations are calculated from the coordinates of the contour pixels and are expressed as the proposition mm(m1,m2,m3,m4). The values are scaled and truncated to become large integers for ease of transmission and processing. The use of the rotation and scale invariant combinations rather that the raw moments does not discard size and orientation information which is available from the area and centroid attributes. It is convenient to use the moment attribute purely to represent shape.

(d) Mean colour values, cr(r,g,b);
The mean colour values found under the track of the contour trace are computed yielding values for red, green and blue. The regions were defined based on a uniformity of colour ratio. The proposition is cr(r,g,b). (e) Plane equation coefficients, pl(a,b,c,d);

If the contour belongs to a region that is approximately planar on the 3-D object, then the equation of the most closely fitting plane, in the form ax+by+cz+d=0 is useful and can be computed from the range image. The calculation chooses (a,b,c,d) to minimise the sum of absolute distances between the plane and the points on the contour. If no acceptable fit is possible the contour is assumed to be non-planar and no attribute is given otherwise the propositional form is pl(a,b,c,d).

The a,b,c,d values are based on a coordinate system after perspective projection because of the viewing conditions. Work must be done to recover a plane equation valid in the real world coordinate system.

(f) Limited shape classification;
The 4 rotation and scale invariant body centred moment combinations are used to form a minimum distance classification of the contour against the three references of a circle, square and triangle. This crude classification can save time in further feature processing. The propositions are sqr or cir or tri.

(g) External/internal contour flag;
The discovery of contours in the image allows a determination of whether they are internal or external. This property is indicated with the proposition ext or int.

The image is a projection onto 2-D of the original 3-D object via the optics of the camera and the quantisations implicit in the image capture device. The attributes are calculated using the image coordinate system. The known problems of camera imperfections, [4], are not dealt with in the present implementation.

(3) Creating the arcs
The arcs of the relational graph are labelled with codes that show the geometrical relation between the contours.

(a) Creating arc labels
The existence of an arc between two nodes implies that they are within a certain threshold distance (between centroids) of each other. Without this filtering, the graph would be totally connected.

In the present implementation, the label on each arc is one of 16 direction codes. These correspond to compass headings North, North by North East, North East, North East by East etc..

An arc can also carry one of the labels 'outside' or 'inside'. These, if present, show that there is a significant intrusion of one contour into the bounding rectangle of the other.

(b)Creating arc attributes
Attributes on the arcs indicate the importance of including the arcs in the matching. They are used for scheduling matching attempts, [2], as part of a technique to partition the matching and monitor its progress [5].

2.3 Algorithms

The block with a cylinder mounted on it, figures 2 and 3, will continue to be used as a running example. The creation of a relational graph as a reference can begin either with a geometric model such as a CAD/CAM model, or can employ ray tracing, [6]. The use of a CAD/CAM model requires issues of surface visibility and colour to be handled explicitly. In the examples here, a z-buffer algorithm is used to determine visibility and colour values are appended after all geometric processing has been completed. When ray tracing is employed, these issues are handled implicitly after the 'studio definition' has been given. Whatever the synthesis, the resulting image is processed by the same feature extraction software that is applied to real image data. The sequence of steps in creating the attributed relational graph is (1) capture or synthesize an image, including a depth image; (2) Grow regions in the colour image and extract region masks; (3) Construct the relational graph using contours in the region mask to provide the graph's nodes.

3. BUILDING A CONTROL STRUCTION FOR MATCHING

One of the main motives for studying structural features is to automate the building of a control strategy, shown in figure 1, for the matching, [7], [8].

4. CONCLUSIONS

Structural features seem to provide a robust and concise description of objects. The use of the same feature extraction methods on real and synthetic images requires the image synthesis to be realistic. The use of structural features in matching seems to be extremely flexible from the point of view of software development and the ability to merge data from the multi-sensor sources, illustrated here by colour and range images.

5. REFERENCES

[1]. F. van der Heijden, Image Based Measurement Systems, Wiley, 1994.

[2]. R. E. Blake, Development of an Incremental Graph Matching Device, Pattern Recognition Theory and Applications, NATO ASI Series, Vol 30, pp355-366, Springer, Berlin, 1987.

[3]. T. Pavlidis, Algorithms for Graphics and Image Processing, Computer Science Press, 1982.

[4]. N. Thune, Stereo Vision: An Integrated Approach, dr.ing. thesis, University of Trondheim, 1991.

[5]. R. E. Blake, Partitioning Graph Matching with Constraints, Pattern Recognition, Vol 27, No.3, pp 439-446, March 1994.

[6]. J. Foley, A. van Dam, S. Feiner and J. Hughes, Computer Graphics: Principles and Practice, Second Edition, Addison-Wesley, Reading, MA., 1990.

[7]. R. E. Blake and P. Boros, The extraction of structural features for use in computer vision. Proceedings of the Second Asian Conference on Computer Vision, Singapore, December 1995.

[8]. P. Boros and R. E. Blake, The calculation of the aspect graph from a CAD model. Proceedings of the Second Asian Conference on Computer Vision, Singapore, December 1995.

Tao Hongjiu is a full professor at the Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan, Hubei Province, China. He graduated from Tianjin University of Technology in 1980; and earned Engineering Doctor Degree in Pattern Recognition and Artificial Intelligence, at the Institute for Image Processing and Intelligent Control, from Huazhong University of Science and Technology in 2003, China. He has published one book, over 30 Journal papers. His research interests are in the computer image processing, pattern recognition, streaming media technology, digital signal processing etc.

A Distributed Video Proxy System Based on Cache

JIA JIONG, ZHU JIAN-XIN
Institute of Computer Science, Zhejiang Normal University
Jinhua, zhejiang,China 321004
Email: jia@mail.zjnu.net.cn Tel.: 0579-2283892

ABSTRACT

According to the current situation of video services on WWW, a distributed video proxy system based on cache is described in this article. The storage policy of the video files and the response of user's request are discussed in this system. Through the simulative experiments and result analysis on the campus local network, we got the most important characteristics of the system: the report of the total cache performance and the relative conclusion.

Keywords: Video, Cache, Distribution, Proxy.

1. INTRODUCTION

With the increasing demand of multimedia information in network, people begin to pay attention to the video stream in the Internet. To transmit the real time video, enough network width is need. With fantastic development of the Internet exploded property and improving the network width. It is possible to transmit video data on time under supported by the contract of operation (for example RTP, RTCP, SIP and RTSP). The application of the video stream in LAN is to coming forth. For example, there appears a great number of the ordering program in campus network. Because the application based on the video stream has many problems that impact the feature of the net, need amount of the bandwidth and induce easily the net congestion etc in internet, this paper puts forward a new distributed video proxy system based on cache and discusses its storage policy of video files and response for user's request. Through the simulative experiments and results of analysis on the campus local network, the report of the total cache performance is obtained.

The characteristics of storage policy of video files on the Internet are the high requested bandwidth and the video width is hard to change, therefore it is effective to use the cache. The documentary [1] indicates that it is need to sustain about 1Mbps in order to stream the video files stored on the web today. These bandwidth requirements make video files susceptible to Internet brownouts. By caching video files close to the clients, the system simultaneously mitigates the unreliability of the Internet, improves access latency, and reduces overall traffic. From point of view server, the server load is reduced validly in cache technology. From point of view network, the utilized rate of network bandwidth is improved validly.

The documentary [2] shows: (1) A clients is always scanning the beginning part of a video in order to ensure whether he likes it or not. If the video is valuable to the client, it will be watched, or else stopped. The research discovers that about 60% video ordered programs can be last finally and other 40% will be stopped at once. (2) Temporal Locality: LRU stack depth analysis of the traces shows that the video files appear the extreme temporal locality [3]. If a video has been accessed recently, it would be accessed again at once.

Through large amount of observation, the requests to a video server tend to exhibit locality of reference. Hence, it is possible to exploit the cache techniques that reduce redundant video accesses to the same server. Distributed video proxy system based on cache is a cache video service machine cooperated with work. This server is the proxy server troop, whose function is cache the video files on a well-connected network. Through cooperating the cache, the video server will own amount of assembly caches and every proxy server will stand the minimum load. For example, if each of 1000 clients on a college campus takes part in the proxy server cluster, and each client has 100MB cache, the total cache size would be 100GB, enough for caching a significant number of video files. But since each client rarely views more than one video file at a time, the typical load on any one client would be less than that requested to service one video file (about 1Mbit/sec).

2. THE SYSTEM DESIGN

2.1 System Component Configuration

Figure 1 shows how to use the cache technology in the system that contains of a remotely located web server and two machines with a high degree of connectivity. For example, the user is not connected to the remotely located web server directly, but a running local proxy server P_1 (step 1). The proxy server P_1 checks to see whether the local is cache or not. If not, the proxy server can be connected directly with the long-range server (step 2). Then it starts to receive the video. P_1 stores a copy of the video locally while simultaneously forwarding another copy to the client browser (step 3). Suppose a user need the same video to the second machine, P_1 checks whether local is cache or not (step 4). P_1 saves the copy, P_2 contacts P_1 (step 5). Accesses the video and sends it to local browser (step 6). Thus bypassing downloading the video from original web server.

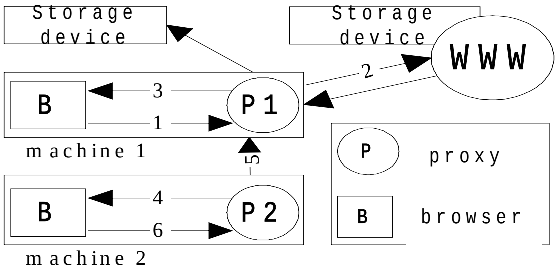


Figure 1: The architectonic of system based on cache

A coordinator process keeps the tracks of storage files in every proxy server and redirect requests accordingly (figure 2). In the previous example, the coordinator coordinates the work of P_1 and P_2. The coordinator also utilizes the proxies to manage the copied video files stored in every machine. If the system has no free spaces, the coordinator will decide which files to be canceled in order to make room.

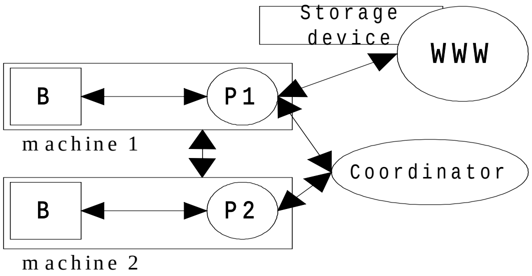


Figure 2: The role of the coordinator

The system contains two components: proxy server and coordinator. A typical configuration contains of a coordinator used in the local area network and some proxy servers. The context in the proxy server and the tactics of deciding to replace the cache in all system are cooperated by the coordinator. The proxy server can join directly to user or management video files.

Two different types of proxies are defined in the system: local and storage proxies. Local proxies are used to deal with user's request together with user in a same machine. Any datum cannot be stored in this proxy, which is similar to the component in the function.

It is the deal condition that every machine can be used in the partial proxy. In fact, this system construction is difficult to realize. Hence if only some selected machines are used to the partial proxy in network when imaginary, the every proxy offer service to the machine in the narrow range. On the other hand, storage proxies are served to the client's request directly and their function is to store data. Storage proxies may stay any places in local area network.

A collection of both types of proxies run in the LAN and organized by a single coordinator together from a proxy cluster. Figure 3 shows an example of this configuration that SP indicates storage proxy server, LP indicates local proxy server, C indicates the coordinator and B indicates the user browser.

The arrangement of system component has a number of advantages in the proxy cluster:

(1) Latency reduction.
(2) High aggregate storage space.
(3) Load reduction.
(4) Scalability.

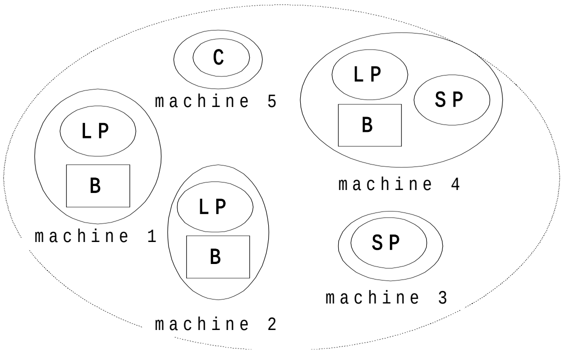


Figure 3: A proxy cluster

There are perhaps two disadvantages in this system architectonic maybe. First, since the local proxies consult the coordinator for every request, their central nature might make them a choke point of this system. But, the problem is beneath for paying too much attention that reached the video that request each other reach the time and the coordinator transmit careless the video data. Second, the coordinator is the most critical component in the total system. In other words, if the coordinator appears faults, it may induce the collapse of the whole system. One possible solution is the coordinator cluster, adding a redundancy coordinator. The redundancy coordinator retains similar condition to using the coordinator in normal times. When the coordinator is collapse, the redundancy coordinator can replace it to guarantee the system well.

2.2 Video Storage Policies

As stated above, the users become even more likely to view the beginning of a movie than to play it back until its end Unlike HTML files, an entire video document does not keep in cache in order to satisfy the need of users. According to this theory, the concept of the partial video cache is absorbed by the system. When the user requests a video file in the cache, the system will sent snippet of the cache to the user, while it will request remainder part and send transparently the to the user from the WWW server.

The video server and the streaming protocol must support to the random visit in order to ensure the partial video cache well. Fortunately, all major streaming protocols allow this.

The cache is divided into two same size blocks of files in the storage system, which allows the cached head to be spread across multiple storage proxies. Representing video as an order sequence of file blocks simplifies the system architectonic considerably.

A convenient mechanism is provided among multiple proxies for spreading portions of a single head. Thus the better load balancing, simplification of cache replacement and partial video realization are reached. If a new title T_1 would be brought into the cache, yet the entire system is full, the concept of the block makes it possible that the system is easy to replace the end portions of an unpopular title T_2 from the cache. Be adverted, the system is not instead of T_2 entirely, but a portion of it present in case it is requested in future. In the same way, the users can determine the size title T_1 in the cache.

The file blocks from multiple proxies can connect to continuous stream, which it might be due to switching decays into blocks. This can cause interruption in the data flow from the proxy to the user. Such switching delays can be eliminated by fetching data at a higher rate and double buffering against latencies.

2.3 Request Responses

The system example in the Figure 4 can be utilized to explain that the independent components are how to exchange each other in the system. The system contains a proxy cluster W and a WWW server in exterior local area network. The proxy cluster includes a cooperating server C and two partial proxies LP_1 and LP_2 that separately serves the client browse B_1 and B_2. There are two storage proxies SP_1 and SP_2 in the system. Server W stores the movie M, which is divided into two logic file blocks M_1 and M_2. How the system deal with the three different scenes (cache not hit, cache hit, request cancellation) will be explained as follow:

2.3.1 Cache Not hit
When B_1 requests the title M, and the request is intercepted by LP_1, the following events will occur:

1. LP_1 is connected to the coordinator C and the server W in the same time.
2. Because M is not cached by the systems, C gives a negative reply. W answers the head information of M.
3. From the header information of M, the characteristic of M is determined by LP_1. LP_1 requests C to obtain an enough cache to store M.
4. If there is enough space in cache, C will answer the block place. If there is inadequate space, C will decide who is the victim according to replace policy in cache and tell the place to LP_1. In this example, SP_1 is chosen.
5. LP_1 begins to receive M_1 (the first block of M) from W, which it then streams to both B_1 and SP_1.
6. When M_1 has been finished in its entirety, LP_1 requests the location of another block from C in order to locally cache M_2.
7. C returns SP_2 as the location for the next block. LP_1 continues receiving M_2 (the second block of M) from W, which it then streams to B_1 and SP_2.

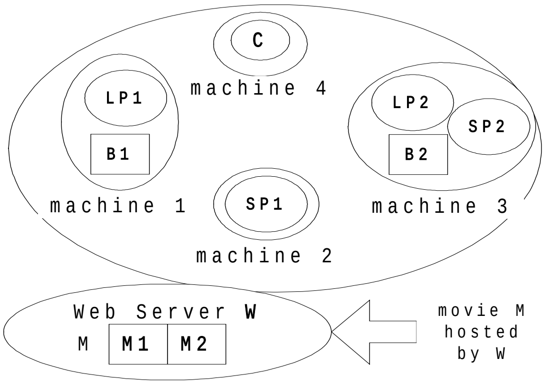


Figure 4: Proxy cluster used in example

To summarize, the movie that is not stored by the system cache is led into the system in divided blocks, according to forms of block to store. Between the proxy and coordinator ensured the annex communicate cost of the every block is the possible fault in the policy. However, compared with the cost of transmitting the total files, the cost is too trivial to mention.

2.3.2 Cache Hit
If M has been stored by the system, the following accident will happen when B_2 requests M.

1. The proxy LP_2 contacts simultaneously to the coordinator C and the server W.
2. C gives the positive reply and returns to the block pointer of the cache in SP_1.
3. LP_2 closes the joiner of W.
4. LP_2 contacts SP_1, and starts streaming from M_1 to B_2.
5. After M_1 has been finished, LP_2 contacts again to request the remainder part of M.
6. C replies and points the pointer to SP_2. LP_2 connects SP_2 meanwhile begins the stream from M_2 to B_2.

2.3.2 Request Cancellation
The users commence to playback of the head of a movie and decide to stop if they do not like it. Stopping the movie results in the browser canceling its local proxy connection while a cache hit or not hit type transaction is taking place.

If there is a cache hit, it is simple to cancel the request. When the proxy checks up the browser whether the message road is closed or not, the proxy only simply closes signal and requests other relevant connects, meanwhile it gives notice to the coordinator.

If there is not a cache hit, different treatment is utilized depending on weather other users in the system are also currently accessing the same file. If only a user B_1 cancels watching, at the moment the user has finished the second part of M. LP_1 checks up the canceled request and notices the coordinator. If no user visits the movie at the moment, the coordinator notices LP_1 to shut the surplus connects. The system processor cancels all of the adductions that point M_2, and M_1 are saved in the local cache.

A more complicate scenario is that B_1 and B_2 visit M simultaneously. When the movie broadcast the second part, B_1 decides to cancel the broadcast. In the same way, LP_1 checks the canceled request and notices the coordinator. Unlike that, B_2 continues to broadcast the movie. The coordinator notices LP_1 to obtain the files continuative, and the files are saved in the local storage proxy.

3. THE SYSTEM ANALYSIS

The system performance is influenced by the factors: the cache replacement policy, the amount of the proxy cluster, the size of the file blocks, all of the size of caches in the system and the mode of the user request etc. Other factors are variable in the system except the last factor. The cache replacement policy is the most critical factor because it affects not only the system performance, but also the load balance among the caches. The load balance is the most important, because the aim of the design system is to make the user machine take part in the cache as proxy. If the load of user machine is always high, it will not take part in.

Before running the simulate system, many system parameters are variable. Only two aspects are discussed here: all of the size of the proxy cache and the cache replacement policy.

3.1 The size of the proxy cache
The long-range video server holds 15.7G video’s files totally. To simply the comparison, there are three systems and every system includes 20 machines. In the first group, every machine has 12M caches. The size of the total cache is 20*12 or 240M. In the second group, every machine has 25M caches. The size of the total cache is 20*25 or 500M. In the third group, every machine has 50 M caches. The size of the total cache is 50*20 or 1.0G.

3.2 The Replacement Policies in cache
In the experiment, many cache replacement policies are used, for example LRU [4](Least Recently Used), LFU [4](Least Frequently Used), FIFO [4](First In First Out) and LRU-k [5]. LRU-k algorithm remains the record of every title in the cache before k times accesses. In the fixed time, for the visit title at present and at the k times, the title of the k-range is regarded as different. LRU-k algorithm chooses the last block of the largest k-range as the next victim. The disadvantage of using LRU algorithm in keeping the candidate within the system is

saved by choosing least recently used title as a victim. LRU algorithm may be regarded as the exception when k value of LRU-k algorithm is 1.

To compare the efficiency of these algorithms, the system adopts two annex policies: perfect and infinite. The forth is an ideal cache replacement mechanism. This mechanism utilizes the future knowledge to replace in the cache the unused block of the longest time future. If the system owns the limited space, this algorithm is the best. The infinite system hypothesis that owns the cache space is larger than the total video files. Running in this system, the highest rate is owned. Whatever it utilizes the cache replacement policy, the system performance is same because no data need to replace from the cache.

3.3 Comparison of the total cache performance

To report the total cache performance in the condition of various parameters and configurations in the system, a conception is introduced----byte hit ratio (BHR). The BHR measure the total caching performance. The BHR is defined as:

BHR=(total bytes served from the cache) / (total bytes read by all clients)

The more the BHR value approaches 1, the higher the system performance. Because most of the bytes requested by the users are served from the local cache.

Through the simulative experiments and result analysis on the campus local network, the following conclusion is obtained: (1) Larger global cache sizes increase the overall hit rate. (2) With the decrease of the cache size, the difference between the replacement policies and perfect becomes less pronounced. This indicates that the scarcity of the storage resource is an obstacle to raise performance, when the size of the cache is small.

5. CONCLUSION

The experimental results show that the LRU algorithm can provides higher cache-hit rates. The LRU-k algorithm can yield good hit rates as well as effective load balancing. To a 1GB video file, the storage space of 20*50M is relatively small. This resulted in high cache-hit rate. From the view of the server, the proxy system reduces load apparently by intercepting a lot of accesses. The effectiveness of the system network increases greatly available bandwidth in the entire video delivery system, and allows servers to serve more users in the same time. In addition, some problems such as the system performance is affected by the rate of system cache and the connecting proxy quantity, are well worth researching further.

6. REFERENCES

[1] S.Acharya, B.Smith. An Experiment To Characterize Videos On The World Wide Web. Proceedings of ACM/SPIE Multimedia Computing and Networking 1998 (MMCN'98), San Jose, Juan. 1998.

[2] S. Acharya. Techniques For Improving Multimedia Communication Over Wide Area Networks. Ph.D. Thesis, Department of Electrical Engineering, Cornell University, Juan. 1999.

[3] V. Almeida et al. Characterizing Reference Locality in the WWW. Technical Report TR-96-11, Department of Computer Science, Boston University,1996.

[4] A. Silberschatz, J. Peterson, P. Galvin. Operating System Concepts. Boston: Addsion Wesley, 1992.

[5] E. O'Neil, P. O'Neil, G. Weikum. The LRU-k Page Replacement Algorithm For Database Disk Buffering. Proc Of International Conf on Management of Data, May 1993

[6] YANGYu-Hai, BIN Xue-Lian, ZHENG Yu-Qiang, A Cooperative Web caching System Based on Hybrid Management, Journal of Computer Research and Development, Vol.40, No.5, May 2003, pp.757-762 (in Chinese)

Jia Jiong was born in 1965. He received the B. Sc. degree in computer science & technology from Zhejiang University in 1987. Received the M. S. degree in computer software & theory from Fudan University in 1995. He is now working as a vice-professor in Zhejiang Normal University. His research interesting includes Neural Networks, Fuzzy Logic, Evolution Computing, Parallel and Distributed Computing, Data Mining, etc.

Zhu Jian-Xin was born in 1972. He received the B. Sc. degree in computer science & technology from Zhejiang Normal University in 1991. Received the M. S. degree in computer software & theory from Zhejiang University in 1999. He is now working as a vice-professor in Zhejiang Normal University. His research interesting includes Parallel and Distribute Computing, Data Mining, etc.

The Building of Inexpensive Large-scale Storage System for Video Applications

DONG Xiao-ming*, XIE Chang-sheng
Key Laboratory of Data Storage System, School of Computer Science and Technology,
Huazhong University of Science and Technology
Wuhan, Hubei 430074, China
* Email: xmdong@tom.com Tel.: +86 (0) 27-87800265

ABSTRACT

Industry has begun to place pressure on the storage system, demanding that it store more and cheaper. To satisfy the requirement of inexpensive large-scale storage for enterprise video applications, we approved two cost-effective designs: NAS cluster and iSCSI based IP SAN. We discussed related design options such as storage medium, networking architectures and storage management issues.

Keywords: disk drive; cluster; iSCSI; storage network; video

1. INTRODUCTION

Computer technologies and television technologies are going closer each other since the growing of the digital TV market recently. It’s a great challenge for computer technologies in broadcasting and television industry to store, manage and share more video, audio, picture and text media information. The issue of building an inexpensive large-scale storage system based on commodity components to meet the dramatically increasing requirements of video storage must to be resolved now. Here we proposed two outline designs based on NAS cluster and IP SAN technology respectively.

The features of the storage system for enterprise level video applications are very large capacity, easy to manage, throughput, scalability and high availability. We look into capacity and manageability in details here.

Capacity: It’s almost impossible to save raw digital video information. The capacity requirement is amazingly huge to store broadcast quality film and video, even after compressed data by MPEG standards. As the current standard for digital TV, MPEG-2 can reach a high compression ratio of 20:1 and variable transmission bit-rate from 4Mbps to 50Mbps. One hour of TV program can consume at least 9GB storage space while estimate as average bit-rate of 20Mbps. The whole storage capacity requirement of a TV station that owns several hundreds of thousands hours of films should be up to several thousands of terabytes. Think about other factors, such as storing MPEG-1 or MPEG-4 format files at the same time for index and browser, then the required space should be larger.

Manageability: The administrative overhead concerns the TCO (Total Cost of Ownership) deeply in enterprises. The management issues include maintenance of storage devices, allocation of storage space and bandwidth, backup and fault tolerance, configurations of user rights and security, etc. The most important issues for storage systems involve hierarchical storage management (HSM).

The current implementations of large-scale storage system on the market for video applications, such as programs editing and broadcasting network based on disks at some TV stations, have several features. As viewed from the selection of storage medium, magnetic tapes and disk drives take dominance absolutely. As viewed from the architecture, the principle is hierarchical multi-level storage system based on Fibre Channel SAN (Storage Area Network). There are online, near-line and offline three levels storage commonly. It’s difficult and expensive to build, and blocked the development of storage network in fields about video applications.

The design of near-line storage system is very important in the whole system because it’s trade-off between online and offline level. The online storage system is built on high performance RAID for the strict requirements on bandwidth and availability at nonlinear editing and broadcasting phases, which cost highest and has smallest capacity (several TBs only). By contrast, the offline storage system is seldom accessed and built on tape library or compact disc library, for data backup, archiving, and recovery only. The huge capacity requirement is mainly satisfied by near-line level, which is built on automated tape libraries in current implementations. When user’s I/O request arrived, it’s HSM software’s duty to migrate relevant data between near-line tape library and online RAID.

Clearly, automated tape libraries have shortcomings such as long response time and expansive cost, which can be overcame by disk storage system. Disks have lower response time and higher bandwidth. It’s time to build large-scale storage system on disks for near-line storage, since the disk technology developing rapidly, and more important, the price dropped almost everyday. At the same time, we proposed two cost-effective designs other than FC SAN architecture: NAS cluster and iSCSI based IP SAN.

The remainder of this article is organized as follows. We discuss today’s prominent storage architectures and selection of storage medium. We describe NAS cluster and iSCSI based IP SAN, the two network storage architectures respectively. We discuss management issue of storage system. The last is some related works and conclusions.

2. DESIGNS

2.1. Storage Medium

The magnetic tape technology is very mature now through decades of year’s research and development. It’s reliable and cheap medium for data storing. But there are some important problems with automated tape libraries near-line storage system. At first, the response time about several seconds to tens of seconds is so slow by contrast with the online storage, which only some milliseconds. This issue is caused by tape’s sequential access manner and hard to resolve. Second, there are lots of mechanical devices including automated robots running at high speed in order to transmit tapes in automated tape libraries, so failures occur more often, and the cost is high.

By contrast with tapes, disks can be accessed random and response time is about 10 ms only or even less. The I/O bandwidth of disk drive is higher than tape drive. For performance and reliability reasons, the markets of online server and enterprise storage have been dominated by drives with SCSI and FC interface for a long period. And their price is higher than ATA drives, whose market is for desktop PCs. In addition, ATA disk arrays are already gaining ground in near-line storage and disk-to-disk backup applications. Analyst of company IDC estimates that about 87% of all drives today using ATA. Economies of scale have made ATA disk arrays at 1 to 2 cents per megabyte, much cheaper than SCSI, at 3 to 5 cents per megabyte.

We even have better choice now. IEEE approved Serial ATA standard in November 2002, and related products of disk drives and array controllers appeared soon. The initial pricing will be about 10% higher than for ATA drives and should benefit from those same economies as ATA. Serial ATA's biggest potential benefit lies in its price/performance. It has several advantages over the parallel, shared-bus master/slave architecture of ATA [1], such as less data lines, longer cable length, native command queuing. Following table 1 compares some features of SATA, ATA and SCSI standard.

SCSI has performance advantage over Serial ATA for server online transaction processing (OLTP). In addition, Serial ATA is lack for supports by related industry standards in fields of backboard connections, device maintenance and enclosure sub-systems, as SAF-TE (SCSI Accessed Fault-Tolerant Enclosures) and SES (SCSI Enclosure Services) for SCSI.

Serial ATA drives can provide high transfer rate for large size files and excellent cost/performance. It's suitable choice to build large-scale storage systems for video applications.

Table 1. Compare with ATA, Serial ATA and SCSI.

	ATA	SATA (1.0)	SCSI (Ultra320)
Transfer rate	100 MB/s	150 MB/s *	320 MB/s
Speed	7,200 rpm	10,000 rpm	15,000 rpm
Devices/channel	2	128	15
MTBF	40,000 h	60,000 h	1,200,000 h
Max cable length	40 cm	1 m	12 m
Hot plug	Not support	Support	Support
Command queue	Not support	Support	Support

** Note: According to the plan of IEEE's Serial ATA Working Group, Serial ATA II is expected to double transfer rates by mid 2004, and a third-generation standard could reach 600 megabyte/second by 2007.*

2.2. NAS Cluster

The most important issue of designing such a large-scale system is, what architecture should be use for inter-connecting of drives?

SAN architectures have been implemented in most of native storage network projects. All of the drive arrays, tape library and application servers are connected by Fibre Channel, forming a high-speed storage network. SAN provide block-based storage services and data sharing for users who connected through Ethernet LAN. Video applications can benefit from high bandwidth, centralized storage and management capabilities of SAN. But there are remain many problems, such as difficulty of sharing storage devices across platforms, cost expensive of FC disk drives and network switch, complexity of technologies.

The principle of our design is use off-the-shelf commodity components as possible to lower hardware costs. It's impossible to connect enough capacity of drives in one enclosure by IDE or SCSI bus, except for designing proprietary backboard, which should increase cost. Then, the approach of distributed storage by connecting nodes through high-speed networks may be the most simple and cost-effective design, just like clusters.

Cluster technology was developed about twenty years ago, for purposes of high performance computing, high availability, or load balancing. In the field of large-scale science computing, for example, mainframe and supercomputer have occupied most of the markets for a long period. And high performance clusters have very important place in the field now. There are 129 cluster systems in the list of top 500 most powerful computers all over the world till June 24, 2003 released recently, about 30% of all. [2]

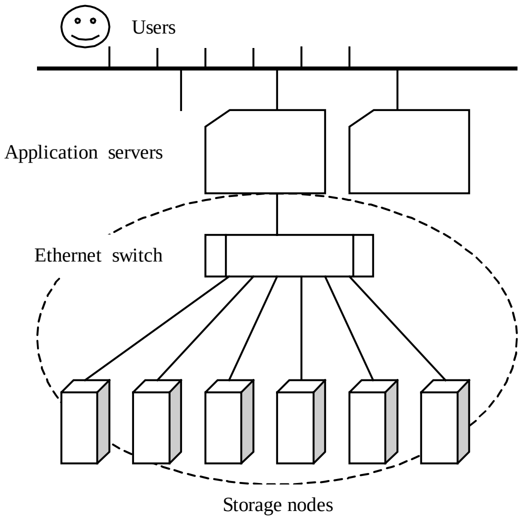


Figure 1. Storage network architecture.

The idea of cluster can be applied to storage systems also. As shown in figure 1, numbers of low-cost PCs act as storage nodes and connect to networks by switch. It seems that each brick a Network-Attached Storage device providing file level data sharing. All of them make up of a storage cluster. The multiple drives in every one node can be independent or organized as RAID array. The advantages of RAID are higher data transfer rates, higher reliability and availability benefit from data redundant and hot spare drives. It increases hardware price at the same time because of RAID controllers.

Traditional NAS servers are independent each other and provide file sharing based on NFS and CIFS protocols. A NAS cluster is not only pile of PC nodes, but also a software system.

It’s important to create cluster software for users who manage and share all of the storage space. In order to hide the distribution of individual components for clients, it’s necessary to implement SSI (Single System Image) that giving some levels of transparency for access, location and scaling.

Access transparency: Clients are unware of local or remote files. Users can login in clusters just as a visual host through a single access point, and it’s not necessary to login on every individual node for services. In addition, it’s necessary to get single graphical user interface (GUI) tool for global management and configuration.

Location transparency: Clients are working in a uniform name space. The most important is to implement single file system hierarchy (SFS). What users see is a hierarchy of whole cluster file system. There is only one single directory tree and all files locate under the same root. It’s not necessary to locate a file to a determinate node.

Scaling transparency: It’s possible to expand system performance and capacity by incremental connecting new storage nodes to cluster, without requirements of modification to system architecture or application algorithms. And the scalability should not increase complexity unacceptable.

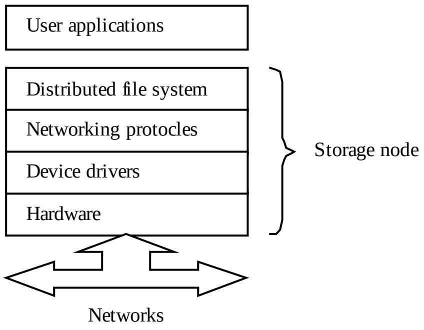


Figure 2. NAS cluster software architecture.

The SSI of clusters can be implemented at different software levels (as shown in figure 2) such as user applications, file system or device drivers level, each has advantage and disadvantage respectively. The approach of distributed file system has capability of full SSI, control of files distribution across storage nodes and optimized performance for file access. On the other side, the cost of using new file systems is high for users. In general, distributed file system is the suitable choice thinking about complexity, performance and cost. Examples for distributed file systems including AFS, Coda, xFS and DAFS, etc.

2.3. Iscsi IP SAN

Another cost-effective solution for storage network is IP SAN. It has advantages of high performance, flexibility and easy to manage. Its hardware platform is as same as the former NAS cluster design. And the software architecture is shown in figure 3.

The Internet SCSI (iSCSI [3][4]) technology got attentions because of the hope to resolve problems involve high price and complexity based on Fibre Channel related protocols and products. The iSCSI is a storage protocol over IP prepared by the Internet Engineering Task Force (IETF). The main idea is transfer data blocks over IP networks by encapsulating SCSI commands into TCP/IP packets. Using iSCSI and Gigabit Ethernet to construct IP SANs makes it possible to replace FC end devices by IP storage devices. It’s advantages over SAN based FC including:

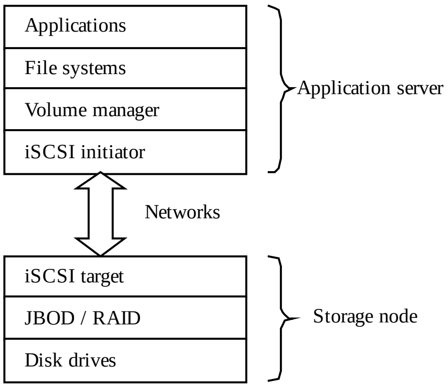


Figure 3. iSCSI software architecture.

Both storage network and LAN are based on TCP/IP related protocols. The benefits are avoiding of building two networks and easy of inter-operation.

Ethernet products are cheap and technologies are well studied. The IP SANs based on Ethernet should have lower hardware cost and less complexity accordingly.

Gigabit Ethernet studied many new technologies from Fibre Channel to gain higher performance. And the management technologies for Ethernet are more mature.

IP SAN break the 10 kilometers limitation for Fibre Channel. It’s easier to implement remote mirroring and disaster tolerance.

Besides, there are problems to resolve for IP SANs. The main issue is that TCP/IP protocols were not designed for transmission of storage data blocks. Then it is difficult to transfer data as efficient as FC and SCSI. It requires more capability to process I/O load. It’s useful to add TOC (TCP Offload Engine) chip to network interface cards for resolving the problem.

2.4. Storage Management

The Information Lifecycle Management (ILM) is a hot topic in storage industry recently and some corporations like EMC and StorageTek approved their ILM policies. What ILM does is to deal with producing, storing, access, share and deletion of enterprise’s data from a dynamic and periodic view. It’s really a software system for managing of information resources in enterprises around the approaches to get data. ILM maybe has various forms with the core thinking of management unchanged. It’s implementation for broadcasting and video industry is Media Assets Management Systems building on storage management systems. MAMS improves the traditional methods for sorting, saving and searching files manually in a TV station.

The most important thing for storage management system is to satisfy data access and share by hierarchical storage management. HSM can do data migrations based on user-defined rules, and provide storage visualizations function aimed at showing a uniform space to clients. If user required data is located at near-line or offline level storage, it's HSM software's duty to copy data to online storage system transparently and return to user. A disk based near-line storage system can decrease access delay obviously.

3. EVALUATION

The above two designs have the same hardware platform of using inexpensive PC containing Serial ATA drives as storage nodes and networking based on Gigabit Ethernet.

From view of software architectures, NAS cluster provides file services above file systems, and IP SAN based on iSCSI protocol provides data block level services. This difference has effect on final I/O performance. In addition, it's known that current HSM and backup management software for SAN require block devices to run. Then it's easier for iSCSI design to migrate those softwares without too many modifications. For example, it's possible to do some modification in iSCSI drivers to visualize disk drive to tape drive for the original HSM software of automated tape libraries. The solutions of remote mirroring, backup and disaster recovery for IP SAN are acquirable easier.

On the other side, the scalability of iSCSI maybe limited by the maximum number of SCSI targets that operating systems could support, 128 for Microsoft Windows 2000 in our test, for example. Though it's possible to resolve the problem by improving software architecture, it's not necessary to implement unlimited scalability if the capacity requirement can be satisfied under the limitation.

4. RELATED WORK

Dr. Koch of Computertechnik AG has developed a cheaper, yet very flexible solution of backup system for the University of Tübingen: 70 TB capacity on standard IDE/ATA hard drives [5]. While the biggest consideration in a backup solution is its reliability, and reliability is not drives with IDE interfaces are designed for. Cheaper price is the reason why use IDE instead of SCSI. And it's enough to apply RAID technology to improve reliability. There have 24 nodes with 3 pieces of 3Ware IDE RAID controllers, 576 drives totally. All of the nodes are connected with Gigabit Ethernet, and running backup management software. The approach is very similar to our storage cluster, but lack of SSI.

The Tiger video fileserver developed by Microsoft in 1996 aimed at providing VOD (Video On Demands) for large number of clients [6]. It was constructed using cheap off-the-shelf PCs also. The video data is distributed across storage nodes, called cub, through stripping and mirroring for load balance and fault tolerance purposes. Some controllers running distributed algorithm program to schedule I/O load to cubs. The Tiger lies on QoS of ATM networks in order to ensure sequential and real-time transferring video data block to users.

5. CONCLUSION

To satisfy the requirement of large-scale storage for video applications we approved the NAS cluster and iSCSI based IP SAN outline designs, which have following advantages:

1) Using standard commodity components and PCs as storage nodes to lower cost.
2) Using Serial ATA drives instead of tapes or SCSI drives for better cost/performance.
3) Based on Gigabit Ethernet to build storage cluster or iSCSI protocol IP storage network to simplify architecture and improve manageability.
4) Scaling capacity by adding storage nodes into network.

These cost-effective designs are useful for not only digital video near-line storage but also data backup systems.

6. REFERENCES

[1] "Serial ATA Takes on SCSI." Computerworld, March 31, 2003. http://www.computerworld.com/printthis/2003/0,4814,79769,00.html

[2] TOP500 Supercomputer sites. http://www.top500.org/

[3] Julian Satran, et al. iSCSI. 2003.2. http://www.ietf.org/internet-drafts/draft-ietf-ips-iscsi-20.txt

[4] Eric Blair, Michael Mesnier, and Quy Ta. Intel iSCSI Reference Implementation. http://sourceforge.net/projects/intel-iscsi/

[5] "Hard Drives Instead of Tapes? 70 TB Backup RAID at the University of Tübingen." Tom's Hardware Guide, April 25, 2003. http://www.tomshardware.com/storage/20030425/index.html

[6] W. Bolosky, et al. "The Tiger video fileserver." 6th NOSSDAV Conference, Zushi, Japan, April 1996.

A Distributed Volume Visualization Architecture on the Grid

Yaolin Gu Zhe Cao
School of Information Engineering, Southern Yangtze University
Wuxi 214036, Jiangsu, P.R. China
Email: gyl627@sytu.edu.cn Tel: +86-0510-5863635

ABSTRACT

In this paper, we present a distributed software architecture for volume visualization that utilizes conventional PCs to generate high-quality interactive graphics. We present a method for enabling progressive client–server volume visualization of data from the computing grid. Rendering is performed on clients, while servers on the grid provide wavelet compressed volume data. It provides a modular framework that can accommodate a wide variety of rendering algorithms and data formats. Demonstration modules that implement ray tracing, fractal rendering, and volume rendering algorithms were developed to evaluate the architecture. Results are encouraging, the system can interactively render simple to moderately complex data sets at modest resolution. Excellent scalability is achieved.

Keywords: distributed architecture, volume rendering, ray-tracing, grid computing

1. INTRODUCTION

In recent years, some types of visualization and rendering techniques (such as ray tracing [6]) are still too slow to be used for real-time applications on desktop computers. These techniques are desirable as they produce imagery of superior quality and realism, making them useful for a range of applications from medical visualization to architectural walkthroughs. Fortunately, visualization/rendering is a problem that lends itself very well to parallel processing [5]. By using large multiprocessor computers, techniques such as real time ray tracing are feasible. Parker et al. have developed a real-time ray tracing system that was demonstrated on a 60-CPU Silicon Graphics Origin 2000 [12]. Their system performs well enough to be useful for many applications—one of their examples is a visualization tool for CT Scan data sets [9], [8]. Realistically, few organizations have easy access to equipment more advanced than desktop computers and small single or dual-processor servers.

In this article we have concerned with the problem of making interactive volume visualization available to computing grid users. Recent increases in computational power on grids imply an increased value in understanding data, and therefore, an increased need for interactive visualization of data that ' s resident on grids. Making interactive volume visualization more readily accessible would greatly enhance large scale computation on the grid. Much of the data used in grid computations can be best understood with 3D volume visualization. Volume rendering enables direct visualization of data collected from physical measurements. Medical visualization (from, for example, computerized tomography, or CT scans), seismic visualization (displaying subterranean densities derived from sound velocity), and weather visualization (from atmospheric measurement) provide examples where volume visualization adds understanding of an existing physical system by letting users look directly inside volume for important features. Volume rendering also is often the best way to visually understand synthetic data such as in wind tunnel or galactic simulation results.

A significant advance in the grid's usefulness will result when volume-rendering data becomes easily accessible to grid users. It will let engineers and scientists readily understand and communicate their understanding of data collected at remote locations, and let doctors and patients interactively visualize medical problems and preplan corrective surgery. We've already overcome many of the barriers to widespread volume visualization on the grid. For example, we can inexpensively install physical memory adequate for storing large volumes (gigabytes) on desktop computers. The current generation of inexpensive rendering hardware on PCs already perform real-time rendering of moderately sized volumes, and such hardware will soon be capable of interactively rendering gigabyte volumes. However, one significant barrier remains: communication bandwidth limitations. Only the most advanced networking technology can approach the communication rate required for immediate access to volumetric data. As a result, volume visualization of data on a computing grid requires users to use high-powered visualization systems directly connected to the grid.

Our approach to the problem of visualizing data on the grid advances several topics not addressed in the literature. These include a new mechanism for visibility detection, a specification of a new client–server division of effort, and a new mechanism for filtering wavelet coefficients through octal tree prioritization.

2. SYSTEM ARCHITECTURE

Until recently, Mucus' network distributed ray tracer [13] was known in 1987. Wylie [4] has demonstrated a system in 2001 that uses off-the-shelf PCs and graphics cards to provide rendering performance of up to 300 million polygons per second. The Parallel Virtual Machine (PVM) [1] and Message Passing Interface (MPI) [11] software packages are also popular. W.Bethel introduced his Distributed Visualization Framework in 2000 and 2003 [2],[3] .

Our client–server architecture can visualize the results of grid computations. The supplier of data installs volumetric data resulting from a computation at a server site on the grid that ' s suitable for distributing data to visualization clients. The rendering occurs on a client machine with sufficient memory and rendering performance to support interactive volume rendering of local data. We distinguish here between *interactive* and *responsive* volume visualization. A visualization session is interactive.

If users can navigate through the scene, rendering at 10 frames per second or faster. A visualization is responsive if the users experience a delay of 2 seconds or less between when a volume is selected and when they can access a detailed rendered image. Interactive visualization requires adequate rendering hardware; responsive visualization requires sufficient data communication (or local storage) of volume data. Our architecture provides responsive visualization of data on the grid, assuming the users already have access to an interactive volume visualization capability. We implemented the components of this architecture in software to understand the data communication requirements. This lets us simulate the functionality, but not predict the architecture ' s full performance. We do predict the data communication that will result in a visualization session and thereby predict the responsiveness under the assumption that other aspects of the system are interactive.

Fig. 1 illustrates our general architecture. The grid computation results in a volumetric data set. To set up the data server, the wavelet transform is applied to the full volume data set, resulting in a repository of wavelet-encoded data that's maintained at the data server. The client workstation progressively builds and maintains its own 3D array of volumetric data. The client volume data array—initially only a coarse approximation to the original volume data—is progressively populated by reconstructing data from wavelet-filtered coefficients transmitted from the data server. The client data array is a set of power-of-two-aligned subcubes that track visibility. The client prioritizes these subcubes according to their relevance to the current visualization. We call high-priority subcubes *frontal subcubes* because they usually occur in the portion of the volume facing the viewer. The client periodically updates its volume data by obtaining improvements in the data associated with frontal subcubes. The client tracks the downloading progress of this associated data, recording filter thresholds associated with the volume data installed on the client and the visibility of the subcubes.

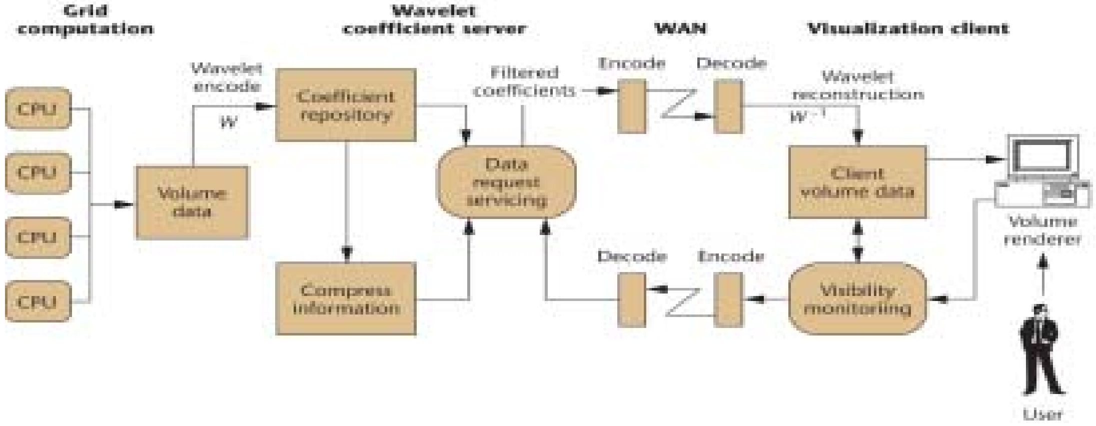


Fig.1 System Architecture of Distributed Visualization

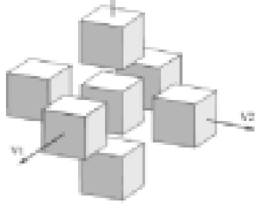
Periodically, as more data is required due to the recent needs of the visualization, the client submits a request to the server. The requests consist of a list of frontal subcubes *S* of the volume, each associated with a filter threshold *TS*, indicating the level of filtering already applied to the wavelet coefficients previously used to reconstruct the client volume data in *S*. The response to this request is a set of wavelet-transformed and filtered coefficients *c* and a new filter threshold *T*new such that all coefficients *c* returned satisfy $TS > |c| \geq T\text{new}$, (here coefficient *c* is associated with subcube *S*). Once the client receives this data, it performs a wavelet reconstruction of the transmitted coefficients, adding the resulting data into the client volume repository.

This process repeats as long as the client needs additional data to support the visualization. When the currently visible data has been sufficiently decompressed so that there's no value in additional updates to the frontal cubes, additional volume data from other (not visible) portions of the volume can be reconstructed, improving the client volume repository's overall accuracy.

3. VOXEL AND WAVELETS

We can measure the importance of a voxel (See Fig. 2) in the data volume by the contribution it makes to the rendered image. To determine voxel visibility, recall the computation of an image pixel in a typical front-to-back volume rendering algorithm, such as ray casting, ray tracing, or shear-warp rendering. During computation, the opacity of voxels along a ray accumulates from front to back to determine the color of a pixel p. Let α be the accumulated opacity of the voxels in front of voxel *v*, and let β be the opacity associated with *v*. Then we measure the contribution of *v* by$(1-\alpha)\beta$. After these values have been accumulated for p, we multiply this contribution by the shading parameters of *v* to determine the color of p. We call this the *importance* of voxel *v*, or *I*(*v*). *I*(*v*) depends on the viewer position and the data's opacity. The approximate important value, computed from the compressed data available at the client, is *I*(*v*). It approximates the actual value of *I*(*v*), and the approximation improves with the accuracy of data installed on the client.

Fig. 2 Vowel data



Assume the data volume is a cube with sides a power of two. The volume's subcubes provide a convenient unit for communicating data between client and server. We use subcubes as the unit for identifying visibility as well as the unit for specifying local data compression. Client requests are expressed in terms of subcubes. During the rendering of the current (last installed) data on the client, the client identifies all the voxels *v* whose importance exceeds an importance threshold value *T*, (that is, voxels satisfying .$I(v) \geq T$). The subcubes that contain these voxels, and subcubes immediately adjacent to them, are the *frontal subcubes*. The client then uses the identification of frontal subcubes to construct a request of the wavelet coefficients that control the frontal subcubes ' contents. Using wavelet compression, the server transmits only the larger coefficients that influence the frontal subcubes. The server determines cutoff value *C* that controls the level of compression of the transmitted coefficients.

Wavelets provide the means to transmit information from the grid to the visualization user. We rely on the wavelet compression's effectiveness to reduce the amount of data needed to support visualization. We also exploit the locality features of wavelets. We can selectively filter the wavelets associated with visible portions of the volume data, applying more filtering to less visible data.

4. COMMUNICATION MECHANISM

We've prototyped and simulated the design for the communication architecture and associated calculations that support the interactive visualization of volumetric data from the grid. We haven ' t deployed the prototype, however, we've implemented the algorithms and measured the communication requirements. An important practical requirement of this architecture is that the servers shouldn't track the client state. There's no maintained session (other than as needed for authentication) between distinct client–server communications. This requirement ensures improved fault tolerance and recovery. If communication is lost, data is corrupted, or a server is disabled, the client can repeat a data request or request the data from another server. While waiting for data, the client can interactively visualize the volume data already received from the grid.

Note that this stateless server requirement limits the client knowledge of the data repository. We could improve the architecture performance if the server were to monitor the client viewpoint and transmit the actual data needed for a rendering, rather than waiting for the client to estimate the data it needs.

Client requests

The client makes the first data request of the server with no information about visibility. The client requests a uniform compression of the entire volume, obtaining the maximum amount of data (the data download size) that can be transmitted without interfering with the client–server communication ' s responsiveness. The client uses this data to initialize the client repository. Subsequent requests are based on monitoring the visibility of the volume data stored at the client. These requests consist of a list of frontal subcubes together with the latest amplitude cutoff, indicating the cutoff value associated with the latest update of each subcube.

Server response

When the server receives a request from a client, it first determines the amplitude threshold *C* that will filter the data to satisfy the request. The server determines a desired compression ratio that ' ll result in a data download size of *D* bytes and then uses its amplitude histogram to find the amplitude cutoff *C* that results inapproximately the desired amount of data. The actual data to be transmitted to the client consists of the following:

*For each frontal subcube *S* with latest amplitude cutoff *CS*, the set of coefficients *c* contained in the wavelet hierarchy associated with *S* so that $C \leq |c| < CS$.

*For all levels of the wavelet tree above the level of subcubes, the set of coefficients *c* so that $C \leq |c| < CP$, where *CP* is the previous amplitude cutoff associated with the most recent request from this client.

*The value of the new amplitude cutoff *C*. The server quantizes and run-length encodes the filtered data, then transmits it to the client.

Client response

When the client receives the data from the server, it applies a wavelet reconstruction to the returned coefficients. It adds the resulting data values into the volume data array. For each updated subcube, the client saves the value of that new threshold in its list of subcubes. After the client has added the new data into the client volume array, the client again determines the new set of frontal cubes, using the latest viewpoint and transfer function determined by the user. This will determine the next client request.

5. DEMONSTRATION PROGRAM

Several interactive programs were written to demonstrate our system and evaluate its performance— Pick, Temple, and CT Scan.

The Pick demo (Fig. 3) uses ray tracing to display a simple scene consisting of two spheres, a planar floor, and a planar wall. The user is able to change the surface types for each of the objects by clicking on them with the mouse. The user can also move the viewpoint to "fly" through the scene.

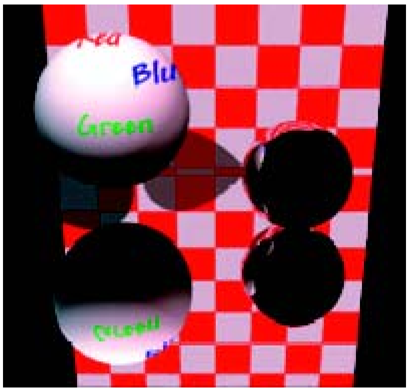


Fig. 3 Sphere Demo

The Temple demo (Fig. 4) uses ray tracing to display an outdoor temple-like structure with 80 objects and four light sources. Like "Pick," the user is able to fly through the scene.

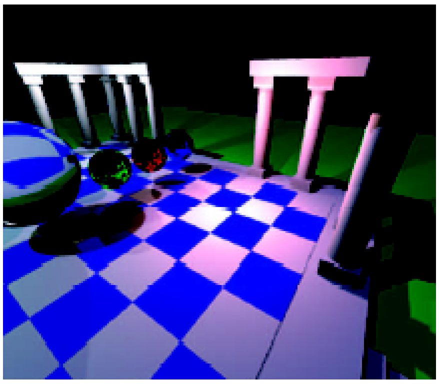


Fig. 4 Temple Demo

Several reflective spheres are animated—they orbit a central glass sphere. Due to the scene's relative complexity, this demo requires much more processing power than the Pick demo, but does not have the large memory footprint of the CT Scan demo

.

The CT Scan demo (Fig. 5) uses ray tracing to display a volume data set containing a CT Scanned human head [7]. The ray tracing algorithm used is similar to that described in [9]. The CT Scan's volume data can be thought of as a tissue density map. Rays are cast through the volume (direct volume rendering), which consists of density or sample values evenly spaced on a 3D grid. The direct volume rendering technique eliminates the need to convert the volume into a polygon mesh before rendering, such as with the Marching Cubes algorithm [10]. This reduces memory requirements and improves interactivity as there is no polygon mesh to recompute when viewing parameters are changed. The Isosurface mode (Fig. 5) that displays a surface representing the locus of a density value specified by the user. This surface is fully shaded and may be self-shadowing.

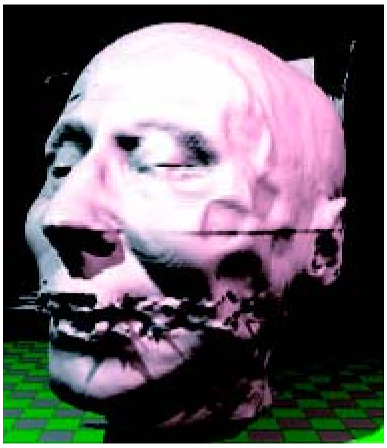


Fig. 5 CT Scan Demo

7. CONCLUSION AND FUTURE WORK

The results show where our progressive communication scheme works well and also indicate areas for further research. We evaluate this architecture's responsiveness by considering only communication time (not calculation time) and assume data is transmitted from the grid to the user at Ethernet rate. Consider how this scheme would operate when viewing on a low speed network.

Our goal is to let remote users of a computing grid interactively and inexpensively perform volume visualization of their data without significant communication delays. This is our future work.

8. REFERENCES

[1] A. Beguelin, A. Geist, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam, PVM: Parallel Virtual Machine, A Users' Guide and Tutorial for Networked Parallel Computing. MIT Press, 1994.

[2] W. Bethel, "Visapult: A Prototype Remote and Distributed Visualization Application and Framework," *Siggraph 2000 Conference Abstracts and Applications*, ACM Press, 2000.

[3]. J. Shalf and E. W. Bethel, "Cactus and Visapult: An Ultra-High Performance Grid-Distributed Visualization Architecture Using Connectionless Protocols," *IEEE Computer Graphics and Applications*, vol. 23, no. 2, Mar./Apr. 2003, pp. 51-59.

[4] B. Wylie, C. Pavlakos, V. Lewis, and K. Moreland, "Scalable Interactive Rendering on PC Clusters," IEEE Computer Graphics and Applications, pp. 62-69, July/Aug. 2001.

[5] A. Chalmers and E. Reinhard, "Parallel and Distributed Photo-Realistic Rendering," Proc. Course #3,SIGGRAPH98 Conf., July 1998.

[6] An Introduction to Ray Tracing. A.S. Glassner, ed., Academic Press, 1989.

[7] SoftLab Software Systems Laboratory, "The Chapel Hill Volume Rendering Test Dataset,"vol. II, Univ. of North Carolina, Dept. Of Computer Science, 2000.

[8] S. Parker, P. Shirley, Y. Livnat, C. Hansen, and P.-P. Sloan, "Interactive Ray Tracing for Isosurface Rendering." Proc. IEEE Visualization '98, 1998.

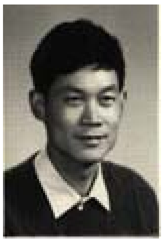
[9] S. Parker, M. Parker, Y. Livnat, P.-P. Sloan, C. Hansen, and P. Shirley, "Interactive Ray Tracing for Volume Visualization," IEEE Trans. Computer, Graphics and Visualization, pp. 238-250, July-Sept. 1999.

[10] W.E. Lorensen and H.E. Cline, "Marching Cubes: A High Resolution 3D Surface Reconstruction Algorithm," Computer Graphics, vol. 21, no. 4, pp. 163-169, 1987.

[11] P. Pacheco, Parallel Programming with MPI. Morgan Kaufmann, 1996.

[12] S. Parker, W. Martin, P.-P.J. Sloan, P. Shirley, B. Smits, and C. Hansen, "Interactive Ray Tracing," Interactive 3D, Apr. 1999

[13] M.J. Muuss, "RT and REMRT—Shared Memory Parallel and Network Distributed Ray-Tracing Programs," USENIX: Proc. Fourth Computer Graphics Workshop, Oct. 1987.



Gu Yaolin : A Full Professor and Deputy Dean of the School of Information Engineering, Southern Yangtze University, P. R. China. He graduated from the Computer Science Department of Shanghai Jiaotong University in 1982. As a Visiting Scholar of the University of Stony Brook in USA from 1994 to 1995, his main research interests are Computer Graphics and Parallel Computing. Now he is a Senior Member of Computer Society, IEEE.

Analysis of Distributed Video-On-Demand System Based on Cluster

Zheng Shijue^{1,2} **Ma Wei** ^{2} **Zhang Jiangling**^{1}
School of Computer Science and Technology, Huazhong University of Science and Technology1
Department of Computer Science, Central China Normal University 2
Wuhan, Hubei 430074, P.R.China
Email: zhengsj@ccnu.edu.cn mw0626@163.com **Email:** jlzhang@hust.edu.cn

ABSTRACT

Video servers are essential in Video-On-Demand and other multimedia applications. In this paper, we propose architecture of the Clustered Video-On-Demand (CVOD) system with high performance, high availability and low cost, which is based on a cluster of personal computers. In this system, we adopt Platform LSF to manage and balance workload of the video servers. It efficiently improves the OoS and optimizes the cost/performance ratio of VOD. This high performance system is mainly designed for a good video education in the university. We present the CVOD architecture and analysis the key factors of the system such as stripping, data placement policy and fault tolerance.

Keywords: CVOD, Platform LSF, Distributed Computing, Server striping policy

1. INTRODUCTION

Developments in computer and communication technologies have made VOD (Video-On-Demand) services a reality. One of the great challenges in VOD domain is to optimize the cost/performance ratio of the designed system [1]. Current VOD systems are commonly designed around the client-server architecture. Under this architecture, a client sends a request to the video server for a video title and then the server transmits video data to the client for playback. As the number of user increases, the server will eventually reach its capacity limit. To further increase the system capacity, one can add more servers to share the load [2].

However it is impossible to satisfy the infinite rapid-increasing requirements with finite addition of servers and bandwidth in the traditional VOD system. Currently most of VOD system with a single expensive high-end server is available for many universities. For example, there are 240 educational video coursewares about 220G stored in VOD server of our university campus network. It mainly provides video education to about 15000 students. Once the server is invalid, the whole system will be paralyzed and it will greatly influence the normal teaching of the university. In this paper, we propose a novel architecture based on CVOD to resolve this problem. In Cluster-type architecture the CVOD system not only benefits from the secure and efficient organization in cluster computing, but also benefits from the cost-effectiveness, flexibility and scalability. Furthermore, We adopt Platform LSF5 to apportion workloads among the clustered servers fairly. It is a cost-effective high performance system for supporting VOD services.

2. RELATED WORK

Theoretically a perfect VOD system should solve the following primary problems:

i) In a complicated VOD system, different users will play the videos in different rates according to the network load and their schemes. In this case, system should provide different qualities of service (QoS).

ii) Sometimes masses of requests focus on a few hot video titles. A good VOD system should deal with hot video problem very well.

iii) Server stripping ensures that loads from each and every client will be evenly shared across all servers on the average.

iv) When some servers are invalid, the system should transfer their assignments to other working servers without influencing normal working.

v) Store more video data in finite storage capacity.

vi) Update the video data and maintain the servers without influencing normal work.

During the past years a lot of efforts have been done in the design of VOD solutions. People proposed many models about VOD system. They presented the models in different ways. Roughly they can be divided into two types. One is client/server model, the other is distributed DBMS model.

Client/Server model

It is a typical VOD model that consists of a set of centralized video servers and geographically distributed clients connected through high-speed networks. In this model, the functions of server include service provider, system management and network management. Clients are only users. And they store video data by a single file system. When the scale is small, the system can work well. But when the number of user increases, the server will eventually exceed its capacity limit.

VS-SOC model (Distributed DBMS model)

VS-SOC model is a popular VOD model based on ATM network. In this model, a high performance DBMS of video servers is the primary factor. To optimize the path from clients to video servers, several video servers with different video data are distributed at the network. Traffic manager and server operating center (SOC) improved the OoS.

No matter which models we adopt, if bandwidth is enough, disk I/O bandwidth and computation power of the servers will be the bottleneck of the system. Faced with the demand of increasing the capacity of the servers, VOD system based on cluster server is a good way to solve the problem.

3. DESIGN OF THE CVOD

3.1 Architecture of the CVOD

The CVOD system is composed of the following modules: clustered server, a storage system, a network and clients.

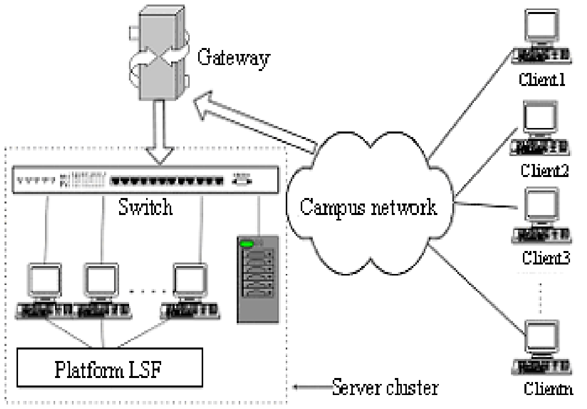


Fig.1 Architecture C/S of the CVOD

Figure1 shows the architecture of the CVOD system, which is based on a client-server model. According to the layout of whole system, dedicated video servers consist of several slap-up PCs which form a typical distributed subsystem. It is in charge of original distribution of video data, initial load balance, security policy, global admission control policy, and overall availability, etc. Each server has its own CPU, memory and disk storage. Each component in this system is eventually connected by campus network..

The distributed architecture is based on a cluster of low-cost PCs attending to a great number of client requests. Clients are not expected to send requests directly to the servers. The requests go first to the router that redirects them to the appropriate server taking into consideration issues such as block location and load balancing.

CVOD system has two important advantages compared with the single computer machine solutions. The first one refers to the ease that a server can be extended using more nodes. The cost of changes that must be performed is negligible, compared to the profit that the addition of new nodes introduces to the system. The second one refers to the requirement that, after some node's failure, the whole system should continue to operate correctly. Our clustered system fully satisfies these two key factors keeping the overall cost low.

We adopt platform LSF to provide a powerful environment for distributed computing. Using Platform LSF, organizations can harness the untapped power of all distributed video resources. It ensures fair sharing of resources among users, and allows policies to be configured to meet the most complex and demanding requirements. Monitoring and reporting tools provide constant, up-to-date information about the state of an organization's computing environment. At the same time, it is a good solution to balance load on the servers.

3.2 Hardware Requirements

COVD is based on a cluster of 10 PC s (IBM NetVista M42, PIV 2.4G processors, 512M memory, 80G hard disk, 100M PCI network card per PC), just like nodes of a distributed VOD server. The interconnection between them is succeeded through a fast Ethernet switch element (Catalyst 3548 XL：48 slot, 10/100M，2GBIC slot), which connects to 1000M of fast Ethernet university campus. Successive switches are connected together creating a tree structure in order to distribute uniformly the bandwidth of the network and to provide more points for clients' connection. A straightforward future upgrade could be the replacement of the backbone network with Gigabit Ethernet. Our primary concern when we started the design of this project was to keep the overall cost low. When the size and the needs of the application grows up to a point that cannot be further satisfied by a single node, a simple addition of a second node can almost double the performance. The results of our experiments showed that the performance of the VOD system increases almost linearly to the number of nodes. In our case, the clients (students) were simple PCs, but generally, it could be any device (Set Top Box) with the appropriate processing power and an Ethernet connection interface. The client application may work fine on a Celeron at 333 MHz with 64 MB RAM and a 10MBit/s network interface card.

3.3 Software Design

We have chosen Windows2000 operating system, and adopted the software of Platform Corporation LSF5 (Load Sharing Facility) building CVOD working platform. Platform is the world's leading distributed computing software provider. Platform LSF5 supports multiple plug in schedulers, providing unlimited extensibility and writing customizable plug-in schedulers to meet site-specific business policies and eliminate the maintenance overhead. A new Resource Leasing Model provides a single system image across all clusters so users administrators can easily manage resource sharing policies across your Enterprise Grid. A customizable Web-based interface facilitates access to users distributed computing resources from any location, via the Internet. In order to develope the application software of system management and users' interface we use Microsoft Visual C++.

The cooperation between these applications is succeeded as follows: During the initialization of the system, a listening port is created at each node for the incoming requests from clients. When a user decides to watch a video, the client-application accomplishes a connection with the listening port of a node. The selection of the proper node is made through a procedure of the client-application, which consults a structure that keeps pairs of movie title - node IP address. All the movies of the system have an entry in that structure. Multiple copies of the same movie have different entries because they are stored into different nodes.

3.4 Video data processing

Video data traverses from the CVOD storages area and ending at the client's stream decoder is presented in this section. Figure 2 shows the situation on each server node. It is easy to see that the disk server runs two threads.

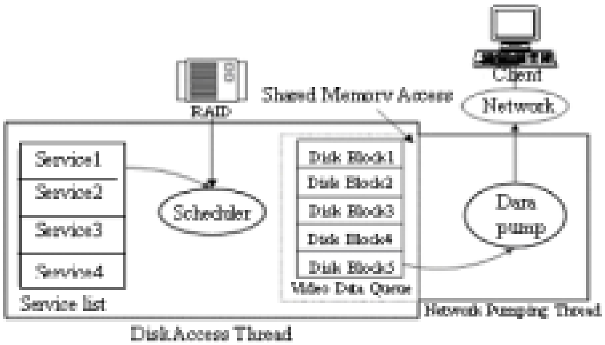


Fig.2 Data Processing in Servers

The Scheduler calculates the order of disk read accesses from the Service List containing a service information entry for each client. After having read the disk blocks containing video data, the Scheduler inserts the video data blocks into the Video Data Queue, a data structure that lies in memory shared by both threads. The Data Pump running in the second thread permanently pushes blocks from the Video Data Queue to the network in a FIFO (first-in first-out) order [3].

3.5 Server striping policies

Striping is a general technique for distributed data over multiple devices to improve capacity, throughput and potentially reliability [4]. In COVD, striping video data over multiple servers increases the system's capacity and potentially improves its reliability through data redundancy. This is called sever striping. Striping a video stream across all servers is commonly called wide striping. There are two types of stripings: time striping and space striping.

3.5.1 Time striping

A video stream can be viewed as a series of video frames. Striping a video stream in units of frames across multiple servers is called time striping. Assume that a stripe unit contains L frames and the video plays at a constant frame rate of F frames per second. In each round of NsL/F seconds, L frames will be retrieved from each server and delivered to a client. In general, the striping size L does not need to be an integer equal to or larger than one. In particular, if $L<1$, then it is called subframe striping, where $L>1$ is simply frame striping. However this striping maybe cause load imbalance.

3.5.2 Space striping

Time striping divides a video stream into fixed-length (in time) stripe units. A second approach would be to divide a video stream into fixed-size stripe units, called striping. Space striping simplifies storage and buffer management at the servers because all stripe units are the same size. Moreover, the amount of data sent by each server in a service round is also the same. A video stream can be striped across the servers independent of the encoding formats and frame boundaries.

This space striping approach is employed by most of the studies already mentioned. Depending on the system design, the stripe unit size can be ranged from tens of kilobytes to hundreds of kilobytes. If the stripe unit size is too large, it maybe congest network and demand large buffers. On the contrary, too small stripe unit size will lead to frequent disk reading. To solve the problem, designers can divide a video stream into variable size stripe units.

3.6 Data Placement Policy

For data placement, a video title is first divided into fixed-size blocks and then equally distributed to all nodes in the cluster. This node-level striping scheme avoids data replication while at the same time share the storage and streaming requirement equally among all nodes in the cluster. To initiate a video streaming session, a receiver node will first locate the set of sender nodes carrying blocks of the desired video title, the placement of the data blocks and other parameters (format, bitrate, etc.) through the directory service. These sender nodes will then be notified to start streaming the video blocks to the receiver node for playback.

3.7 Fault tolerance

The COVD can achieve fault tolerance at server level. That is to say, the system can maintain continuous video playback for all active sections when one or more of the servers become inoperable. As with the disk array and RAID architectures, data redundancy can be added to support failure recovery in a distributed server. The basic idea is to introduce one or more parity units into each stripe. The redundant data allows the receiver to mask a server failure by computing lost stripe units stored in the failed server from the parity units together with the remaining stripe units.

4. PERFORMANCE EVALUATION

In this section, we evaluate the system requirements and performance of the architecture presented in Section 3. We experimented with state-of-the-art video compression algorithms like MPEG-4. First, we present the results of our experiments with a single server node. Furthermore, we present our experimental results with systems of multiple server nodes that prove the linearity between the number of server nodes and the number of served streams. In our experiments, we are concerned with the CPU utilization under these two different conditions. Numerical results are listed in tables.

4.1 Measurements with a single server node.

For our measurements, we used 4 different MPEG-4 movie streams. Although these streams are Variable-Bit-Rate encoded, we can assume that the streaming rate was around 1 MB/s as the variation from that mean value was relatively small. The policy of requests for these different streams was round robin and every 4 seconds a new request was made to the server. In this section we will present the CPU utilization results for one server node.

For our experiments, the delivery protocol was client-pull. We developed a little application that only simulates the network behavior of a client. It is connected with the server and requests for new data at regular time intervals. When it receives these data, it discards them immediately without doing any further processing like decoding and rendering. At each client machine we executed one real-client application in order to check the quality of the video reproduction and a number of simulate-clients to increase the network workload.

Table 1 lists the percentage of CPU utilization of our experiments for this implementation. The client-pull implementation at the 2.4GHz CPU has a good performance, which serves 90 concurrent True-VOD streams utilizing only 21.5% of its total processing power.

Table 1: Analysis of CPU utilization

Concurrent streams	20	40	60	80	90
CPU utilization	4.3%	8.5%	14.2%	18.6%	21.5%

4.2 Measurements with multiple server nodes

Furthermore we present our experimental results for system implementations that utilize more than one server node. We experimented with systems of 2 and 3 server nodes and used data mirroring, copying the same set of movie streams to each one of the disks of the system. The size of client buffers was 2MB and the arrival rate of new requests was 1 every 4 seconds.

If we connect multiple nodes onto the same fast Ethernet switch device, we will theoretically multiply the available network bandwidth by a factor equal to the number of nodes. This is translated to a linear increase of the number of streams that can be concurrently serviced by our system. This hypothesis was confirmed by our measurements that showed that the upper limit of video streams that the system can handle reaches at 172 for the 2-node system and at 241 for the 3-node system. All the measurements were made using the client-pull architectural approach that gave us the best results in the single server implementation.

Table 2: CPU utilization for multiple server nodes

Concurrent streams Server nodes	50	100	150	200
2	4.3%	9.1%	16.2%	24.4%
3	2.1%	6.8%	9.6%	14.7%

In table 2, we see the representation of the mean CPU utilization, for multiple server nodes implementations, with regard to the number of streams that the system can handle. As we expected, the number of concurrent streams was almost double (or triple) at the same level of CPU utilization for the 2-servers (or 3-servers) system in respect to the single-server system.

4.3 Major results

The major results of the CVOD system compared with the system with a single server in our university are as the following:

- It supports much more concurrent clients (lease 3000uers) watching video programs at the same time.
- It is much more reliable. When one or some PCservers are invalid, the system can transfer their assignments to other working servers without influencing normal working.
- Benefit from Platform LSF, workloads are apportioned among the servers fairly.
- The cost of whole CVOD system about 100,000RMB is more cheaper than single server (IBM xSeries 360 8686-3RQXeon 2.5GHz*2/2GB about 190,000RMB which we have used)

5. CONCLUSION

We propose the architecture of the Clustered Video-On-Demand (CVOD) system with high performance, high availability and low cost. This new system architecture supports reuse and standardization of video resource more than before, and achieves well heterogeneity, portability in various platforms. A video on demand server that meets the basic requirements of such a system has been developed: great storage capacity, high bandwidth, predictable response time, large number of concurrent users and fault tolerance. With the development of cluster technology, the research of CVOD will have a leap and the system will become more powerful.

6. ACKNOWLEDGMENT

We thank Professor Liansheng Tan of the Department of Computer Science of Central China Normal University and Professor Hai Jin of the College of Computer Science & Technology of Huazhong University of Science & Technology for their supports of various aspects in setting up CVOD and thanks also go to all members of our research group especially Guangdong Chen and Jianqun Cui for their efforts in the coding task.

7. REFERENCES

[1] Apostolos Papagiannis, Dimitrios Lioupis, Stylianos Egglezos: DESIGN & IMPLEMENTATION OF A LOW-COST CLUSTERED VIDEO SERVER USING A NETWORK OF PERSONAL COMPUTERS, IEEE Fourth International Symposium on Multimedia Software Engineering (MSE'02), December 11-13, 2002

[2] J.Y.B.Lee, R.W.T.Leung: Study of a Server-less Architecture for Video-on-Demand Applications. Proceedings of the IEEE International Conference on Multimedia and Expo 2002, Lausanne, Switzerland, Aug 2002

[3] Jamel Gafsi, Ulrich Walther, Ernst W. Biersack: Design and Implementation of a Scalable, Reliable, and Distributed VOD-Server. Proceedings of the 5th joint IFIP and ICCC Conference on Computer Communications, AFRICOM CCDC, Tunis 1998.

[4] Y.B. Lee: Parallel Video Servers: A Tutorial. *IEEE Multimedia*, pp. 20-18, April-June 1998.

[5] Min-You, Wei Shu, Chow-Sing Lin: Odyssey: A High-Performance Clustered Video Server, Software—Practice & Experience archive,Vol.33 , June 2003, 673-700.

[6] Jack Y.B.Lee, P.C.Wong: Performance Analysis of a Pull-Based Parallel Video Server. IEEE Transactions on parallel and distributed systems. Vol.11.NO.12, December 2000

The Investigation to Distributed Supervision System Based on GPRS

Huie Chen, Congxin Liu, Weilu Zeng, Liqin Zhang
College of Electrical & Information, Three Gorges University
Yichang, Hubei 443002, China
Email: flying_moth@sina.com Tel.: + 86 (0)717- 6397024 + 86 (0)717- 6393376

ABSTRACT

This paper analyzes the difficulties of establishing GPRS supervision system and raises some effective measures. A virtual wireless communication system based on GPRS has been implemented, its capability is being tested and the application prospect is also illustrated in the end of this paper.

Keywords: GPRS, supervision, “always connected”, wireless, wire.

1. INTRODUCTION

With the rapid development of information technology, computers make intelligence melt into many devices. Furthermore, communication technology and network promotes local devices networking. Although wire communication network is reliable and rapid, it has some unendurable disadvantages such as high cost, inconvenient wiring, easily destroyed by violence and so on. At present, many enterprises have workstations distributing in the field where the environment is very bad, which makes no supervision and long-distance control become absolutely necessary. Obviously, using wire as the communication medium is improper both in technology and in cost. That rapidly promotes the application and the development of wireless communication technology in the field of industrial control. GPRS (general packet radio service) (2.5G) is a kind of wireless packet switch technology based on GSM. It provides end-to-end wireless IP (Internet Protocol) connection for a wide area and has the features of speediness, "always connected", etc. GPRS is intermediate between GSM and 3G (the third generation) mobile communication. Since GPRS uses the access and transmission network of China Mobile, the supervision area will be wide and the cost will be low if we establish long-distance supervision system with GPRS technology. This paper has discussed how to extend the space of supervision to mobile networks and the Internet with the help of GPRS and TCP/IP.

2. THE MAIN FEATURES OF GPRS

1) Full usage of the existing resource - China Mobile’s countrywide telecommunication networks like GSM. Long-distance data input with convenience, speediness and low cost. Theoretical data transmission speeds up to 57.6 kilobits per second (kbps). And the maximum speeds are between 115kbps and 170kbps. That can meet the users’ need perfectly. The speeds of GPRS (3G) will be up to 384kbit/s.
2) Short contact time. The latency time for GPRS logging on is short and the connection is established quickly. It is only 2 seconds on average.
3) “Always connected”. The users can always be connected, which makes the access service become very simple and rapid.
4) Charge on flow. The users only occupy the resource when they send or receive data. They can always be online. The charge is based on the flow of data packets that users have received and sent. When there is no data transmitted, the users need pay nothing even if they are online.

Seen from the features stated above, GPRS networks are especially proper to the real-time transmission of small quantity of high-frequency flow. Long-distance industrial data collection system is a case in point [1].

3. HOW TO ESTABLISH COMMUNICATION NETWORKS IN GPRS SUPERVISION SYSTEM

Generally, there are two methods of establishing wireless digital communication system based on GPRS.

(1). Outer Network Mode (SM-MH, DH-MH)

Outer network mode refers to the mode where some components of the system are inside of the Internet. In order to support this mode, renting a special line to meet the need of operation is necessary. The flow of industrial data collection system is discontinuous, and some applications even have hours or days of an upload cycle. That means uneconomical and low usage of the resource if we rent a special line for transmission. The advantages of this mode are that it simplifies GPRS DTU (Data Transmission Unit) design and the whole system is relatively steady and reliable. Of course, you can access the supervision host in the way of ADSL dialup. But the system’s security is hard to guarantee except for the difficulty for dynamic IP addresses to access each other that will be discussed later. The architecture of outer network GPRS system is as Figure 1 illustrates.

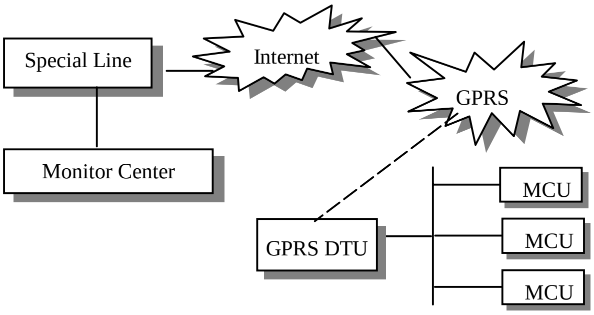


Figure 1 The Outer Network GPRS System

(2). Inner Network Mode (MH-MH)

Inner network mode refers to the mode where all of the important components of the system are inside of GPRS networks and none extend to the Internet. Obviously, the security of the system designed in this mode is much higher than that designed in the mode of outer networks. What's more, the design is much more convenient and efficient and the cost is much lower. The architecture of inner network GPRS system is as Figure 2 illustrates.

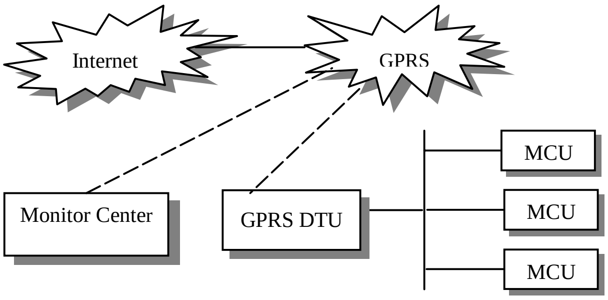


Figure 2 The Inner Network GPRS System

4. THE PROBLEMS OF DESIGNING SUPERVISION SYSTEM IN THE MODE OF INNER NETWORK AND THE SOLUTIONS

(1) The Problem of Dynamic IP Accessing Each Other

Seen from Figure 2, both the supervision hosts and the supervised devices logging on GPRS networks by dial-in service as the cell phones do. And the IP addresses provided by ISP are dynamically allotted, which creates an obvious problem that is how they can find each other. Even if this problem has been solved by manual telephone when the networks initialize, how can the system reestablish quickly by itself when all of the hosts and the clients are offline?

Our solution is: SMS (Short Message Service) + E-mail (Electronic Mail). It ensures that the system can automatically establish or resume the connections within a determinate delay. SMS is able to fulfill the communication conveniently among the components of the system. The disadvantages are that the real-time capability and the reliability can't be guaranteed, so we add E-mail service to the system.

The process of establishing network connection can be described as:

Clients online. At first, GPRS DTU (GPRS Data Transmission Unit) sends short massage to GPRS AS (GPRS Access Server) and reports its IP address. If GPRS DTU can't receive the response from GPRS AS within the determinate delay, it sends e-mail to a commercial e-mail server (e.g., yahoo) immediately. When GPRS AS doesn't receive message from GPRS DTU or the data collected by GPRS DTU within the routine-time and can't control GPRS DTU, it logs on the fixed e-mail server to get the IP address of GPRS DTU via e-mail. This solution can limit the communication delay to a determinate time slice so that the real-time capability of the system can be guaranteed.

(2) The Problem of the Supervision Host Offline and the Software's Automatic Redial:

If the problem of dynamic IP accessing each other were solved successfully, then could such GPRS DTU be put into service? The answer is negative. It is because the features [2][3][4][5] of GPRS networks make the offline phenomenon commonly exist when the supervision host accepts the devices log on, which adds a new demand to the networks designed in the mode of MH-MH. That is applying a static inner IP address for the supervision host from China Mobile, or it is hard to reestablish the system after the host is offline with the available products. The reason is that these products are lacking the uniform command of high-level protocols established according to the features of GPRS link and the system doesn't have the ability of dealing with abnormalities. Unfortunately, there are few cities that have started to provide static GPRS inner IP address service so far. Even if the static GPRS inner IP address service is available, the interruption of communication is unavoidable and the performance of the whole supervision system must be influenced greatly.

Seen from the above analyses, GPRS network's offline would bring many unfavorable factors to the system's stability. But the tougher problem is that the host can only get online again through manual dial once it is offline, because GPRS DTU cannot be aware that it has been offline. This problem can result in the failure of designing networks in the mode of MH-MH and the enterprises will have to establish GPRS supervision networks through a special line.

As for this problem, we adopt the following solution:
First, we divide GPRS data transmission terminals into two groups- GPPS DTU and GPRS AS (as Figure 3 shows).

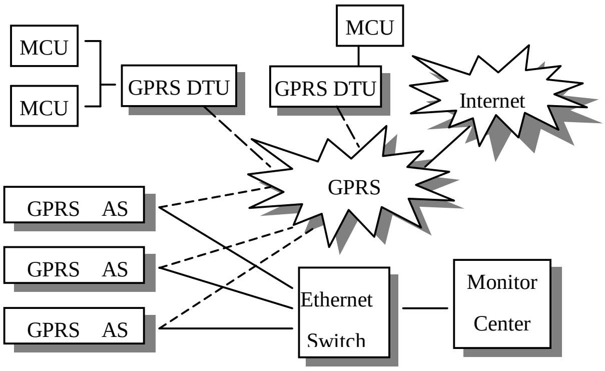


Figure 3 A Useful GPRS Detect System

The working process of the system can be described as:

Steady Transition: In the network topology in Figure 3, every GPRS AS runs self-defined RBP (Redundancy-Backup Protocol). When one GPRS AS is offline, the attached GPRS DTUs can switch to other GPRS AS because GPRS ASs possesses self-adapt mechanism. The GPRS AS that has received data from extra GPRS DTUs sends a signal to the offline GPRS AS promptly and asks it to dial again immediately. After the offline GPRS AS is back online, it will inform the overburdened GPRS AS of its IP address. And then the overburdened GPRS AS gives the new IP address of the offline GPRS AS to the extra GPRS DTUs. GPRS DTUs that have received the address start to send data to the original GPRS AS.

Smooth Transition (when all GPRS ASs are offline): Multiple GPRS ASs connect to each other through 10M/100M Ethernet, and intercommunicate their laden messages and all kinds of dynamic tables that will be described later. If the time that their burdens are all zero exceeds a minimum time limit, all GPRS ASs dial again one by one. And then through the similar strategy described above, the system allots the flow gradually and transits smoothly. In order to avoid the storm of short messages and misoperations, GPRS DTU redials after 2 or 3 minimum time limits since it can't communicate with any GPRS AS, and contact with the certain GPRS AS directly. In the process of steady transition, the switch is very natural. The system's capability is influenced slightly and the data flow of the system increases very little, which goes on that the offline cycle of GPRS AS is longer than the time used by GPRS AS dialup. The probability that the process of smooth transition happens is very low and the flow increases slightly. There is a short delay before the system resumes normality.

(3) The Limitation to the Capability

In fact, this problem is solved at the same time as the problem of the supervision host offline is solved. When users want to expand the capability, they just need to add more GPRS ASs to the system. The more GPRS ASs there is, the stronger the system is.

(4) Some Technique Measures in Enhancing Security

On the basis of analyzing the working principle of the OTP[6] (One-Time Password) system, we plan to enhance data's reliability by the following steps:

A. The format of long-distance instruction is ID (16bit) + operative code (16bit). ID is the symbol of the controller's identity and operative code is the encoding to the operation.

B. Establish special forms consisting of server ID, encryption key (32bit) and the combining measure in GPRS ASs. And establish the corresponding forms consisting of server ID, encryption key and the contrary combining measure in GPRS DTUs.

C. Define a set of encryption algorithms proper to MCU (supposed the amount is N) and the corresponding decryption algorithms. They are respectively saved in GPRS ASs and in GPRS DTUs.

D. Data sender combines the original data with the corresponding encryption key to produce combined data packet.

E. Data sender stochastically produces a sequence number between 1 and N, and then uses the corresponding algorithm for this number to encrypt the combined data packet, and let the number become the sequence number of TCP data packets. The data receiver finds the corresponding decryption algorithm according to the sequence number.

F. After decrypting twice, the receiver can get the original data- ID + op code, and then check if the ID is the same with the local ID. If not, discard this packet and send N-ACK packet to the sender. Otherwise, check if the op code belongs to the local table of operative codes. If it does, operate. Otherwise, send N-ACK packet. When the sender receives N-ACK packet, it will choose a different encryption key and a different combining algorithm and then redo D-F.

G. If the instruction is remote, send multiple data packets and ensure the correctness of data transmission by checking back in the receiver. In order to avoid some packets becoming the reference of decryption after they are intercepted, every original piece of data must experience the steps of D-F.

This measure has the features as: Reject complex encryption but invent encryption collection proper to MCU (Micro Control Unit) according to the features of SCM (Single Chip Microcomputer). Create encryption, combining algorithm and encryption key dynamically, which makes the original data have different data packets when it transfers in the networks. The transmission information that should have been transmitted directly is hid in the header of TCP packet so that the hacker won't even notice.

At least, this encryption ensures the integrity of data when it transmits in GPRS links. When a damaged data packet arrives, the rejected rate, *P*, is given by Eq. (1):

$$P = 100 * (1 - 4 * 20 / 2^{32})\% \qquad (1)$$

Note: Supposed that there are 4 GPRS ASs and 20 kinds of remote operations.

5. SUPERVISION SYSTEM DESIGN AND REALIZATION

According to the measures above, we have designed and implemented a set of supervision system based on GPRS.

(1) GPRS DTU (GPRS Data Transmission Unit)
Hardware: 8051, MC35 (Siemens GPRS Modem)
Embedded develop environment: μCos/II
Protocols mended: TCP/IP, PPP, POP3 and SMTP. Through simplification and expansion, the protocols possess determinate intelligence so that GPRS DTU can automatically resume the communication with the supervision host once the supervision system is offline.
Function realized: Access devices to GPRS networks. Transmit collected data in multiple working modes, such as "always connected", timed transmission, AS calling, data driving and power saved. If the communication is abnormal, establish the contact with GPRS AS and the administrator rapidly. Receive and perform the instruction sent by the supervision host and GPRS AS. Support embedded program control. The architecture of GPRS DTU is as Figure 4 illustrates.

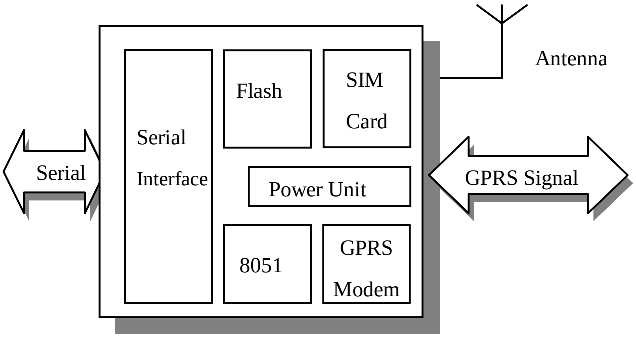


Figure 4. The Architecture of GPRS DTU

(2) GPRS AS (GPRS Access Server)

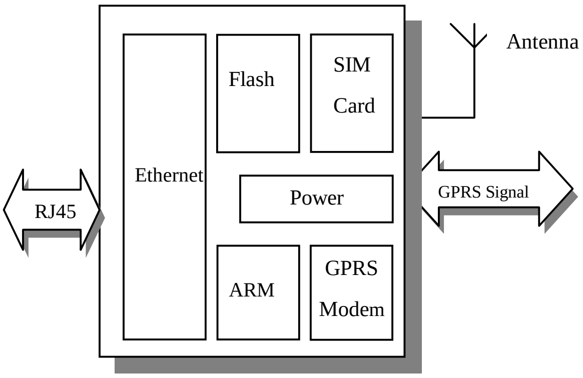


Figure 5 The Architecture of GPRS AS

Hardware: S3C4510B, MC35
Embedded develop environment: µClinux
Protocols mended: TCP/IP, PPP, POP3, SMTP and RBP. These protocols have also been mened and possess determinate intelligence. They can work regularly under the control of established strategy. The whole supervision system is quite steady and dependable.
The main function realized: Work as the communication center of the whole supervision system. Take in charge of long-distance log-on of the devices. Assort with all parts and get them right when the system communication is abnormal. Balance and allot the flow among multiple GPRS ASs. Allow multiple GPRS ASs to work for one or multiple supervision hosts. Establish forms consisting of GPRS DTU IP addresses and the devices' ID. Realize the transparent data transmission between the supervision host and GPRS AS. Reserve the connection message for every GPRS DTU automatically, and realize the virtual persistent connection between GPRS AS and GPRS DTU.

(3) The SMS model in the supervision host
In this system, the supervision host can send the important data collected to the administrator's mobile phone after the form of short message. In the supervision host, we've still established short message bank and the corresponding management tools.

Table 1

Mode \ Item		Connect Expenses of each Host (RMB Yuan per month)	Special Service of Mobile or ISP	Possibility of Being Served	Devices Scale	Supervision Host Offline	Establish Cost (relative)	Security
SM-MH (Special Line)		2000 (Per 64K)	Special Line	Generic	Unlimited	No	High	High
DH-MH (ADSL)		600	Shared Dynamic IP	Generic	Unlimited	Yes	Relative High	Low
MH | MH	Present Project	200	Inner Static IP	Small	Limited	Yes	Middle	High
	This System	**200 - 400**	**None**	**100%**	**Unlimited**	**No**	**Low**	**High**

(4) The Comparison in Technique Capability and in Cost between this System and the Present Projects (as Table 1 Shows)
Note: SM: static host, DH: dynamic host, MH: mobile host. The quotes come from the subsidiary company of China Mobile in Yichang.

Through the simple comparison in table 1, we can find out that only SM-MH can be put into service now. The stability and dependability of our system can be compared with SM-MH. And our system shows a lot of preponderance, when it comes to the convenience of establishment, the scale of devices and the cost. Seen from the working cost, with the same investment our system may accommodate devices 10 times more than the SM-MH does. When lack of burden, SM-MH will lose predominance thoroughly. The cost for SM-MH to expand capability is very high. Using our system, users can decide whether to expand or not. They can expand the capability freely. And there is nearly no limit and the extra cost is quite low. So our system would possibly be used widely in medium and small enterprises when they establish GPRS supervision networks.

6. CONCLUSIONS

This paper has made an in-depth discussion about the most important questions of establishing GPRS industrial supervision system. An actual supervision system has been designed. The further work is to expand its capability so that it can possess the function of VPN (Virtual Private Network) and can provide technique support for the enterprises to establish supervision system cross province.

7. REFERENCES

[1] Feng Dongqin, Jin Jianxiang, He Jian, "Ethernet and Industrial Control Network", Instrument Academic, No 1, 2003, pp. 600-603. (in Chinese)

[2] Zhe Wei, The analyses and improvement to TCP capability in lineal environment, Master's Thesis of Beijing Posts and Telecommunications Engineering University, March 2001. (in Chinese)

[3] Du Jirong, The analyses and improvement to TCP in GPRS environment, Master's Thesis of Zhejiang University, March 2002. (in Chinese)

[4] Zhu Jing, Niu Zhisheng, "The analyses and service quality control to TCP in unreliable network environment" , Electronic Academic, November 2000. (in Chinese)

[5] Liu Haipeng, Zhang Gendu, Li Ming, "The technology summary about enhancing TCP capability in wireless mobile networks", Computer Research and Development, Vol. 39, No 6, 2002, pp. 641-648. (in Chinese)

[6] N.Haller, C.Metz, "A One-Time Password System", RFC2289, IETF,February 1998.

[7] Jean J. Laborosse, MicroC/OS-II The Real-Time Kernel, Second Edition, Lawrence: CMP Books Inc. Pub., 2002.

[8] Andrew S. Tanenbaum, Distributed Operating Systems, Beijing: Tsinghua University Press, 2000.

Huie Chen is working for a master's degree of Embedded Operating System. She graduated from Three Gorges University in June, 2003 with specialty of computer science & technology and was granted a privilege to enter its graduate school without entrance examinations in Oct., 2002. She is fully developed morally and has been awarded "three As' student" annually at college. She still got a second in the 2nd "English Weekly Cup" National Internet English Writing Competition in 2002; a third in the 3rd in 2003. Her research interests are in RTOS, distributed processing, networks and information technology.

Distributed Virtual Reality Environments Based on VRML

Fang Hua , Yaolin Gu
School of Information Technology, Southern Yangtze University, Wuxi, Jiangsu, China
Email: hf_daisy@etang.com **Tel:**+86(0)13812285815

ABSTRACT

Nowadays computer graphics technology has been developed to create and simulate Virtual Reality Environments. Meanwhile more and more researchers are focusing on the extensibility and interaction of VEs. Ideally Virtual Environments should be dynamic, mutable and attractive for immersed users. As such environments can be designed easily by Virtual Reality Modeling Language (VRML), here we propose a distributed Virtual Reality (VR) system that is based on an interactive animation system using VRML for geometric and behavioral modeling. The emphasis is on concepts and extensions for the integration of user immersion and interaction, and system extensibility into a VRML-based animation environment. In this paper we have used VRML 2.0 Version to implement the modeling of a virtual theater with several desks and chairs, a rostrum, a specialized computer with multimedia devices, fans and fluorescent lamps, as well as interactive avatars. The case study served here illustrates the proposed concepts and extensions, and helps us elicit the conclusion that any Virtual Environment has its great system interactivity and extensibility and user immersion, and can be a VRML-based Distributed Virtual Reality Environment.

Keywords Virtual Reality Modeling Language, Distributed Virtual Reality Environment, computer interaction, extensibility, computer graphics

1. INTRODUCTION

The term “virtual world” (a.k.a. “virtual reality”, “virtual environment”) generally refers to a human-computer interface allowing users to experience a computer-generated environment that is interactive and three-dimensional. The virtual world and its objects can be represented by 3-D stereoscopic images and sounds. They are directly manipulated by a person through hand or body movements, and sometimes spoken words. Objects in the world may be linked to collections of data or concurrently running simulators [1], [2], [3], [4]. The technology brings together diverse elements including networked systems (possibly heterogeneous), specialized input and output devices, 3-D graphics, hand or body gesture and speech recognition, data visualization, distributed processing, simulation and the possibility of more than one user sharing the environment [4], [5].

In recent years, technological development has provided powerful computers and networks enabling us to create complex immersive and interactive virtual environments for games, data visualization, and simulation. There exist many excellent software tools for creating VR systems. Today, most virtual environments are designed manually and created by means of interactive mouse-based editors. Handcrafted worlds are simple to build and the quality of such worlds can be quite high. The resulting environments are mostly static in the sense that only very few of the objects within them change over time, such as user avatars and scripted objects. In [6], however, the authors state that, in order to make virtual environments compelling and attractive, they must be mutable. As is known, current immersive virtual environment systems use a wide variety of programming methodologies and scene description languages ranging from textual scene description languages to visual ones within 3D immersive virtual environments. One of the approaches proposed in [6] to make mutability feasible is “designer change”, which means a special effort from designers is required to create the dynamic environments.

Modern 3D graphics systems allow a rapidly growing user community to create and animate increasingly sophisticated worlds [7]. TBAG presented in [8], for example, is a toolkit for rapid prototyping of interactive, animated 3D graphics programs. The narrow role of modifiable state in the paradigm allows applications to be run in a collaborative setting without modification.

Another toolkit – Virtual Reality Distributed Environment and Construction Kit (VR-DECK) presented in [9] – is an object-oriented system utilizing distributed message passing for communication. Rewriting Systems (a.k.a. production systems, Lindenmayer-systems) in [10], [11], [12] is also efficient technique, especially in modeling fractals or self-similar objects.

In this paper, we propose a concept for a Distributed Virtual Reality Environment (DVRE) based on Virtual Reality Modeling Language (VRML), which explicitly supports interactivity and mutability requirements. As the pivotal techniques of VR are modeling and simulation, VRML [13], [14] is a Web standard. Its first incarnation, VRML 1.0, was a static 3D scene description language where users could only move the point of view. Its successor, VRML 2.0, is an open standard and introduced animation and programmable behavior. In [14], VRML is employed in Distributed Virtual Environments (DVEs). The authors describe how to use the approach they presented to provide general DVE support and its use in implementing VRML DVE architecture.

The core of our work is to show that VRML-based interactive systems can also be used as a DVRE system and to describe how it can be done. In Section 2, we introduce the concept of DVREs and its significance in computer graphics. Then we focus on particular requirements for DVREs. In Section 3, we discuss the language, i.e. VRML necessary and useful for Distributed Interactive VEs (DIVEs)[15]. Finally, in Section 4, a virtual theater paradigm is presented to serve as a case study and to illustrate the proposed concepts.

2. DISTRIBUTED VIRTUAL REALITY ENVIRONMENTS

DVRE is particularly efficient on modeling large-scale interactive virtual systems. In the next subsection, we present the concept of DVREs. Later, we focus on requirements and architecture necessary for DVEs.

2.1 Concept of DVE/DVR

What is "Distributed VR"? The idea behind DVR is very simple: a simulated world runs not on one computer system, but on several. The computers are connected over a network (possibly the global Internet) and people using those computers are able to interact in real time, sharing the same virtual world [16]. Today, there are two main research directions on it: one is the DVR on the Internet, e.g., tele-virtual shopping based on VRML; the other is dedicated networks invested by the military, e.g., American military DSI using ATM.

2.2 General Requirements for DVE applications

In our work on DVEs we aim to support the following requirements:

Interaction: objects should be activated through the interactive devices such as Space Ball, Mouse, Keyboards, Data Gloves, Speech Recognition systems, or FOBs, and make corresponding actions.

Animation: it achieves the animation effect ensuring the avatars in VEs are mutable and dynamic.
Immersion: it makes users have a perception that they were in the truly environments.

Extensibility: it should be possible to add to or modify an existing DVE by means of embedding small-scale objects into large ones. Note that the key point we use here is called "modular". It means each object in VEs responds to one module [9]. Whenever one module needs extended, it includes what should be included. Therefore, a layered architecture is needed as well.

Content-independence: graphics and geometry are important, but they are not the only data a DVE is concerned with. All forms of data should be treated equally [17].

2.3 Architecture

As is mentioned above, our implementation of DVE support follows a layered architecture. So an application development environment is supposed to be modular and dynamic [9]. By modular we mean that at the bottom exist "unit modules" which offer the smallest objects built for reuse. Then some "mid-scale modules" consisting of several unit modules are above them. The top layer describes the whole environment to build. It is similarly composed of several mid-scale modules. In short, modules can be thought of as the building blocks of a virtual world, and they can be nested. Note the modules should be self sufficient and capable of operating independently as well as in conjunction with other ones. Fig. 1 shows the hierarchy model.

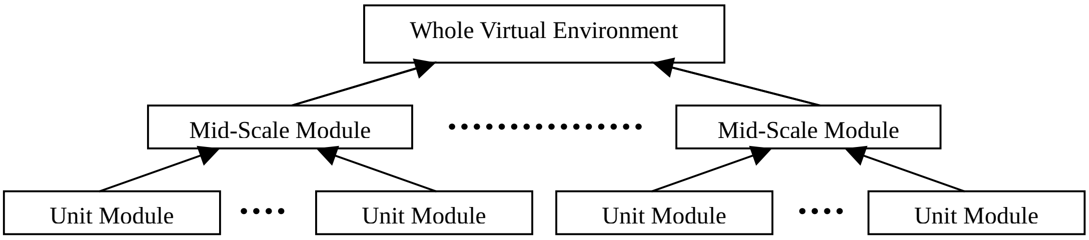


Fig. 1 The Layered Architecture for Modeling A DVE

3. VIRTUAL REALITY MODELING LANGUAGE

VRML is an authoring standard for 3-D content on the Internet [13], [14]. It defines a number of nodes that describe 3D geometrical objects, materials to be applied to these objects, light sources, cameras, transformations and so on. These nodes may be connected together by "*routes*" along which information, in the form of VRML events, can flow, e.g. the output of a *TimeSensor* node could be routed into a *Transformation* node to create moving geometry. In VRML, there are nodes used commonly as follows: Appearance, Background, Geometry, Group, Inline, Material, NavigationInfo, PointSensor, Script, Shape, Sound, TimeSensor, ViewPoint, and many others that are used commonlessly [18], [19].

While implementing the environments with VRML, the user just deals with familiar entities such as geometric shapes, and operations like translations and rotations. In addition, the core set of nodes may be extended through the use of a native prototyping mechanism. Objects designed using 3D-MAX or other tools may be imported and used [18] as well.

When writing codes in VRML, users can select almost any kind of text editor such as Notepad, WordPad, Microsoft Word, and some dedicated editors for VRML, e.g., VrmlPad. Users can also use Internet Space Builder (ISB) as a visual toolkit.

VRML worlds can include content from URLs, and contain hyperlinks to other VRML worlds or any URL. VRML has its own mime type, and is normally viewed using a Web browser plug-in such as Intervista's WorldView or SGI's CosmoPlayer. This makes a VRML world appear very much like a JPEG image or Java applet [13] from a user's perspective.

In [18], we have seen features of VRML when applied in DVEs: perfect readability, transplantability, and maintainability; platform-independent; and its reality and interaction. Note again, VRML supports user interaction and immersion, as well as system animation and extensibility by employing some given nodes, in particular, the *Inline* node and the *Script* node with scripting languages, i.e., JavaScript or VrmlScript.

4. A THEATER PARADIGM

Now we present a theater paradigm based on VRML to illustrate the concepts and extensions proposed above. Here we use VrmlPad 2.0 as our editor and Cosmo Player as browser.

We suppose the theater include several nested equipment which makes it possible to achieve various functions both basic and multimedia. For instance, while immersed in the theater the avatar can walk around and visit its configuration. He can sit on the chair, observe the lantern slide or the entire space of classroom, turn on/off the slide projector, optionally manipulate the computer device connected with multimedia device by switching the display interface, adjust the position of desk lamp and the direction and luminance of lamplight, control the power source of desk lamp as well as the fans on the ceiling, manipulate the keyboard, adjust the distance between the computer and the avatar by moving or rotating the chair, and so on. Besides, some extended effects are added in the system, e.g., avatars can "enter" into the computer and observe its internal structure, seeing CPU, main memory, etc. Thus users would absolutely sense the interaction, immersion and animation of virtual environment.

While analyzing and designing it, we have also considered its content-independence. For doing this, we put emphasis not only on the entities but the sense of reality. By immersion and interaction, we add some given nodes such as TimeSensor and TouchSensor; by animation we use some other sensor nodes and Script nodes embedded with JavaScript or VrmlScript; and by extensibility, a layered architecture with different modules is exploited here. Fig. 2 shows the hierarchy we proposed:

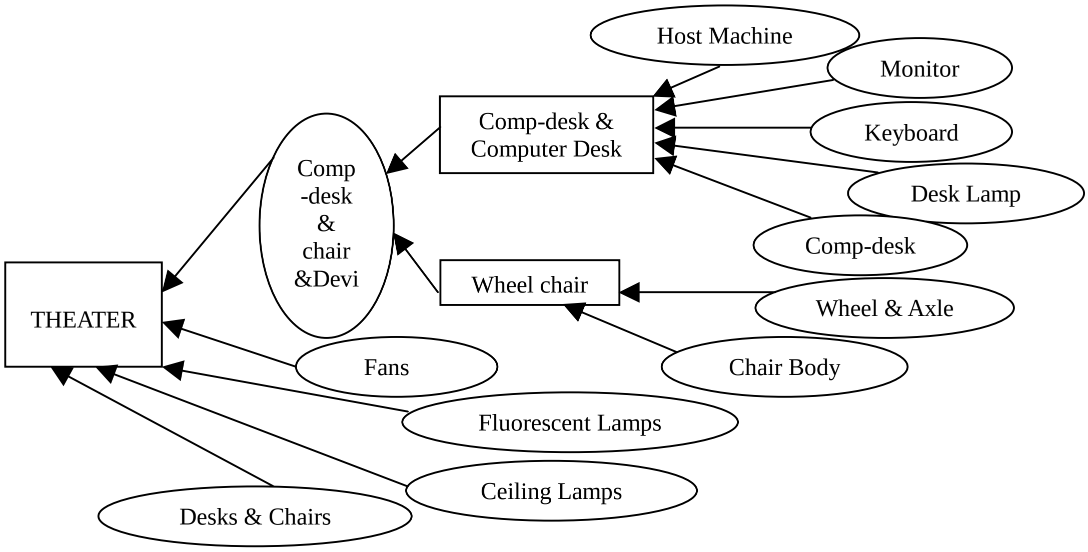


Fig. 2 Layered Architecture Designing for the Virtual Theater (Left->Right Represented as Top->Bottom Described Above)

We now present some resultant scenes in Fig. 3 and Fig. 4 as follows:

Fig. 3 Structure in Computer

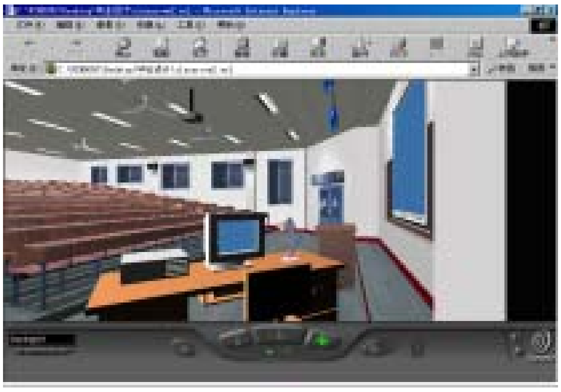


Fig. 4 One Scene in Virtual Theater

5. CONCLUSION AND FUTURE WORK

We have successfully demonstrated the feasibility of a distributed VR system based on VRML. VRML is a very convenient language of programming for visually interactive and dynamic 3-D environments and animations. It offers a large potential to satisfy the mutability paradigm for extensible virtual environments.

In the future, we prospect to have the ability to perfect the heater by adding more timed actions and some proper sound events.

6. REFERENCES

[1] Zeltzer, D., Pieper, S. and D. Sturman, "An Integrated Graphical Simulation Platform", ***Proceedings of Graphics Interface '89***, pp. 266-274, 1989.

[2] Appino, Perry A., Lewis, J. Bryan, Koved, Lawrence, Ling, Daniel T., Rabenhorst, David A. and Codella, Christopher F., "An Architecture for Virtual Worlds", ***Presence***, vol. 1, no. 1, 1992.

[3] C. Shaw, J. Liang, M. Green and Y. Sun, "The Decoupled Simulation Model for Virtual Reality Systems", ***CHI '92 Conference Proceedings***, pp. 321-328, ACM, May 1982.

[4] C. Codella, R. Jalili, L. Koved, J. B. Lewis, D. T. Ling, J. S. Lipscomb, D. A. Rabenhorst, C. P. Wang, A. Norton, P. Sweeney and G. Turk, "Interactive Simulation in a Multi-Person Virtual World", ***CHI '92 Conference Proceedings***, ACM, May 1992.

[5] Blanchard, C., Burgess, S., Harvill, Y., Lanier, J., Lasko, A., Oberman, M. and M. Teitel, "Reality Built for Two: A Virtual Reality Tool", ***Proceedings of the 1990 Symposium on Interactive 3D Graphics (Snowbird, Utah)***, pp. 35-36, New York: ACM, 1990.

[6] B. Anderson and A. McGrath, "Strategies for Mutability in Virtual Environments", ***Virtual Worlds on the Internet***, J. Vince and R. Earnshaw, eds., pp. 123-134, Los Alamitos, Calif.: IEEE CS Press, 1998.

[7] E. Gobbetti and J. F. Balaguer, "An Integrated Environment to Visually Construct 3D Animations", ***Proc. SIGGRAPH***, 1995.

[8] C. Elliott, G. Schechter, R. Yeung, and S. Abi-Ezzi, "TBAG: A High Level Framework for Interactive, Animated 3D Graphics Applications", ***Int'l Conf. Computer Graphics and Interactive Techniques, Proc. 21st Ann. Conf. Computer Graphics***, pp. 421-434, July 1994.

[9] C. F. Codella, R. Jalili, L. Koved, and J. B. Lewis, "A Toolkit for Developing Multi-User, Distributed Virtual Environments", ***Proc. IEEE Virtual Reality Ann. Int'l Symp.***, pp. 401-407, 1993.

[10] H. Noser, C. Stern and P. Stucki, "Distributed Virtual Reality Environments Based on Rewriting Systems", ***Proc. IEEE Transactions on Visualization and Computer Graphics.***, pp. 213-225, vol. 9, no. 2, April-June 2003.

[11] P. Prusinkiewicz, M. S. Hammel, and E. Mjolsness, "Animation of Plant Development", ***Computer Graphics Proc., SIGGRAPH '93, Ann. Conf. Series***, p. 351, 1993.

[12] H. Noser and D. Thalmann, "The Animation of Autonomous Actors Based on Production Rules", ***Proc. Computer Animation '96***, pp. 47-57, June 1996.

[13] B. Roehl, J. Couch, C. Reed-Ballreich, T. Rohaly, and G. Brown, ***Late Night VRML 2.0 with Java***, Emeryville, Calif.: Ziff-Davis, 1997.

[14] Mike Wray and Rycharde Hawkes, "Distributed Virtual Environments and VRML: an Event-based Architecture", HP Labs (Bristol), Filton Road, Bristol, BS12 6QZ, UK. ***In Proceedings of the Seventh International WWW Conference (WWW7)***, Brisbane, Australia, 1998.

[15] Carlsson C. and Hagsand O. (1993) DIVE - a Multi-User Virtual Reality System. ***Proceedings of the IEEE VRAIS '93 Conference***: 394-400.

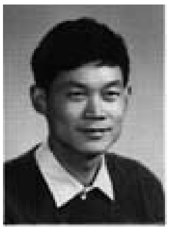
[16] Bernie Roehl, "Distributed Virtual Reality -- An Overview", June, 1995, http://ece.uwaterloo.ca/~broehl/distrib.html

[17] Hawkes R. (1996) "A Software Architecture for Modeling and Distributing Virtual Environments". Ph.D. Thesis, University of Edinburgh.

[18] VRMLSite Managazine(VRML Java VRML more VRML and VRML Virtual Worlds 3D Virtual Reality)： (http://www.vrmlsite.com/)

[19] Huotari, J.|Niemela, M. "Enhancing graphical information system models with VRML" 2002

Hua Fang is a graduate student in School of Information Technology, Southern Yangtze University.



Gu Yaolin : A Full time Professor and Deputy Dean of the School of Information Engineering , Southern Yangtze University, P. R. China. He graduated from the Computer Science Department of Shanghai Jiaotong University in 1982. As a Visiting Scholar of the University of Stony Brook in USA from 1994 to 1995, his main research interests are Computer Graphics and Parallel Computing. Now he is a Senior Member of Computer Society, IEEE.

A Method of Computing Fractal Dimension*

Ren Wei [1], Liu Dan [1] and Xie Ling [2]
[1] Institute of Nautical Science & Technology, Dalian Maritime University
[2] School of Electronics & Information Engineering, Dalian University of Technology
Dalian, Liaoning Province, 116026, China
E-mail: monsoons@126.com **Tel:** (86) 0411-84729651

ABSTRACT

To describe the curves of the nature in computer vision system, curve interpolation is necessary. Fractal curve interpolation could create lots of complicated curves or surfaces. Fractal Dimension is also one of the important parameters for describing the scene's Fractal structures; it has been used in some field such as Computer Vision, Digital Image Processing. A method to compute Fractal Dimension is given in this paper.

Keywords: Box Dimension; Douglas-Peuker Algorithm; Algorithm of random midpoint displacement.

1. INTRODUCTION

In practice, we usually could get some observational data or random data arrays. To find more details of the curves, we must add sufficient points based on the original curve or surface, and the local feature should be presented precisely. Data interpolation could use polynomial method, spline method, least square method, finite element method (FEM) etc. Using the frontier methods, no matter what the curve is complicated, if magnifying the curve, we will find that every little area of the curve is a line. There are some complicated or abnormity curves in the nature, it is difficult to be described by the traditional interpolation method.

In the 70's, Benoit B. Mandelbrot created a new branch called "Fractal" in the mathematics field. Using Fractal method, we could protract some complicated and rough curves or surfaces.

Fractal Dimension has been found useful in describing the "complexity" of random spatial patterns. [1][7] There are various definitions of a "Fractal Dimension" proposed in the literature, such as Box Dimension, Hausdorff Dimension, capacity Dimension and packing Dimension. They may or may not have the same value for a given Fractal depending on the properties of the Fractal. [2]

Estimating of the Dimension of a Fractal has become an important and interesting statistical problem. [3] The Box-counting technique is commonly used to estimate the Fractal Dimension because of its simplicity.

A method to count Box Dimension is given.

2. METHOD AND THEORY

2.1 Definition of Box Dimension

The similarity dimension is meaningful only for exactly self-similar sets. For more general sets, including experimental data set, it is often replaced by the Box Dimension. For any bounded (nonempty) set A in E-dimensional Euclidean space, and for any $\delta > 0$, a δ-cover of A is a collection of sets of diameter δ whose union contains A. Denote $N_\delta(A)$ the smallest number of sets in a δ-cover of A. Then the Box Dimension d_{box} of A is

$$d_{box} = \lim_{\delta\to 0} \frac{\log(N_\delta(A))}{\log(1/\delta)} \tag{1}$$

When the limit exists. When the limit does not exist, the replacement of lim with lim sup and lim inf defines the upper and lower Box Dimensions:

$$\begin{aligned} \overline{d_{box}} &= \lim_{\delta\to 0} \sup \frac{\log(N_\delta(A))}{\log(1/\delta)} \\ \underline{d_{box}} &= \lim_{\delta\to 0} \inf \frac{\log(N_\delta(A))}{\log(1/\delta)} \end{aligned} \tag{2}$$

The Box Dimension can be thought as measuring how well a set can be covered with small boxes of equal size, because the limit (or lim sup and lim inf) remain unchanged if $N_\delta(A)$ is replaced by the smallest number of E-dimensional cubes of side δ needed to cover A, or even the number of cubes of a δ lattice that intersect A. Section V describes methods of measuring the Box Dimension for physical data sets. [4]

2.2 Douglas-Peuker Algorithm

The Douglas-Peuker algorithm could estimate Fractal Dimension. The main method of the algorithm:

1. Connecting the first point and the last point of the curve by a dashed line. Computing the distance of each point to the dashed line(d_1,d_2, $\cdots$,d_n). A threshold value ε should be provided.
2. Computing the value of d_i related of ε, if the value is negative, then remove the point, otherwise, the point should be reserved.
3. Connecting the reserved point after step2. Repeating the step1 and the step2 until all the distances of the points to the dashed line are bigger than the threshold value.

Figure 1 is the example, given the threshold value ε, the

* This paper is supported by Fok Ying-tong Education Foundation.(91077)

point 4 is removed, then the new curve is created.

Changing the threshold value, each value should create its own curve and points array. Then the Fractal Dimension of the curve could be calculated. [5]

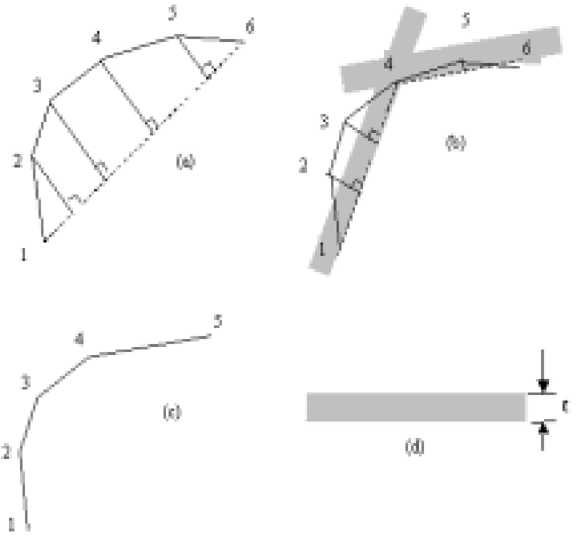


Figure 1 The algorithm of Douglas-Peuker method

2.3 Algorithm of random midpoint displacement

Algorithm of random midpoint displacement is a kind of recursive algorithm. Connecting the points, get a line, move the midpoint up and down. Connecting the new points as a new curve. Repeat the steps.

Figure 2 is an example. Different interpolative curve would be created after many times of iterations.

Actually, Algorithm of random midpoint displacement is a contrary process to Douglas-Peuker Algorithm. [6]

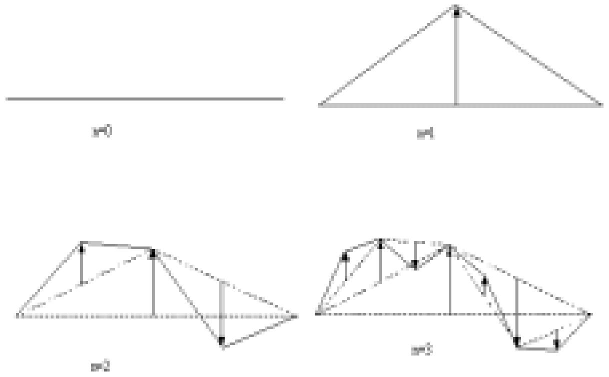


Figure 2 The algorithm of random midpoint displacement method

2.4 Computing the Box Dimension

To compute the Box Dimension of the curve or the surface, we could set an appointed area. Figure 3 is an example, the area’s size is 512×512.

Putting the curve in the area, when the scale is 512, the curve will be in the area. Changing the scale to 128, the curve will be in 4 rectangular areas. If changing the scale to 32, the curve will be in 16 rectangular areas. Changing the scale to 8, the curve will be in 48 rectangle areas. Then we could get the relation between the scale and the number of the rectangular areas the curve located in.

Step1. Recording the area bound as a rectangle, the scale should be recorded at the same time.

Step2. Testing each of the rectangular area. If the point is in a rectangular area, recording the area, the counter adds one.

Step3. Reduce the scale to half, repeating the step 1 and step 2.

The data array including the scale and the number of the rectangular areas the curve located in should be acquired.

According to the equation (1), $\frac{\log(N_{\delta}(A))}{\log(1/\delta)}$ could be computed. Then the Box Dimension of the curve could be fitted by the least square method.

Actually, it is difficult to count the Box Dimension by the algorithm of Douglas-Peuker with a computer. It is easy to be counted by the upper method, but the speed of this algorithm is not too fast. If the image is big, or needing many times of iterativeness, it would cost us so much time to wait for the result.

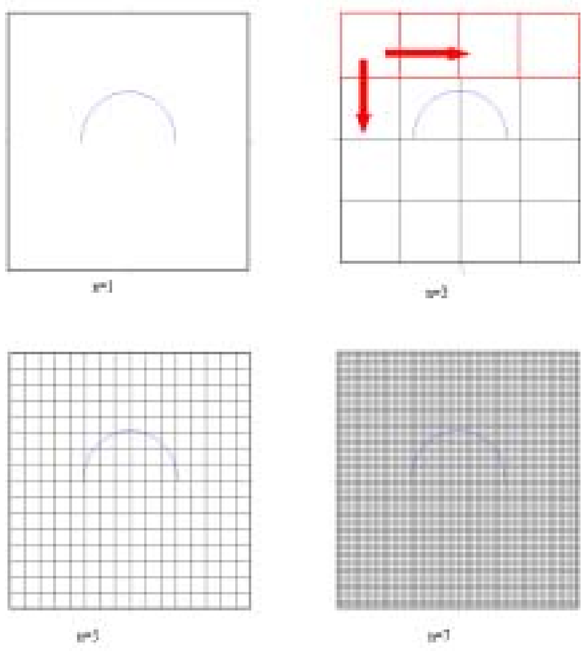


Figure 3 Computing Box Dimension

In Figure 3, the result of n times of iterativeness has been computed by testing the point fill in the designed area or not.

3. CONCLUSIONS

To simulate the curves of the nature in the computer vision, Fractal has provides a good method to do this. Fractal Dimension is also one of the important parameters for describing the scene’s Fractal structures. To create a Fractal curve, computing the Fractal Dimension of the curve is the first thing.

Box Dimension counting is popular problem in Fractal field. In this article, the author gives a method to compute the Box Dimension; the algorithm could be executed in personal computer. But the algorithm will cost much of time when it processes a big image. So it is critical to improve the speed of the algorithm.

4. REFERENCES

[1] C. Taylor, S. Taylor, Estimating the dimension of a Fractal, J. R. Statist. Soc. B 53 (1991) 353-364.

[2] C. Cutler, A review of the theory and estimation of Fractal Dimension, in: H. Tong (Ed.), Dimension Estimation and Models, World Scientific, Singapore, 1993, pp. 1-107.

[3] A.J. Roberts, A. Cronin, Unbiased estimation of multi-Fractal Dimensions of finite data sets, Phys. A 233 (1996) 867-878.

[4] Benoit B. Mandelbrot Michael Frame Fractals Encyclopedia of Physical Science and Technology.

[5] Guohua Zhang etc. A new Fractal interpolation method of the geographical lines. Mapping transaction. 2002. August. Vol 31 No.3 (Chinese Edition)

[6] Fangrong He etc. A summarization of 3D Fractal terrain creating technology. Wuhan chemical academe transaction. 2000. August Vol 24 No.3 (Chinese Edition)

[7] Jin Yi-wen , Lu Shi2jie. Fractal Geometry Theory and Application[M] . Hangzhou; Zhejiang University Press ,1998. (Chinese Edition)

Ren Wei is a graduate student of Institute of Nautical Science & Technology in school of Dalian Maritime University Dalian, Liaoning Province, China. His Special is Traffic Information Project & Controlling, and his researching field is: VR (Virtual Reality), Computer System Simulation.

Adaptive Partition and Hybrid Method in Fractal Video Compression*

Wang MeiQing [1] **Liu Rong** [2]
College of Mathematics and Computer Science, Fuzhou University
Fuzhou，Fujian 350002，China
Email: [1] mqwang@fzu.edu.cn ,[2]liu_r@fzu.edu.cn **Tel:** 0591-7543785

ABSTRACT

Fractal image compression is a relatively recent image compression method. Although it does not work as good as the state-of-the-art compression technology, its main advantage that the decompressing algorithm is very simple，makes it suitable for the situation of one encoding and many decoding, such as video on demand, archive compression, etc. There are two basic fractal compression methods, namely the cube-based and the frame-based methods, being commonly used in the industry. However there are advantages and disadvantages in both methods. This paper discusses the two basic fractal and other algorithms extend from them. Experimental results show that the algorithm based on adaptive partition can obtain a much higher compression ratio compared to the algorithm based on fixed partition while the qualities of decompressed images are similar, and the hybrid algorithm improves the compression ratio and the quality of decompressed images.

Keywords: Fractal, video compression

1. INTRODUCTION

Fractal image compression is a relatively recent image compression method developed in the late eighties [1]. It reduces the redundancy of images by using the self-similarity properties of images, the main advantage of decompressing a fractal compressed image only needs to compute the fixed point of a fractal transform operator equation, and it can decompress to any image resolution quickly, which is very simple and suitable for the situation of one encoding and many decoding.

As far as video compression is concerned, there are two basic fractal compression methods being used most frequently. One is being known as the cube-based compression [3] and the other is frame-based compression [2]. In the cube-based compression, a sequence of images is divided into groups of frames, each of which in turn is partitioned into non-overlapped cubes. The compression code is computed and stored for every cube. Although this method was proposed in 1993, it has such a high computing complexity that it was difficult to implement due to the limit of the computer at that time. In the frame-based compression, the compression code is computed and stored for every frame, but intra-frame or inter-frame self-similarity may be used. The former method may be used to obtain high quality decompressed images, but the compression ratio is relatively low; the latter method may be used to obtain a high compression ratio, but the current frame is relate to the previous frame which introduces and spreads errors between frames.

In [9], a hybrid compress algorithm was proposed, which consists of an interaction of the frame-based algorithms and the cube-base compression algorithm. In [10], the cube compression algorithm based on adaptive partition was proposed, which can obtain a much higher compression ratio compared to the algorithm based on fixed partition while the qualities of decompressed images are similar. In this paper we propose an adaptive hybrid algorithm, which combine the adaptive partition to the hybrid compression method and compare these different fractal video compression methods with videoconference images and movie images. Experimental results show that the new method can obtain more high compression ratio.

The paper is organized as follow. First the mathematic theory for the fractal compression of video sequence is presented. Second two basic fractal video compression algorithms are given. Third the hybrid method and adaptive method for motion images are discussed. Then an adaptive hybrid method is proposed. Last section lists the experimental results of various kinds of methods on two sequences of images. The first sequence of images is related to a videoconference and the second sequence of images is an extract from a film.

2. THE FRACTAL COMPRESSION THEORY OF VIDEO SEQUENCE

A digital image f of size $N \times M$ can be expressed as a matrix Q, where $Q(i, j)$ is the intensity of the pixel at the location (i, j) . Especially, a video sequence of s frames $Seq = \{ f_t, 1 \le t \le S\}$ can be expressed as a 3-dimension matrix P of size $N \times M \times S$, where $P(i, j, t)$ is the intensity of the pixel (i, j) of the frame t . If $S = 1$ then the sequence is just a still frame.

The principle of a fractal compression method is to compute compression codes according to the self-similarity of an object, in other words, for every small part of an object, one another similar part of the object can always be found using a method called collage coding. As to a video sequence Seq, that means, for a part of Seq, we can find another similar part in Seq . Thus, Seq needs to be partitioned into non-overlapping small cuboids, called range cuboids. Then for every range cuboid, the approximate part needs to be found and be used to compute compression codes.

In general, the size of a range cuboid is $n \times m \times l$, where n, m can be 16, 8, or 4; and l can be 8, 4, 2 or 1 according

* This work was supported by the Educational Ministry of Fujian Province, China under Grant No. JA02147

to the rates of image motion. For convenience, let $N = n \times k_n$, $M = m \times k_m$ and $S = l \times k_l$. Then the 3-dimension matrix $\mathbf{P}_X$ is partitioned into sub-matrixes $\{P_{i,j,t}; 1 \le i \le k_n, 1 \le j \le k_m, 1 \le t \le k_l\}$ respectively, where $P_{i,j,t}$ stands for the range cuboid $R_{i,j,t}$. A matrix U can be considered to be related with a vector V_U if the values of U are collocated using a row-wise data structure, which leads to a vector V_U . The principle of fractal compression is using the fixed-point theory of Iterated Function System (IFS), that is, for a range cuboid R , there is an IFS whose fixed point is just the vector V_R related to R . Therefore, a similar part D_R of R needs to be found whose related vector V_{D_R} satisfy that $V_R \approx \alpha V_{D_R} + \beta I$, where α, β is called scaling factor and offset factor respectively and the corresponding affined transformation $W(X) = \alpha X + \beta I$ is a generator of some IFS, where X is a vector of $n \times m \times l$ -dimension space $\Re^{n \times m \times l}$ and I is the identity of $\Re^{n \times m \times l}$.

As a generator of an IFS must be a contracted transformation which means $|\alpha| < 1$, a simple part of video sequence Seq may not generate such α . In practice, Seq is partitioned into overlapping small parts called domain cuboids. Generally the size of a domain cuboid $\tilde{D}$ is $2n \times 2m \times l$. $\tilde{D}$ is also related to a sub-matrix of P . The partitions of range cuboids and domain cuboids are shown as figure 1.

The dimension of $V_{\tilde{D}}$ is not the same as that of V_R so they can’t be compared. A domain cuboid $\tilde{D}$ is shrunk by averaging the intensities of four neighbouring pixel of disjoint groups and leads to a $n \times m \times l$ array denoted symbolically as D , which is also known as a codebook cube. If $\tilde{D}$ is related with the matrix

$$V_{\tilde{D}} = \tilde{U}(\tilde{i}, \tilde{j}, \tilde{t}), 1 \le \tilde{i} \le 2n, 1 \le \tilde{j} \le 2m, 1 \le \tilde{t} \le l$$

and D is related with the matrix

$$V_D = \{U(i, j, t), 1 \le i \le n, 1 \le j \le m, 1 \le t \le l\} ,$$

then

$$U(i, j, t) = (\tilde{U}(2i, 2j, t) + \tilde{U}(2i+1, 2j, t) + \tilde{U}(2i, 2j+1, t) + \tilde{U}(2i+1, 2j+1, t))/4$$

where $1 \le i \le n, 1 \le j \le m, 1 \le t \le l$ 。

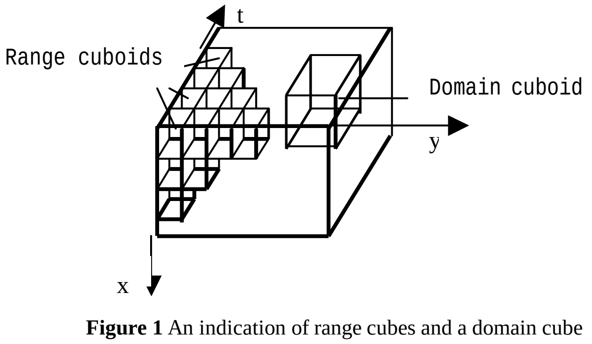


Figure 1 An indication of range cubes and a domain cube

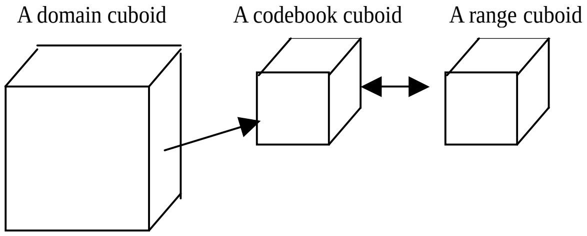


Figure 2 A domain cuboid is shrunk to a codebook cuboid

Figure 2 shows the relationships among a domain cuboid, a codebook cuboid and a range cuboid.

For every range cuboid R and codebook cuboid D , suppose V_R and V_D are the related vector respectively. Then minimal square method can be used to solve following minimization problem:

$$E(D,R) = \min_{\alpha,\beta} \| V_R - (\alpha V_D + \beta I) \|$$

Suppose $K = n \times m \times l$, $V_R = (r_1, r_2, \Lambda, r_K)$, and $V_D = (d_1, d_2, \Lambda, d_K)$, then the optimal values of α, β and the minimal *rms* error $E(D,R)$ can be computed as follows[4]:

$$\alpha = \begin{cases} \dfrac{K\sum_{i=1}^{K}(d_i \times r_i) - (\sum_{i=1}^{K} d_i)(\sum_{i=1}^{K} r_i)}{K\sum_{i=1}^{K} d_i^2 - (\sum_{i=1}^{K} d_i)^2} & \text{if } K\sum_{i=1}^{K} d_i^2 - (\sum_{i=1}^{K} d_i)^2 \ne 0 \\ 0 & \text{if } K\sum_{i=1}^{K} d_i^2 - (\sum_{i=1}^{K} d_i)^2 = 0 \end{cases}$$

$$\beta = \frac{1}{K}(\sum_{i=1}^{K} r_i - \alpha \sum_{i=1}^{K} d_i)$$

$$E(D,R) = \sqrt{\frac{1}{K}[\sum_{i=1}^{K} r_i^2 + \alpha(\alpha\sum_{i=1}^{K} d_i^2 - 2\sum_{i=1}^{K} d_i r_i + 2\beta\sum_{i=1}^{K} d_i) + \beta(K\beta - 2\sum_{i=1}^{K} r_i)]}$$

.

For a given range cuboid R , all possible codebook cuboids need to be compared to find an optimal approximation. That is, we will find a codebook cuboid D_R which satisfy that:

$$E(D_R, R) = \min_{\alpha,\beta} \| V_R - (\alpha V_{D_R} + \beta I) \| = \min_{D} E(D,R)$$

α, β and the index of D_R are known as the compression codes of R .

The quality of decompressed images compared with original images may be described by the value of the *PSNR* (Peak-Signal-Noise-Ratio). The *PSNR* of an 8-bit gray image may be computed by using the formula [8]:

$$PSNR = 10 \times \log_{10} \frac{255^2}{\frac{1}{2^{2N}} \sum_{i,j} (\hat{u}(i,j) - u(i,j))^2} ,$$

Where $u(i,j)$ and $\hat{u}(i,j)$ are the intensities of the original image and decompressed image respectively at the pixel (i,j) . Experience shows that the value of *PSNR* of a decompressed image can be as high as 38 to 40 db [5, 6, 7].

3. BASIC FRACTAL VIDEO COMPRESSION

There are two basic video compression methods, one for cuboid-based compression method and the other is frame-based compression method.

3.1 The Cuboid-based Fractal Compression

When the frame numbers of a video sequence Seq is too big, the RAM space of a computer will be occupied a lot, and the handling speed changes into slow. In a cuboid-based compression a big sequence is divided into groups of frames (GOF), i.e., $Seq = \{GOF_g : 1 \le g \le k_s\}$ which may then be compressed and decompressed as an entity. According to current available machinery velocity, it is appropriate that the frame number S of one GOF is less than 16. For a given $GOF_g\,(1 \le g \le k_s)$, the set of its all range cuboids is $\{R_{i,j,t}; 1 \le i \le k_n^g, 1 \le j \le k_m^g, 1 \le t \le k_l^g\}$ where the size of $R_{i,j,t}$ may be $8 \times 8 \times 4$, $4 \times 4 \times 2$ or $4 \times 4 \times 1$

On the other hand, in order to reduce the compression time, the search of an optimal codebook cuboid for a given range is limited in the adjacent area of the range cuboid. The set Ω_g includes of all codebook cuboids of $GOF_g\,(1 \le g \le k_s)$.

Algorithm 1: The cuboid-based fractal compression for video.

```
Given the image sequence SEQ
Prepare $GOF_g\,(1 \le g \le k_s)$
For $g = 1, \Lambda\ , k_s$
  Begin
    Prepare $\Omega_g$;
    For $t = 1$ to $k_l^g$ do
    For $j = 1$ to $k_m^g$ do
    For $i = 1$ to $k_n^g$ do
    Begin
       For each $D_k \in \Omega_g$ do
          Begin
   $(\alpha_k, \beta_k)$ := Solve $\min_{\alpha,\beta} \| V_{R_{i,j,t}} - (\alpha V_{D_k} + \beta) \|$;
                 Compute $E(D_k, R_{i,j,t})$;
          End
      Compute the compression code:
      $E(D_{opt}, R_{i,j,t}) = \| V_{R_{i,j,t}} - (\alpha_{opt} D_{opt} + \beta I) \|$
      $:= \min_{D_k} \{E(D_k, R_{i,j,t})\}$ ;
       Store $\alpha_{opt}, \beta_{opt}$ and the index of $D_{opt}$;
      End- For
  End -For.
```

Note that the cube-based fractal compression algorithm may obtain high quality decompressed images, but the compression ratio of the resulting compressed object is relatively low.

3.2 The Frame-based Fractal Compression

In a frame-based compression, the compression codes are computed and stored for every frame. Each frame can be considered as a single sequence. In this case, range cuboids and domain cuboids are just blocks. The approximate transformation for every range block is obtained by means of inter-frame similarity, i.e. the domain blocks from the previous frame are used in computing the approximate transformation for the range blocks of current frame. As such the transformation is not necessary a contractive map and the domain blocks, which are not required to be shrunk, from the previous frame may be of the same size as the range blocks and can be used as codebook blocks. In the decompression process, the approximate transformation is used on the previous frame once only without the need of iterations. For a video sequence $Seq = \{f_0, f_1, f_2, \Lambda\ f_S\}$, the decompressed image sequence can be denoted as $\{DF_0, DF_1, \Lambda\ DF_S\}$. For a given frame $f_g\,(1 \le g \le S)$, the set of all range blocks is $\{R_{i,j,g}; 1 \le i \le k_n^g, 1 \le j \le k_m^g\}$. The set Ω_g includes of all domain blocks from DF_{g-1}. The range blocks and the domain blocks are depicted as in Figure 3.

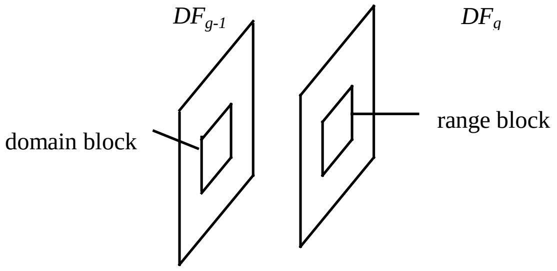


Figure 3 Range blocks and the domain blocks in a frame-based compression.

Algorithm 2: The frame-based fractal compression for video.

Given the image sequence $Seq = \{f_0, f_1, f_2, \Lambda\ f_S\}$:-

Apply {**Algorithm 1:** the cuboid-based fractal compression method for video} to sequence $\{f_0\}$;

```
Do $g = 1, \Lambda\ , S$
    Compute $DF_{g-1}$;
    Prepare $\Omega_g$;
     For $j = 1$ to $k_m^g$ do
     For $i = 1$ to $k_n^g$ do
     For each $D_k \in \Omega_g$ do
          Begin
           $(\alpha_k, \beta_k)$ := Solve
 $\min_{\alpha,\beta} \| V_{R_{i,j,g}} - (\alpha V_{D_k} + \beta) \|$;
           Compute $E(D_k, R_{i,j,t})$;
          End
          Compute the compression code:
          $E(D_{opt}, R_{i,j,t}) = \| V_{R_{i,j,t}} - (\alpha_{opt} D_{opt} + \beta I) \|$
                  $:= \min_{D_k} \{E(D_k, R_{i,j,t})\}$ ;
           Store $\alpha_{opt}, \beta_{opt}$ and the index of $D_{opt}$;
     End-For
End-Do.
```

Two obvious disadvantages of the frame-based compression are the error due to the use of the previous frame will inevitably spread to the latter frames and a delay between frames during decompression. However it should be noted that the frame-based compression results to a high compression ratio which is particularly suitable for video transmission through the internet.

4. THE HYBRID FRACTAL COMPRESION

A hybrid algorithm [9] combining the cuboid-based method and the frame-based method is proposed in order to bring the advantages of the two compression methods. Suppose each GOF_g of $Seq=\{GOF_g : 1 \le g \le k_s\}$ is partitioned into disjoint subsets forming the series $\{G_1, G_2, G_3, \Lambda\}$, which may be decoupled into two disjoint series $sub-GOF_e = \{G_2, G_4, G_6, \Lambda\}$ and $sub-GOF_o = \{G_1, G_3, G_5, \Lambda\}$. A cube-based compression algorithm may be applied to $sub-GOF_o = \{G_1, G_3, G_5, \Lambda\}$ and its decompressed image series is denoted as $\{DG_1, DG_3, DG_5, \Lambda\}$. The compression of the subset G_{2i} of $sub-GOF_e$ is then obtained by using the frame-based method compared with DG_{2i-1}.

Algorithm 3: The hybrid compression algorithm.

```
Given the image sequence $Seq=\{GOF_g : 1 \le g \le k_s\}$ :-
      Do  $g = 1,2,\Lambda, k_s$
          Obtain $\{G_1, G_2, G_3, \Lambda\}$;
          Construct $sub-GOF_o = \{G_1, G_3, G_5, \Lambda\}$,
                    $sub-GOF_e = \{G_2, G_4, G_6, \Lambda\}$
                    ;
      Apply cube-based compression to $sub-GOF_o$;
          Compute $\{DG_1, DG_3, DG_5, \Lambda\}$;
          Do  $i = 1, 2, L$
            Apply frame-based compression to $G_{2i}$;
          End-do
   End-do
```

Note that elements in $\{DG_1, DG_3, DG_5, \Lambda\}$ and $\{DG_2, DG_4, DG_6, \Lambda\}$ may be computed simultaneously. As a result when the video is being displayed the time delay between frames can be as little as possible.

5. THE ADAPTIVE PARTITION COMPRESSION ALGORITHM

The compression ratio of the cuboid method is not high due to the partition with a same size of cuboids. In practice, the rate of inter-frame and intra-frame variation is not the same. The compression qualities may not be changed and high compression ratio may improve if bigger cuboids can be used in the regions where intensities of the pixels change slowly otherwise smaller cuboids will be used to improve compression qualities. So we introduce the adaptive partition to cuboid compression method. First, a sequence of image is partitioned into bigger non-overlapped cuboids. For a given cuboids, if the rms error related to the optimal codebook cuboid is less than the tolerance given initially, then the compression codes are the last codes; otherwise the cuboid is partitioned to 8 smaller sub-cuboids and search a optimal codebook cuboids for every small sub-cuboid. The partition continues until the rms error is small enough or the cuboids can't be partitioned again.

Algorithm 4: The adaptive partition compression algorithm

```
    Given the image sequence $Seq$ :-
    Prepare $Seq=\{GOF_g : 1 \le g \le k_s\}$;
    Given tolerance $\varepsilon$, the maximal partition $\rho^{\max}$,
and the minimal partition $\rho^{\min}$;
    For $g = 1, \Lambda, k_s$
    Begin
          For every possible partition $\rho$, prepare $\Omega_g^\rho$;
          For $t = 1$ to $k_l^g$ do
          For $j = 1$ to $k_m^g$ do
          For $i = 1$ to $k_n^g$ do
             Begin
                 $\rho = \rho^{\max}$; $R^\rho = R_{i,j,t}^{\max}$;
                 Call Octant ( $\rho, R^\rho$ );
             End
    End.
```

```
Procedure Octant ( $\rho, R^\rho$ ):
  $\varepsilon^\rho = 10000$;
  While ( $\varepsilon^\rho > \varepsilon$ ) and ( $\rho \ne \rho^{\min}$ ) do
    Begin
        For each $D_k \in \Omega_g^\rho$ do
          Begin
            $(\alpha, \beta)$ := Solve $\min_{\alpha,\beta} \| V_{R^\rho} - (\alpha V_{D_k} + \beta I) \|$;
            Compute $E(D_k, R^\rho)$;
          End;
        Compute the minimal rms error:
            $\varepsilon^\rho = E(D_{opt}, R^\rho) = \min\{E(D_k, R^\rho) \mid D_k \in \Omega_g^\rho\}$;
        If ( $\varepsilon^\rho \le \varepsilon$ ) or ( $\rho = \rho^{\min}$ ) then
            Store $\alpha_{opt}, \beta_{opt}$ and the index of $D_{opt}$;
        Else
           Begin
             Store tag bit 1;
             New Partition $\tilde{\rho}$:
             Partition $R^\rho$ to 8 small sub-cubes;
             For each sub-cube $\tilde{R}$
              Call Octant ( $\tilde{\rho}, \tilde{R}$ );
           End
    End;
```

6. AN ADAPTIVE HYBRID COMPRESSION METHOD

Because the adaptive partition algorithm can obtain a much higher compression ratio compared to the algorithm based on fixed partition while the qualities of decompressed images are similar, we combine the above hybrid compression method with an adaptive partition. That is, we compress the sub-sequence $sub-GOF_o = \{G_1, G_3, G_5, \Lambda\}$ with the adaptive partition cuboid method instead of the fix-size cuboid method. The numerical experiments show that it can obtain a much higher compression ratio and qualities than the first hybrid fractal method.

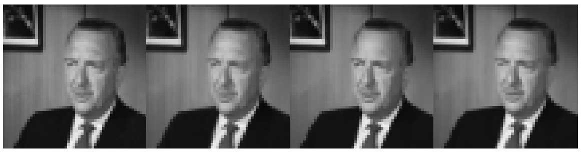
7. NUMERICAL EXPERIMENTAL RESULTS

There are two sequences of motion images in the experiments, one from a videoconference and the other from a movie. The original sequence of motion images from the videoconference consists of frames of 8-bit gray image each of 256×256 pixels, The original sequence of motion images from an extract of a movie consists of frames of 8-bit gray image each of 720×576 pixels, Compression ratio, which is defined as the ratio between the storage size of the original image to the storage size of the compressed image, is used to compare the efficiency of compression.

The values of Compression ratio and PSNR of various compression algorithms are listed in the Table 1.

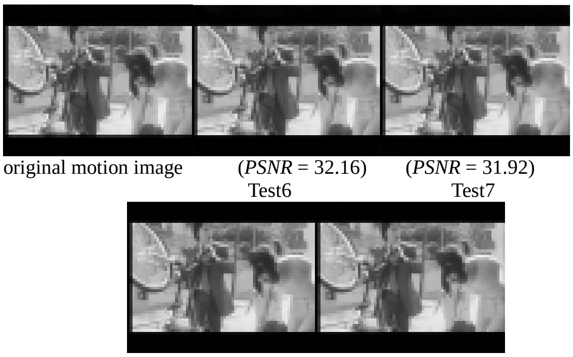
Table 1: Compression ratios and *PSNR* obtained from various methods. V(1): videoconference; V(2): movie

Images	Compression test	Partition	Compression ratio	*PSNR*
V(1)	Test1(Cuboid-based)	4×4×2	6.72	36.15
V(1)	Test2 (Hybrid)	4×4×2	11.76	33.79
V(1)	Test3 (Adaptive)	16×16×4	17.47	34.86
V(1)	Test4 (Adaptive Hybrid)	16×16×4 8×8×1	16.02	35.4
V(1)	Test5 (Frame-based)	8×8×1	16	34.27
V(2)	Test6(Cuboid-based)	4×4×1	3.65	31.96
V(2)	Test7 (Hybrid)	4×4×1	6.06	30.70
V(2)	Test8 (Adaptive)	16×16×4	8.62	30.33
V(2)	Test9 (Adaptive Hybrid)	16×16×4 8×8×1	12.12	29.37



(*PSNR* = 35.38) (*PSNR* = 34.23) (*PSNR* = 34.49) (*PSNR* = 35.25)
Test1 Test2 Test3 Test4

Figure 4 the second frame of decompressed videoconference using all kinds of the algorithm depicted above



original motion image (*PSNR* = 32.16) (*PSNR* = 31.92)
Test6 Test7

(*PSNR* = 30.62) (*PSNR* = 29.75)
Test8 Test9

Figure 5 the third frame of decompressed movies using all kinds of the algorithm depicted above

The compression ratios improve much by using hybrid methods and adaptive partition in basic fractal video compression methods while the qualities of decompressed videos keep well. But it is still a big problem that the compression processes consist of a high computational complexity, especially when the number of frames in GOF_g is large. Note that the sequence of motion images from the movie in the present tests has a difficulty to overcome. This involves a fast moving front wheel as well as a relatively slower motion of the body movement. However such body movement is still much faster compared to the images in the videoconference. As such the compression ratio of decompressed images of the movie is not as good as that of the videoconference. The PSNR obtained for the movie shows no visual effect to a general audience.

8. CONCLUSIONS

A hybrid compression algorithm, which merges the advantages of a cube-based fractal compression method and a frame-based fractal compression method, and an adaptive partition instead of fix-size partition are discussed in this paper. Then an adaptive hybrid method is developed to handle sequences of motion images typically from a video or a movie. Numerical tests were performed for a videoconference and an extract from a movie using various fractal video compression algorithms. The hybrid compression algorithm exhibits relatively high compression ratios for the sequence of motion images from a videoconference. The algorithm shows a minor weakness when dealing with a very fast motion of certain background in addition to a relatively slower body motion. The numerical tests also show no visual difference to the audience of the decompressed sequence compared to the original sequence slowly.

Acknowledgement
The motion images used in the second set of the numerical experiments were provided by Meiah Entertainment Ltd, Hong Kong, China.

9. REFERENCES

[1] K. U. Barthel, T. Voyé. Three-dimensional fractal video coding. In Proceedings of ICIP-95, 1995.

[2] M. Barnsley, L. Hurd. Fractal Image Compression. A K Peters, Wellesley, 1993.

[3] Y. Fisher, D. Rogovin and T. P. Shen. Fractal (self-VQ) Encoding of Video Sequences. In Proceedings of the SPIE Visual Communications and Image Processing, Chicago, USA, Sept. 25-28, 1994.

[4] C.-S. Kim, S.-U. Lee. Fractal coding of video sequence by circular prediction mapping. In NATO ASI Conf. Fractal Image Encoding and Analysis, Trondheim, July 1995.

[5] M. S. Lazar and L. T. Bruton. Fractal coding of digital video. IEEE Transactions on Circuits and Systems for Video Technology, 4: 297 - 308, June 1994.

[6] D. T. Lee. JPEG2000 requirements and profiles – Coding of Still Pictures, Progress report ISO/IEC JTC1/SC29/WG1 N1803, Hewlett Packard, 2000.

[7] Tai-Chi Lee, Patrick Robinson, Michael Gubody, Erik Henne. Software/hardware co-design implementation for fractal image compression. In Proceedings of the 37th annual Southeast regional conference, ACM Press, New York, NY, USA, 1999.

[8] D. Saupe, R. Hamzaoui, H. Hartenstein. Fractal image compression- an introductory overview. In Fractal Models for Image Synthesis, Compression and Analysis, ACM SIGGRAPH'96 Course Notes 27 (Ed: D. Saupe and J. Hart), New Orleans, Louisiana, Aug. 1996.

[9] Meiqing Wang, Choi-Hong Lai. A Hybrid Fractal Video Compression Method. Computers and Mathematics with Applications, submitted.

[10] Wang Meiqing. A Fractal Video Compression Algorithm Based on Adaptive Partition, Journal of Software, submitted.

An Improved Fractal Image Compression Approach by Using Iterated Function System and Genetic Algorithm

Liu Guan rong, Zheng Yang, He Hua
Institute of Computer Science and Technology, Wuhan University of Technology, Wuhan, 430070
Email: qingfeng_1114@sina.com **Tel:** 62925571

ABSTRACT

The paper introduces the basic theory of iterated function systems (IFS theory) and genetic algorithm (GA), on the basis of which, we present an improved method to automatically generate a binary image affine IFS and achieve a fractal image compression. We adopt a natural variable-length genotype encoding to represent the individual. And the multi-object fitness function is also applied in this algorithm. Both theoretical analysis and experiments show a higher compression ratio and better quality images by using this algorithm.

Keywords Fractal Image Compression Iterated function system (IFS) Genetic Algorithms

1. INTRODUCTION

Fractals are geometric patterns that are self-similar, which can be used especially in computer modeling of irregular patterns and natural phenomena. The mathematician named John Hutchinson in his paper [14] developed some new idea from a mathematical theory called iterated function systems (IFS) [1]. IFS were later successfully used in modeling of natural patterns such as clouds, leaves and trees and fractal image compression. Michael Barnsley and his research group who realized the potential usability of iterated function systems. They have indicated that IFS can be used for fractal image compression and developed an interactive system for the generation and solution of inverse problem for IFS [2, 8, 18, 19]. Arnaud Jacquin, a graduate student of Barnsley, has completed an automatic fractal encoding system scheme in his dissertation and proposed a new theory called local or partitioned iterated function system (PIFS) [20].

Now Fractal image compression has been a popular technique for very high compression ratios. An increasing number of researchers build further research and applications on them. Fractal objects are objects whose geometry, natural beauty is generally a result of their self-similar structure [3] and can be zoomed infinitely. Under the condition of satisfying a certain limit, the target of compression is to acquire the simplest description of the image. Image can be regard as an array of attractors of contraction maps on a complete metric space. Recently, new research on seeking out these efficient iterated function systems focuses specifically on wavelet fractal, selections and optimization of fractal definition domain, heuristic fractal and IFS segment. With some new developed methodologies, the evolutionary has emerged as a solution for fractal image compression.

The evolutionary computation is a parallel problem solution that uses ideas of and gets inspirations from the natural evolutionary process [5]. Due to its intrinsic parallelism and some intelligent properties such as adaptation, self-organizing and self-learning, the genetic algorithm (GA) and more generally the evolutionary algorithm (EA) are currently known as efficient stochastic optimization tools, and are widely used in various application fields. Based on ideas of the genetic algorithm, we construct a search match algorithm, which called genetic search method. With the characteristics of the fractal image compression, we adopt a variable-length encoding for individual gene, and propose a new selection scheme about fitness function and location of crossover mutation. Experimental results have shown that our method yields a better compression ratio, improves fidelity and overcomes efficiently some disadvantages of the traditional search method.

2. THEORETICAL FOUNDATIONS OF FRACTAL IMAGE COMPRESSION

Fractal image compression algorithm is based on the fractal theory of self-similar and self-affine transformation. In this paper, to present the basic theory involved in fractal image compression, we restrict attention to complete metric spaces which can be represented, for convenience, by the region $I^2 = [0, 1]^2$.
Some basic definitions are [8]:

(1) Iterated function system definition:
A (hyperbolic) iterated function system consists of a complete metric space (X, d) together with a finite set of contraction mappings $w_n = \{w_1, w_2, \ldots, w_n\}$ (w_n: X→X), with respective contraction factors c_i, for i=1,2,…,N. The abbreviation “IFS” is used for “iterated function system.” The notation for this IFS is $\{X; w_1, w_2, \ldots, w_n\}$ and its contraction factor is $C = \max\{c_1, c_2, \ldots, c_n\}$.

(2) The IFS Condensation Theorem:
Let $\{X; w_1, w_2, \ldots w_n : C \in [0,1]\}$ be a hyperbolic IFS (X is typically [0,1] or $[0,1]^2$). Then the transformation W: H(X)→ H(X) defined by

$$W(B) = \bigcup_{i=1}^{N} w_i(B), \quad \forall B \in H(X),$$

is a contraction mapping on the complete metric space(H(X),h) with contraction factor C (This distance function h is known as the Hausdorff metric), that is
$h(W(A), W(B)) \le C * h(A, B), \forall A, B \in H(X)$
Its unique fixed point $A \in H(X)$ satisfies

$$A = W(A) = \bigcup_{i=1}^{N} w_i(A)$$

This is given by $A = \lim_{n \to \infty} W^n(B)$, for $\forall B \in H(X)$.
Then the fixed-point A is called the attractor of the IFS.

(3)The Collage Theorem:
Let (X, d) be a complete metric space. $F \in H(X)$, and let

$\varepsilon \geq 0$ be given. If we choose an IFS {X; w$_1$, w$_2$, …, w$_n$:C} $(0 \leq c \leq 1)$ obeys:

$$h(L, \overset{n}{\underset{i=1}{Y}} w_i(L)) \leq \varepsilon$$

Where h is the Hausdorff metric. A is the attractor of the IFS.

Then $h(F,A) \leq h(L, \overset{n}{\underset{i=1}{Y}} w_i(L))/(1-c) \leq \varepsilon/(1-c)$,

for all F $\in H(X)$. (An example is shown in Figure 2.1)

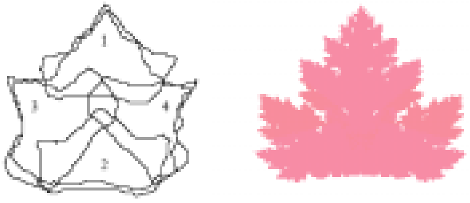

(a) (b)

Figure 2.1 An example of approximating a fern leaf based on the college theorem (see b), which was determined by the four IFS maps. We can see the leaf is viewed as an approximate union of shrunken copies of itself (see a).

2.2 Affine transformation in I^2

Consider the IFR{I^2; w$_1$,w$_2$,…,w$_n$}, where the w$_i$ are affine transformations in I^2 (in two dimensions, e.g. I^2=[0,1]^2). We define the w as follows:

$$w\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} = Ax + T \qquad Eq.1$$

$$((x, y) \in I^2, a, b, c, d, e, f \in R \;\; A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \;\; T = \begin{bmatrix} e \\ f \end{bmatrix})$$

Then (x, y) is coordinate of each pixel, n is number of linear affine maps, and (a, b, c, d, e, f) are transformation constants. Using elementary matrix algebra, we know that a, d stretches, shrinks and mirrors the given point horizontally or vertically. And b, c are used for skewing and rotation, e, f are used for translating the given point [6].

For instance, with Eq.1 we can construct interesting fractals such as the famous spleenwort fern (see Figure.2.2). Then the IFS {I^2, w$_1$, w$_2$, w$_3$, w$_4$}, can be expressed as shown in Table 2.1

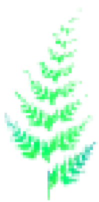


Figure 2.2 A spleenwort fern

In random iteration algorithm, the value of P can be taken to be $p_i \approx \dfrac{|a_i d_i - b_i c_i|}{\sum_{i=1}^{N} |a_i d_i - b_i c_i|}$ for i=1,2…N, where N is number of maps. But other situations you can be treated empirically (and for convenience, by the P=1/4=0.25). In the previous exciting episode, this fern may be described totally in terms of only 4*7+1=29 IFS parameters (N is considered)!

Table 2.1 IFS code for a spleenwort fern

w	a	B	c	d	e	f	p
1	0	0	0	0.16	0	0	0.25
2	0.85	0.04	-0.04	0.85	0	1.6	0.25
3	0.2	-0.26	0.23	0.22	0	1.6	0.25
4	-0.15	0.28	0.26	0.24	0	0.44	0.25

According to the collage theorem, this paper deals with fractals, which are binary image (represented, for convenience, by the region I^2= [0,1]^2). There is an associated grey-level function g(x, y), which may assume a finite nonnegative value. From the point of view of continuous binary image B, each pixel can assume only two discrete values with 0 or 1(white and black are respectively represented 0, 1). To ensure that is generating automatically suitable linear maps and these maps are contractive, the six parameters are described in table 2.2.

Table 2.2 coefficients are constrained

Parameters satisfied conditions
$0 \leq e \leq 1, 0 \leq f \leq 1$ $-e \leq a \leq 1-e, -e \leq b \leq 1-e$ $-f \leq c \leq 1-f, -f \leq d \leq 1-f$ $0 \leq a+b+e \leq 1, 0 \leq c+d+f \leq 1$

The contraction factor δ is calculated by taking the norm of the linear transformation A, and the norm of linear transformation is defined as:

$$\|A\| = \sup_{x \in I^2} \frac{\|Ax\|}{\|x\|} \qquad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad Eq.\ 2$$

When $\|A\|$<1, the affine transformation is contractive.

2.3 Similarity Measurement

How measure an affine IFS whose attractor is close to or looks like a target image? To binary image, let us define S metric function (the function that measures the quality of an approximation is called a metric) as follows:

$$S(A,B) = \frac{|A\,\mathrm{I}\;\;B|}{|A\,\mathrm{Y}\,B|} \qquad Eq.\ 3$$

$|A\,\mathrm{I}\;\;B|$ represents number of black pixel after image A crosses image B, and $|A\,\mathrm{Y}\,B|$ is a black pixel union of image A and B.

Eq.3 expresses the fact that S is smaller, the similarity is worse. In this paper, we present an extension of this function to measure similarity. The associated IFS operator will have the form

$$S(B,\bigcup_{i=1}^{n} w_i(B)) = \frac{\left| B \cap (\bigcup_{i=1}^{n} w_i(B)) \right|}{\left| B \cup (\bigcup_{i=1}^{n} w_i(B)) \right|} \qquad \textit{Eq. 4}$$

The IFS mapping w is contractive and possesses a fixed point "attractor" B that satisfies the property

$$B = \bigcup_{i=1}^{n} w_i(B)$$

3. GENETIC ALGORITHM FOR FRACTAL IMAGE COMPRESSION

Genetic algorithm (GA) is procedure based on the principles of natural selection and natural genetics [10], which has been applied successfully to numerical optimization and combinatorial optimization, statistical computation. Furthermore, GA has proved to be quite efficient searching for approximations to global optimal in the huge and complex spaces in relatively short time. For fractal image compression, we hope ultimately to find "the best" IFS, so that this task is considered the process of optimization.

3.1 Chromosome Encoding

In brief, for fractal image compression, the encoding of chromosome (which is individual of population) is actually the coding of IFS. In second section, we have given definition of affine transformation in I^2 (I^2= [0,1]^2) about binary image (see Eq.1). An efficient IFS (which satisfies conditions in table 2.2) is a workable chromosome. The encode of a gene is defined by

$$w_{ki} = [a_{ki}, b_{ki}, c_{ki}, d_{ki}, e_{ki}, f_{ki}]$$

Let

- k be the kth gene belong to chromosome.
- i be the ith contraction affine transformation of IFS.
- w be regarded as an affine transformation.

A chromosome can be encoded by the sequence of gene, the sequence:

$$F_k = [w_{k1}, w_{k2}, w_{k3}, ..., w_{kn}]$$

Let

- k be locus of the chromosome in population.
- n represents number of efficient affine transformations in the IFS.
- F be regarded as a IFS (there called chromosome). See the description in Java as follows:

```
class Chromosome
{
static int MAXMAPS=22, NPARS=6;
double gene[ ]=new double[MAXMAPS][NPARS+1];
int mapnum;
}
class Population
{int POPSIZE=50;
// POPSIZE is variable-length
Chromosome p [ ] =new Chromosome[POPSIZE];
…
//genetic operation
…}
```

Let

- POPSIZE be number of chromosomes made of population.
- MAXMAPS be the maximum number of affine transformations.
- NPARS be number of transformation parameters (see Eq.1 NPARS=6).
- mapnum be actually number of affine transformations of the IFS.

3.2 Fitness Function

The evolution is driven by a fitness function that is maximized during the process. The fitness value is selected to reflect a desirable trait in the member of the population [7]. In the case of fractal image compression, we measure three function values for evaluating chromosomes. The fitness function is the embodiment of multi-object optimal problem. We have tried three objects:

(1) to maximize similarity measurement S (see Eq.4)

(2) to minimize compression factor β

We have designed a penalty function R_η :N→[0,1]

$$R_\eta(\beta) = \exp(-\frac{\beta^2}{4\eta^2})$$

Parameter η is expected value of β, which determines penalty standard value (that is standard compression factor, let STDCP=0.5). In this paper, parameter beta (β) equals to number of affine transformation and will be modified between runs in the rang [0,1] as follows:

```
STDCP← set penalty standard value;
if beta > STDCP then
This function will penalize beta value;
else if beta <STDCP then
This function will guerdon beta value;
end if
```

(3) to minimize contraction factor δ

We have designed other penalty function

P_σ :[0,1]→[0,1] , the function is described as

$$P_\sigma(\delta) = (1-\delta^{10})\exp(-\frac{\delta^2}{4\sigma^2})$$

For the same reason, the parameter σ is expected value of δ, which is standard number of affine transformation (let STDIFSN=20). In second section, we have known the contraction factor δ is calculated by Eq.2, and modified the method as above.

For a given space I^2, let B be a target image, and c be attractor of the IFS, then the fitness function can thus be written as follows:

$$F_B(c) = S(B,\bigcup_{i=1}^{n} w_i(B)) R_\eta(\beta) P_\sigma(\delta) \qquad \textit{Eq.5}$$

3.3 Genetic Operators

GA includes three basic operators: selection, crossover (recombination) and mutation.

◆**Selection** is a evolution operator that chooses a chromosome from the current generation's population for inclusion in the next generation's population. Before making them into the next generation's population, selected chromosomes may undergo crossover or mutation, which depend upon the probability of crossover and mutation. The offspring consist of the next generation's population. We adopt roulette wheel selection strategy (in which the chance of a chromosome getting selected is proportional to its fitness) and Heuristic selection strategy. Heuristic selection can acquire better individuals and avoid the evolution of population may degenerate during evolution process. This mean that population set = best * BestParent + r * offspring, and then POPSIZE=best + r. This paper adopts best=5 which represents the new generation of population containing the five best parent chromosomes.

◆Crossover

Crossover is a genetic operator that combines or mates two chromosomes (parents) to produce a new chromosome (offspring). The idea behind crossover is that the new chromosome may be better than both of their parents if it takes the best characteristics from each of the parents. Crossover occurs during evolution according to a user-definable crossover probability (P_c). In our experimentation, we adopt one point crossover operator that randomly selects a crossover point within a chromosome then interchanges the two parent chromosomes at this point to produce two new offspring (see Figure 3.1).

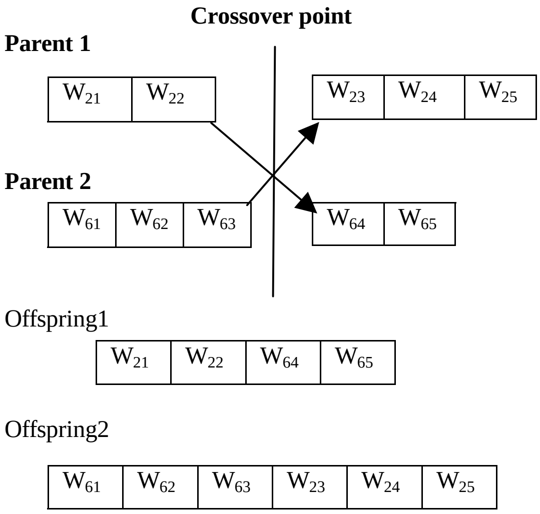


Figure 3.1 Crossover Operator

There is possibility that the length of offspring beyond the allowed maximum, we should measure and modify offspring.

◆Mutation

Mutation is a genetic operator that alters one or more gene values in a chromosome from its initial state. This can result in entirely new gene values being added to the gene pool. With these new gene values, the genetic algorithm may be able to arrive at better solution than was previously possible [12]. Mutation is an important part of the genetic search as it helps to prevent the population from stagnating at any local optima. Mutation occurs during the process of evolution according to a user-definable mutation probability (P_m), and this probability should usually be set fairly low (0.01 is a good first choice) [8].

Affine transformation in any plane that can be decomposed product of four types of basic transformation [13]. According to the characteristics of the contraction affine transformation, we designed four types of affine transformation mutator gene: rotate A_θ, scale A_s, stretch A_e, cut A_c, that is, $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ or $T = \begin{bmatrix} e \\ f \end{bmatrix}$ multiplies associated transformation matrix (such as A_θ, A_s, A_c, A_e). To acquire better offspring, the jitter mutation was considered. The jitter range function is

$$\alpha_{new} = \alpha_{old} + (1 - F_B(c))\varepsilon$$

$F_B(c)$ is the fitness of IFS (see Eq.5), and where ε is a scaling factor chosen uniformly at random over an interval [-1, 1].

```
if fitness→1 then
  Transformation result in only minuteness mutation;
end if
```

In the special case, it is possible that there are some unsuited transformations during the mutation operator. We must measure and modify the new chromosome after mutation operator.

3.4 Algorithm Introduction

```
// initialization population
    Population  pop [MaxGap];
    //MaxGap be maximum generation
    int t←0;
    pop[0]← generate(population)
    /*generate randomly POPSIZE suitable chromosome */
    For (i=0; i<POPSIZE; i++)
    Evaluate (pop[0]); //compute fitness
while (max fitness < target fitness or t>max generation) do
    pop[t+1]←select(pop[t],best);
    /*according to evaluation result, picking out five (best=5)
best individuals into the next generation.*/
    pop[t+1]←crossover(pop[t],$P_c$);
    /* according to the IFS crossover probability $P_c$, picking
   out two parents to generate offspring (that is new
   individuals which are belong to the next generation). */
    pop[t+1]←mutation(pop[t+1], $P_m$);
    /* apply the mutation operator to some individuals in
   pop[t+1] by $P_m$, creating new offspring.*/
    Modified (pop[t+1]);
   // assuring that chromosomes are suitable.
    Evaluate (pop[t+1]);
    t←t+1;
    end while
```

4. RESULT

Our test was made using the simple algorithm described above. The target binary image " maple leaf", a 128×128 pixel array, 2 bits (0=white,1=black) per pixel(see Figure 4.1).

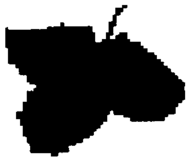


Figure 4.1 target binary image "maple leap"

Table 4.1 shows control parameters used in GA method. We used these parameters for 500 generations and 10 runs (Pentium II processor, 128 M memory). Table 4.2 lists the coefficients of the best solution about the linear of transformations, and the best images of generations 500 shown in Figure 4.2 (other parameters solution see Table 4.3).

Table 4.1 Definition of Control Parameters

parameter	Representation	value
POPSIZE	Population size	100
Best	Selection number of best individual	5
p_m	IFS probability of mutation	0.05
P_c	IFS probability of crossover	0.5
STDIFSN (σ)	standard number of affine transformation	20
STDCP (η)	standard contraction rate factor	0.5
MaxGap	maximum generation	1000

Table 4.2 Parameters of the best solution of the IFS for the maple leaf

w_i	a	b	C	d	e	F
1	0.591	0.000	0.094	0.573	0.189	0.212
2	-0.214	0.326	-0.125	0.107	0.574	0.144
3	0.085	-0.490	0.327	-0.465	0.669	0.447
4	0.490	0.064	0.113	-0.177	0.061	0.447
5	0.402	-0.223	-0.130	-0.367	0.198	0.988
6	0.205	0.022	0.084	-0.164	0.345	0.584
7	-0.226	0.138	0.167	0.174	0.822	0.273
8	-0.287	-0.246	-0.491	-0.059	0.906	0.675
9	0.151	-0.461	-0.519	0.293	0.764	0.598
10	0.436	0.054	0.117	0.786	0.051	0.070
11	-0.788	-0.007	0.342	-0.482	0.935	0.573
12	-0.058	0.356	0.309	0.114	0.407	0.439
13	-0.141	-0.375	0.176	0.023	0.772	0.707
14	0.109	0.313	0.490	-0.173	0.319	0.337
15	0.21	0.328	-0.312	-0.363	0.060	0.750
16	-0.151	-0.175	0.233	0.207	0.589	0.225

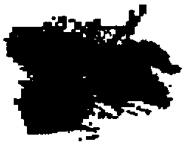


Figure 4.2 decoding image “maple leap”

In spite of our system did not manage to produce a 100% correct solution for the Maple leap problem, we can represent the IFS of Maple by only using 16×6=96 real numbers which just about the fractal compression encoding of maple image.

Table 4.3 results parameters of generation 500

Parameters	value
contraction factor β	0.631611
similarity metric S	0.866465
fitness f	0.490455

5. CONCLUSION AND FURTHER WORK

The results presented in the paper have shown that the variable-length individual genetic algorithm, despite its some suboptimal case, is a good solution for a very important problem in fractal image compression, which could be used to find efficient and good IFS that encode an image. We have defined a discrete data type for image, so this method is more efficient to digital image.

Compare to the current some methods using the partition for the GA, we can’t get a better approximation or compression ratio. It is noteworthiness that we provide information about how to modify some of the GA parameters in order to improve its performances such as multi-object fitness function, design of genetic operators, and chromosome suitable representation. And this method has the great advantage of providing a simple model of the GA behavior.

Further work is considered as follows:
(1) We will focus on color image compression. The color image is described as a function z=f(x,y), and z is the grey-scale level of the image at point(x,y). Fisher [21] calls this type of IFS a partitioned IFS (PIFS) (**see Eq. 6**).

$$w_i \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & s_i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \\ o_i \end{bmatrix} \qquad Eq.6$$

In Eq.5 s_i and o_i denote the contrast adjustment and the brightness respectively [6].

(2) We will go several parallel and distributed implementation strategies into genetic algorithm.

6. REFERENCES

[1] Dietmar Saupe, Raouf Hamzaoui, Hannes Hartenstein, *Fractal Image Compression An Introductory Overview*, ACM SIGGRAPH'96 Course Notes, 1996

[2] Vrscay, E. A hitchhiker’s guide to “fractal- based” function approximation and image compression, Math Ties: Faculty of Mathematics Alumni News 10-13, http: //links.uwaterloo.ca /hitchiker.html, 1995

[3] Anargyros Sarafopoulos, Automatic Generation of Affine IFS and Strongly Typed Genetic Programming. 149–160 Electronic Edition Sp- ringer, 1999.

[4] Evelyne Lutton. Genetic Algorithms and Fractals, Evolutionary Algorithms in Engineering and Computer Science, 1999.

[5] Macarie Breazu, Gavril Toderean, Regionbased Fractal Image Compression using Deterministic Search,

International Conference on Image Processing, 742-746, 1998.

[6] Kirk Saranathan, Fractal Image Compression, ID # 98105923 PMATH 370, Prof.W. Gilbert.

[7] L. Vences and I. Rudomin, Genetic Algorithms for Fractal Image and Image Sequence Compression, in Proceedings Computacion Visual, 1997

[8] Michael F. Barnsley, Lyman P. Hurd, Fractal image compression, A.K Peters, Wellesley, Mass, 1993.

[9] Saupe D, Raul M. Evolutionary Fractal Image Compression [J]. IEEE Trans on Image Procession, 1996, (9): 12-15.

[10] R. Shonkwiler, F. Mendivil, A. Deliu, Genetic Algorithms for the 1-D Fractal Inverse Problem, Proceedings of the Fourth International Conference on Genetic Algorithms, San Diego,1991.

[11] 章毓晋，图象处理和分析，清华大学出版社，1999.

[12] 孙树栋.遗传算法原理及应用[M].北京：国防工业出版社，1999.

[13] Ruhl M, Hartenstern H. Optimal Fractal Encoding is NP-hard [M]. New York: IEEE computer society Press, 1977.

[14] J.Hutchinson, Fractals and self-similarity, Indiana Univ. J.Math. 30, 713-747,1981.

[15] [15] Xiong Shengwu,Guo Jinglei, A evolutionary algorithm of grey image compression, SPIE proceedings, 4553:392-397, 2001.8

[16] Davoine, F., Robert, G., J. –M. Chassery, How to improve pixel-based fractal image coding with adaptive partitions, Proceedings Fractals in Engineering, Vehel, J –L., Lutton, E.,Tricot, C. (eds.), Springer Vderlag, London, 1997, PP. 292-306.

[17] D. Dasupta, G. Hernandez, F. Nino, An Evolutionary Algorithm for Fractal Coding of Binary Images, IEEE Transactions on Evolutionary Computation. Vol.4.No.2. July 2000.

[18] M.F. Barnsley, Fractals Everywhere, Academic Press, New-York, 1998.

[19] M.F.Barnsley and S.Demko, Iterated function systems erlag, 1995.and the global construction of fractals, Proc. Roy. Soc. London A399,243-275, 1985.

[20] A.Jacquin, Image Coding Based on A Fractal Theory of Iterated Contractive Image Transformations, IEEE Trans. Image Proc. 1 18-30,1992.

[21] Y. Fisher, Fractal Image Compression, Theory and Application, Springer-V

Synchronization of Decompression and Display in Fractal Video Compression System

Cheng Hang Shu Zhibiao Fang Yan Wang Meiqing
College of Mathematics and Computer Science, Fuzhou University
Fuzhou, Fujian 350002, China
Email: szb@fzu.edu.cn **Tel:** 0591-3376095

ABSTRACT

Image files with RAW format are simple、no header and no color palette, and each byte directly stands for a pixel intensity. So many researchers who study image compression or image processing algorithms constantly use RAW format file. But it is difficult to display intuitively the qualities of the images with RAW format after image compression or image processing. In this paper we convert image files with RAW format to BMP format in a fractal Video decompression system and display the decompressed image sequence as a section of Video. Furthermore, we use multi-thread technology to implement decompression and display synchronously so that end users of the fractal Video decompression system don’t need to wait a long time to view the decompressed Video.

KEYWORDS: Fractal, Video compression, Multi-Thread

1. INTRODUCTION

About 80% of the information that people obtain is from images according to the research of the authoritativeness organization and the result of the statistics. Thus the research of image process is important in the scientific research and technical application. Moreover it brings the human tremendous economy and societal benefits [1]. In the image compression research, some of new methods emerge with the development of applied mathematics. One of them is fractal image compression. The fractal image compression reduces the redundancy of images by using the self-similarity properties of images, in other words, one part of an image can always be found using a method called collage coding. Its main advantage is that, fractal decompressing a fractal compressed image only needs to compute the fixed point of a fractal transform operator equation, which is very simple and suitable for the situation of one encoding and many decoding. Its compression ratio is purportedly in excess of 10,000:1 in the condition of manual intervention, and still has good visual effect after the decompression [2]. In fractal image compression system, the original images and the decompressed images are stored as RAW formats. Image files with RAW format are simple、no header and no color palette, and each byte directly stands for a pixel intensity. So many researchers who study image compression or image processing algorithms constantly use RAW format file. But it is difficult to display intuitively the qualities of the images after image compression or image processing. Especially when Video image compression is discussed, there are no software tools to display a sequence of motion images or Video stored in RAW format.

In this paper the authors use Visual C++6.0 to develop a displaying tool to display successive RAW format images and combine the tool to a fractal Video decompression system. Firstly we convert the RAW format images to BMP format images, and then display the BMP images with a speed of 25 frames/s to imitate Video [3][4]. Thus we can intuitively compare the qualities of decompressed images with original images. Moreover we adopt multi-thread technology to realize the synchronization between decompression and playing in the system of fractal Video compression. Experimental results show that we can fully utilize the computer resources and reduce waiting time, so the whole efficiency of the program can be improved greatly.

2. IMAGE FORMATS

Generally speaking, image data and header file are both defined in an image file. The header file contains important information about the type, size, color palette data of the image and so on. The other section of the image file is image data, which often is compressed in order to reduce the disk storage required for an image [5] [6]. In this paper we convert RAW format files to BMP format files for imitating Video. So we give a brief summary of the BMP format and the RAW format.

2.1 BMP (Bitmap)

The bitmap-file (BMP) is a standard format in Microsoft Windows operating system that is supported by any type of image processing software in the Windows. It is characterized by its compression or un-compression method, encoding speed quickly, being easily processed by common PC graphics and image processing software.

In reality the bitmap is a layout of pixel values that are stored in an array of bytes. Each value can be stored using 1、4、8 or 24 bits. For example, the single color image just use a bit to store a pixel value. Only when using 24 bits to store a color image or 8 bits to store a gray image, the value stands for the intensity of the pixel. When using smaller bits, such as 4 bits, to store a color image, the value is the index of the color palette.

Table 1: DIB file structures

Bitmap file structures	Structure name	Sign
Bitmap-file header	BITMAPFILEHEADER	bmfHeader
Bitmap-information header	BITMAPINFOHEADER	bmiHeader
Color palette	RGBQUAD	bmiColors[ ]
Image data	BYTE	BitArray[ ]

The bitmaps can be processed as Device-Dependent Bitmaps (DDB) or Device- Independent Bitmaps (DIB). In this paper we adopt DIB. The DIB possess characteristics of trans-platform that has nothing to do with computer operating

Systems and devices. The default filename extension of a Windows DIB file is .BMP [7]. The table 1 gives The DIB file structure. The above-mentioned DIB file structures are defined in the Windows.h.

2.2 RAW

Because the RAW format images are simple and no header, many researchers who study image compression or image processing algorithms constantly convert other format files to RAW format files for simplifying the algorithms.

At present, we can't open these RAW format files with a normal photo package without using a Plugin. And some famous visual programming languages(such as Visual C++, Visual Basic) do also not back up RAW format. The above disadvantages of RAW format results in the difficulties of displaying directly images with RAW format files. Though some software tools such as Photoshop start to support displaying a single image with RAW format, they cannot play Video sequences based on RAW format.

In this paper we use Visual C++ to develop a software tool which converts a sequential images with RAW format files to BMP format files and displays them as a Video. It is necessary for conversion to define a file header (including Bitmap-file header, Bitmap-information header, color palette) and transfer the RAW file data to BMP file data. The concrete steps and VC's source of the transformation is as follows:

2.2.1 Definition of BMP File Header

Because a RAW format file has no a file header, so we need to define a file header for conversion to a BMP file. A BMP file header includes two parts: BITMAPFILEHEADER and BITMAPINFOHEADER. The first part define the type, size and the start point of the image data; and the second part define the width, height, number of bits used to store a pixel, etc.

BITMAPFILEHEADER

```
typedef   struct   tag BITMAPFILEHEADER
   {
      WORD     bfType; //signature, must be 4D42 hex
      DWORD      bfSize; //size of BMP file in bytes
      WORD     bfReserved1;//reserved, must be zero
      WORD     bfReserved2;//reserved, must be zero
      DWORD      bfOffBits;//offset to start of image
                                    //data in bytes
   } BITMAPFILEHEADER;
```

BITMAPINFOHEADER

```
typedef   struct   tag BITMAPINFOHEADER
   {
      DWORD    biSize;// size of BITMAPINFOHEADER
                          //structure, must be 40
      LONG       biWidth;//image width in pixels
      LONG       biHeight;//image height in pixels
      WORD        biPlanes;//number of planes in the
                               //image, must be 1
      WORD        biBitCount;//number of bits per pixel
                                    //(1, 4, 8, or 24)
      DWORD      biCompression;//compression type
      DWORD      biSizeImage;//size of image data in
                                       //bytes
       LONG         biXPelsPerMeter;//pixels per meter in
                                                // horizontal direction
       LONG         biYPelsPerMeter;// pixels per meter in
                                                // vertical direction
      DWORD      biClrUsed;//number of colors in image,
                                    //or zero
      DWORD      biClrImportant;//number of important
                                             //colors, or zero
   } BITMAPINFOHEADER;
```

2.2.2 Definition of BMP File Data

In a gray image file with BMP format, there is a color palette to store the color index for the intensity of every pixel. Therefore we have to define a color palette:

(1) The structure of a Color Palette:

```
      typedef struct tagRGBQUAD
         {
            BYTE    rgbBlue;//defined intensity of blue
            BYTE    rgbGreen;//defined intensity of green
            BYTE    rgbRed;//defined intensity of red
            BYTE    rgbReserved;//reserved,must be zero
         }RGBQUAD;
```

(2) Data Conversion:

Since the fact that the storage order of the intensities of pixels in a BMP format file is contrary to that of a RAW format file, we have to reorder in order to convert the data from a RAW file to the data of a BMP file. Among following program, the function "fread" reads Sizeof(BYTE)' length of data to the memory field assigned by "&value", which is obtained from the assignable file of "fp":

```
BYTE value;
for (int i=0;i<bitHeight; i++)
{
      for (int j = 0; j < bitWidth; j++)
       {
           fread (&value,1,sizeof(BYTE),fp);
           BitArray[1078+ (bitHeight - i - 1)*bitWidth+ j] =
           value;
       }
}
```

3. FRACTAL DECOMPRESSION SYSTEM

Fractal Video compression uses the self-similarity of the images to reduce the spatial and the time redundancy of Video. When compressed, a sequence of Video images $\{f_1, f_2, \Lambda , f_n\}$ with RAW format files is partitioned into small cubes called range cubes, each one denoted as R. In order to find the similarity the sequence is also partitioned into domain cubes, each one denoted as $\overline{D}$. Usually the size of every edge of a domain cube is two times of the size of the relative edge of a range cube. Because a domain cube is not the same size as a range cube, so it is necessary to be shrunk to the same size. The shrunk domain is called as a codebook cube D. In the compression process, for every range cube R, there are three compression codes: scaling factor α, offset factor β and the index of a codebook cube D which make the following rms error

$$\| R - (\alpha D + \beta I) \|$$

minimal [8]. Then all the compression codes are stored in a compression file.

In the decoding process, an initial sequence is given, then the parameters α, β and the index of the codebook cube D

are read from the compression file to compute the optimal approximation of range cube $R \approx \alpha D + \beta I$. Then R is substituted by $\alpha D + \beta I$. The procedure needs to compute iteratively several times. At last we obtain the decompressed sequence of the original sequence with RAW format files.

Compared to the compression time, which is about one hour or more, the decompression time is a little. But when the decompressed sequence needs to display as soon as it is decompressed, we need to consider the time delay as end users can't bear an idle time even if only several seconds.

The decompression time is increasing with the increase of the number of the frames of the compressed sequence. At the same time, it needs to spend some times, maybe several seconds or more, to display the decompressed sequence as the sequence is stored as RAW format files and needs to be converted to BMP format files when displaying. So end users of a fractal Video decompression system need to wait some time to view the Video composed of the decompressed sequence. There are two ways to reduce the waiting time of end users. One way is dividing a big sequence to a couple of small sequences and compressing every small sequence to a single compression file; so all these compression files can be decompressed simultaneously in multi-thread technology; the other way is implementing decompressing and displaying synchronously which also needs multi-thread technology. In this paper we use multi-thread technology to implement the latter method.

4. MULTI-THREAD TECHNOLOGY IN FRACTAL DECOMPRESSED SYSTEM

4.1 The relative concept of multi-threads

Windows 98 and Windows NT's operating system are the typical systems with the multi-job and multi-user, which can make multi users perform with multi applications, then improve greatly the utilization ratio of a CPU. The so-called multi-task usually consists of two great types: multi-process and multi-thread. The process refers to a running application in the system; and the thread is an independent cell in the process, which is a dispatched cell. The process at least consists of a thread, which is usually called the main thread. A process creates a thread or multi threads from the beginning of the main thread, which is so-called multi tasks based on multi threads. In fact, while multi-thread provides exactly the right programming paradigm to make maximal use of the computer resource, it can be effectively used as a way to add more flexibility to any program. It improves the whole performing efficiency of the program and the stability of the computer system with fully using the free time slices of CPU and little wait-time for users. The multi-thread is a method of allowing concurrent execution in a same virtual single address space, and shares the global variable with the system resource, so threads make complex programs easier to write.

4.2 Programming based on MFC

In this paper we use the multi-thread technology of Microsoft Foundation Classes (MFC) in Visual C++ to implement fractal Video decompressing and displaying synchronously. In MFC the thread is divided into two kinds: User Interface Thread and Worker Thread. The former supplies message mechanism to handle the import of users and answer the messages and events created by users, and the latter only is usually used for performing the back calculation and protection missions instead of message mechanism. Because there is no message transmission between fractal Video decompression and Video playing, we just create two Worker Threads in the main thread, one for fractal Video decompression, called Decompression Thread; the other for playing the Video, called Playing Thread.

4.2.1 Implementation of a Worker Thread

In the multi-thread technology, there are two key steps to implement a Worker Thread that are creating a control function and running the thread in the program. The following section gives the programming procedure of the Decompression Thread. The implementation of the Playing Thread is similar as that of the Decompression Thread.

(1) Create a control function for the thread

The control function of a Worker Thread is a running point of the access of the function. The thread ends with the function coming to close. In the control function DecomThread of Decompression Thread we just create an instance of fractal Video decompression class CFracDecomp and call its decompression function Decompress() as follows:

```
UINT DecomThread(LPVOID pParam)
{
    FracDecomp=new CFracDecomp();
    FracDecomp->Decompress();
    FracDecomp->~CFracDecomp();
    delete FracDecomp;
    return 0;
}
```

Where the parameter of "pParam" is a 32-bit value transmitted to the structure function.

(2) Start the Worker Thread

We use following function to start the Worker Thread Decompression Thread:

```
AfxBeginThread
( DecomThread,NULL,THREAD_PRIORITY_NORMAL+1);
```

Where DecomThread() is the control function of Decompression Thread defined as above, and THREAD_PRIORITY_NORMAL+1 is used to define the priority of the thread.

(3) Terminate the thread

When the function terminates，it will return a value of Unsigned integer showing the reason to termination. 0 means a success call，the other values mean some mistakes occurred during a call. The function AfxEndThread() is used to terminate the thread.

4.2.2 Communication among the threads

The decompressed image sequence obtained from fractal decompression class is a set of files with RAW format. So in the Playing Thread, it needs to be converted to the set of files with BMP format and then displayed in the screen according the speed of 25 frames per second. We have to solve the communication problem between Decompression Thread and Playing Thread since the data for Playing Thread comes from the data produced by Decompression Thread. We adopt the global variable technology to implement communication between the two threads, namely, declaring a global variable str[50] with CString type, which store file names with RAW format in Decompression Thread and the data of the file is read to do image format conversion in Playing Thread. Thereby we can obtain the communication between the threads.

Figure 1 gives a flow diagram of Video playing with multi-thread:

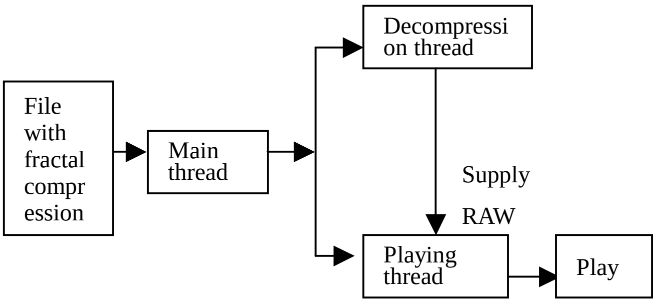


Figure 1: Flow diagram of Video playing with multi-thread

5. EXPERIMENTAL RESULTS AND CONCLUSIONS

In the experiments we use an 8-bit gray image sequence of 25 frames $\{f_1, f_2, \Lambda , f_{25}\}$ from a Video film. Each frame is a 720 × 576 image with RAW format. Firstly the Video sequence is divided to 2 sub-sequences, one for 12 frames, and the other for 13 frames. Each sub-sequence of images is compressed as a compression file, named Comp_file1 and Comp_file2. In the fractal Video decompression system, Comp_file1 is decompressed firstly to image files $\{f_1, f_2, \Lambda , f_{12}\}$. Then multi-thread technology is used to play the sequence $\{f_1, f_2, \Lambda , f_{12}\}$ and decompress Comp_file2 synchronously. As soon as the first 12 frames are displayed, the last 13 frames have been decompressed and can be used to display. Figure 2 gives the programming interface of the Video play based on RAW format using VC++ 6.0.

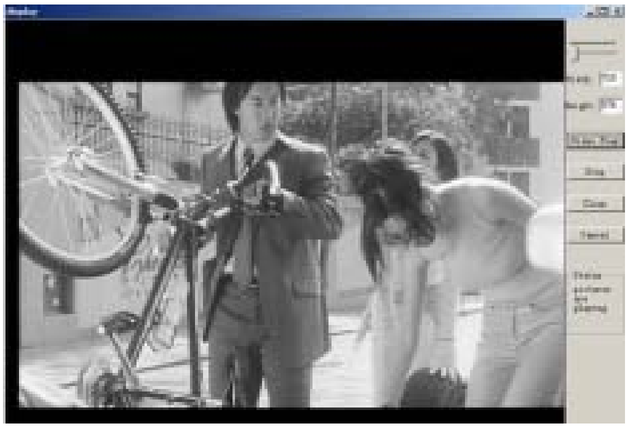


Figure 2: interface of the Video playing

In this paper we discuss the image files with RAW and BMP format. And we implement the conversion from a RAW image file to a BMP image file so we can play a Video composed of fractal decompressed images with RAW format.

Furthermore, we use multi-thread technology to reduce the waiting time of end users who use a fractal decompression system that decompresses Video compression files and plays the Video synchronously. Thus we can evaluate the quality of decompressed Video intuitively. In future research, we will use multi-thread technology to decompress compression files simultaneously so the decompression time will be reduced further.

6. REFERENCES

[1] DeSheng Fu, Yihe Shou. Graphics and Image Processing. Southeast University Press .2002 (01). (in Chinese)

[2] M. Barnsley, L. Hurd. Fractal Image Compression. A K Peters, Wellesley, 1993.

[3] Jie Zhang, Na Yu, Wenxiu Li. Several Practical Class is Convenient for Image Programming Using VC++. Computer Study .2000 (03). (in Chinese)

[4] Creative Work Group. Visual's C++.NET's Multi-medium Whole Application. Post & Telecom Press. 2001 (01). (in Chinese)

[5] Wayne E^Carlson. A Survey of Computer Graphics Image Encoding and Storage Formats. Computer Graphics, 1991, 25(2).

[6] Weigu Zhang, Fuzong Lin. Image File Format ---Windows Programming. Tsinghua University Press. (in Chinese)

[7] Hongbing Wu. Analyze Three Image Format of Windows -- DIB 、 DDB and DIB Section. Microcomputer and Application. 2002 (05). (in Chinese)

[8] M. Wang, C.-H. Lai. A hybrid fractal Video compression method. Computers and Mathematics with Applications, submitted.

The Cache-Multicast Method of Proxy Cache for Streaming Media

Xu Zhiwen, Guo Xiaoxin, Pang Yunjie, Wang Zhengxuan
Faculty of Computer Science and Technology, Jilin University,
Changchun City, Jilin province 130021, China
Email: xuzhiwen@public.cc.jl.cn **Tel:** 86-431-5669210

ABSTRACT

The transmission of large capacity and high byte rate for streaming media becomes a challenging study problem for the Web application. The proxy cache for streaming media is an efficient method to solve this problem. In this paper, we proposed the method of dynamic cache-multicast based on partial cache method such as the segmented cache. The idea of the method is using a dynamic cache to keep partial media object requested by users in an interval time. The advance is that reference frequency of partial media in interval time is higher than that of full media requested by clients. The cache-multicast's method enhances the efficiency of proxy cache for streaming media, mitigates network burden from content server, and saves the traffic resource for network backbone. Event-driven simulations are introduced to evaluate this algorithm is very efficient.

Keywords: Streaming media, proxy cache, dynamic cache.

1. INTRUDUCTION

At present, web proxy cache has broad application, since the proxy cache technology for caching text and image objects does not apply to streaming media. The reasons are follows: First of all, video files require large memory volume. A single file may demand 10 M to 10 G memory volume, which is determined by the quality and length of the video. Most content cached should be stored in the hard disc, and hard disc for proxy cache and memory cache must be organized with great care. Secondly, the fact that real-time media transmission requires obviously large disc volume, bandwidth of network and support within a long period of time, which demands that effective cache algorithm should be used in proxy cache so as to avoid using too much disc volume for caching new content. The characteristic of the streaming media determines that streaming media is its caching objects instead of web objects. The research of streaming media caching technology is a challenging subject.

2. RELATED WORK

2.1 Prefix Cache

As to the study of proxy cache of streaming media, we are still at the stage of theoretical and experimental research, needing further development. In order to solve the problem of startup delay and realize smooth data transmission, Z.Miao propose the method of prefix cache [1,3,4]. When transmitting the media, they divide the media into two parts. The smaller preceding part is called prefix cache, while the latter part is called suffix cache. Generally prefix cache is stored in proxy cache. When the users make an application, what we store in the prefix cache will be played first. Meanwhile, the part stored in suffix cache is transmitted from the content server to the proxy cache. When media stored in prefix cache is over, the part stored in the suffix cache starts to play. In this way, the problem of startup delay is solved effectively. The literatures [3] give the introduction to the research in such areas as the management and organization of proxy cache based on prefix cache, the connecting schemes of the server, the schedule of batch and patch in the proxy cache.

2.2 Segmented Cache

The initial portion of a media stream stored in the proxy cache is more important than the latter portion. The result that the initial portion is of great importance and the media objects should be cached partially applies to most media objects, which leads Wu Kun-Lung [2] and his team to develop a segmentation-based approach to proxy cache media objects. The blocks of a media object received by the proxy server are grouped into variable-sized, distance-sensitive segmentations. In fact, the segmented size increases exponentially from the beginning segment. For simplicity, the video i is 2^{i-1} blocks and contains media blocks 2^{i-1}, $2^{i-1}+1$, …, 2^i-1 ($i\in\{1, 2, …, M\}$). The motivation for such exponentially sized segmentation is that we can quickly discard large blocks of media objects cached. Caching managed in this way can make a quick adjustment on partial objects cached. For example, the proxy cache can release 1/2 of media object cached. The number of segmentations each cached object contains is dynamically determined by the cache admission and replacement policies. Different segmentations are given different caching values.

3. THE CACHE-MULTICAST ALGORITHM

In the proxy cache of streaming media based on segmentation, the size of the media cached varies with the changes of the requested frequency of the certain media. In general, only a part of the media object is saved in the proxy cache, whereas the other part is not saved. When a user requests this media object, the part not saved should be released from the content server to proxy cache and then can be transmitted to the client. When the media has request by many clients at the same time, it should be transmitted from the content server to the proxy cache for several times. We put forward a dynamic efficient organizing algorithm of cache-multicast, utilizing the characteristic of proxy cache for segmented streaming media. Within the certain duration of time, if two or more users request the same video material, we may cache a segmentation of the video dynamically to meet those users' requirements. It is unnecessary for the proxy cache to apply for the video material frequently. Once is enough to satisfy all the users. The length of this duration of time should be shorter than that of the video material and the permitted over caching threshold.

Fig.1 shows the process how the server and proxy handle the clients' requests. At time t_1, t_n, $t_{m,}$ different clients request to media. If t_n -t_1 < the width of largest patch window W, represented by W request at time t_n will be handled by patch channel. If t_m -t_1 > W, the request at time t_m will be handled by the regular channel. In both cases, clients at time t_n, t_m both take up the network and server resource. Fig.2 is the process in which the proxy cache makes use of cache-multicast algorithm. ABC means the dynamic caching windows moving with time. The length of AB equals to t_n -t_1. In order to handle the request at time t_n, we cache this segmentation so as to save the resource that patch channel takes up in the process of web transmission and the server. The length of AC equals to t_m - t_1. In order to handle the request at time t_m, we cache this segmentation so as to save the resource that usual channel takes up in the process of web transmission and the server.

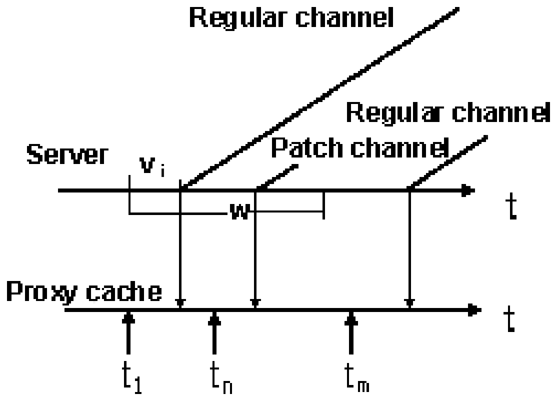


Fig.1 process of the server and proxy

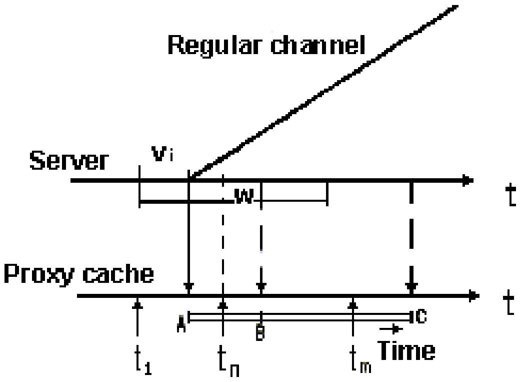


Fig.2 process of cache-multicast

How should we apply the policy of the dynamic cache-multicast so as to ensure he highest byte hit ratio? The length of video i with popularity λ is L_i. And than cached length in the proxy cache is V_i. Clients make request respectively at time t_0, t_1, ···, t_j. Setting t_0=0, then when j=1, 2, ··· and (t_j-t_{j-1}) ≤ the threshold of cache-multicast, we make used of cache-multicast for all the requests from the clients so as to ensure that the proxy cache has the benefits for every user's video request. That is to say we make use of cache-multicast for clients' requests. The method of cache-multicast will help to save network resource and the load of server. J represents the times of clients' applications managed by cache-multicast. Proxy cache saves server's management and network transmission resource for j times. $\Delta= t_j- t_0$ is the length of time between two requests, which equals to the time's length of cache-multicast. When $\Delta \leq V_i$, the proxy cache uses cache-multicast at time t_0+V_i. The time length of the cache-multicast is Δ. When $\Delta > V_i$, the proxy cache saves, at time t_0+V_i dynamically story the media. The time length of the cache-multicast is Δ, and the length of time of prefetch cache in the proxy cache is Δ- V_i. Suppose this is the Poisson process. The users' average request time is $1+V_i\lambda_i$. The j represents the average request time in the dynamic cache-multicast, and $j =\Delta_i \lambda_i$. L_i represents the length of the media. C_s and C_p are respectively the transmission value coefficients, from the server to the cache and from the cache to the clients. The media's segmentation [V_i, L_i], should be taken out from the original server. $C_i(V_i)$ is the average value of delivering video i. Then

$$C_i(V_i) = (C_s \frac{(L_i - V_i)}{1+V_i\lambda_i}\frac{1+V_i\lambda_i - \Delta\lambda_i}{1+V_i\lambda_i} + C_p L_i)\lambda_i B_i$$

The former item is the delivering value from server to proxy cache. The latter item is the delivering value from proxy cache to the clients. Our main aim is to reduce the resources of the backbone network, that is to say the smaller the value of the first item, the better. We considered if two or more applicants apply for the same video within time V_i, we may save the resource of the backbone network for j times on average, and improve the byte hit ratio of proxy cache, by making those latter applicants not take up network resources of the backbone.

In the segmented strategy, according to the application frequency for the certain media, we cache video based on segmentation with different length. This strategy considers the significance of the initial part of a requested media, and ensures higher efficiency of the proxy cache. However, it doesn't consider adequately the effect imposed on it by the media for requested clients in duration. We design the algorithm of dynamic cache-multicast with the main intention to consider adequately the case when many clients apply for the same media within the duration of time and on the basis of segmented cache, dynamically cache the segmentation of the media on demand by multi-user so as to ensure that it is necessary for only the very first applicant to take out the part of media not cached in the proxy cache from the server and cached this segmentation in the proxy cache. This length of time is regarded as the length of the media cached. We adopt FIFO replacement policy to make this part of cached media meet the need of multi-user within the duration of time. The efficiency of proxy cache will be influenced by the determination of this time length. If we determine this time length by calculating users' behavior, then the proxy cache allocates the caching length according to this time length. The method mentioned above will bring some troubles to the management of proxy cache, and therefore, permission control and replacement strategy will become more complex. Moreover, calculating the most efficient time interval according to the users' behavior, will on the one hand bring burden to the proxy cache, and on the other hand save the resources of the network and server.

4. PERFORMANCE EVALUATION

4.1 Methodology

We utilize an event-driven simulator to stimulate the proxy cache service and furthermore to evaluate the algorithm of the dynamic cache-multicast based on variable-size. The algorithm of LRU and FIFO stack were used to keep track of media objects in the cache. LRU stack was to track the initial segmentation and the segmented-cache, whereas the FIFO was to track the cache-multicast. C_{init} represents the initial segmentation, and C_{multi} is used to represent dynamic cache-multicast segmentations. We calculate byte hit ratio and request time with startup delay. The byte hit ratio measures the total bytes of the objects cached over the total bytes of objects requested. When a request arrives but the initial K_{min} segmentation is not cached in the proxy cache, then there will be a start delay. Let's suppose that the media objects are video, and the size of the video is uniformly distributed between 0.5B and 1.5B blocks, where B represents video size. The default value of B is 2000. The playing time for a block is assumed to be 1.8 seconds. In other words, the playing time for a video is between 30 minutes and 90 minutes. The size of cache is expressed on the basis of the quantitative description of media blocks. The default cache size is 400000 blocks. The inter-arrival time distributes with the exponent λ. The default value of λ is 60.0 seconds. The requested video titles are selected from a total of the distinct video titles. The popularity of each video title M accords to the Zipf-like distribution. The Zipf-like distribution brings two parameters, x and M. the former has something to do with the degree of skew. The distribution is given by $p_i=C/i^{1-x}$ for each i∈{1, ···, M}, where

$$c = 1/\sum_{i-1}^{M} 1/i^{1-x}$$

is a normalized constant. Suppose x = 0 corresponds to a pure Zipf distribution, which is highly skew. On the other hand, suppose x = 1 corresponds to a uniform distribution with no skew. The default value for x is 0.2 and that for M is 2000. The popularity of each video title changes with time. It is very likely that a group of users may visit different video titles at different periods of time and the users' interest may be different. In our simulations, the distribution of the popularity changes every request R. The correlation between two Zipf-like distributions is modeled by using a single parameter k that can be any integer value between 1 and M. First, the most popular video in the first Zipf-like distribution finds its counterpart, the r_1-th most popular video in Zipf-like distribution 1, where r_1 is chosen randomly between 1 and k. Then, the most popular video in the second Zipf-like distribution finds its counterpart, the r_2-th most popular video. r_2 is chosen randomly between 1 and min (M, 10), except r_1. The rest may be deduced by analog. When k represents the maximum position in popularity, a video title may shift from one distribution to the next. k = 1 expresses perfect conformity, and k = M expresses the random case or unconformity.

We compared the cache-multicast approach with the full video approach, the variable-sized segmented approach, and the prefix schemes in terms of the impact they imposed on byte hit ratio and startup delay from the following aspects: the cache size, the skew of the video popularity, users' viewing behavior and other related system parameters.

4.2 Impact of Cache Size

We study the impacts imposed by the cache size on the byte hit ratio and startup delay. For a fairly wide range of cache size, the dynamic cache-multicast strategy has the highest byte hit ratio and the lowest fraction of requests with delayed starts, whose byte hit ratio is higher than the variable-sized segmented approach and the prefix/suffix schemes with the same startup delay. Fig. 3 shows the impact cache size imposes on the byte hit ratio. Fig. 4 presents the impact imposed by cache size on the fraction of requests with startup delay. The full video approach and the prefix have comparable byte hit ratio, with the full video approach having a slight advantage over the prefix scheme. For a smaller cache size, the advantage of the byte-hit ratio managed by the variable-sized segmented approach is quite evident. The dynamic cache strategy proves to have the highest byte-hit ratio. Even though the full video and the prefix approaches perform almost equally in byte-hit ratio, they differ dramatically in the fraction of requests with startup delay. The full video approach has a significantly higher fraction of requests with startup delay (Fig. 4). For example, for a cache size of 400,000 blocks, 60% of the requests cannot start immediately using the full video approach. However, only 15.6% of applicants encounter startup delay using dynamic cache, variable-size segment and prefix approaches. Within the whole range of cache size, the effect of the dynamic cache-multicast approach, variable-size segmented strategy and the prefix strategy are basically the same. They all effectively solve the problem of startup delay.

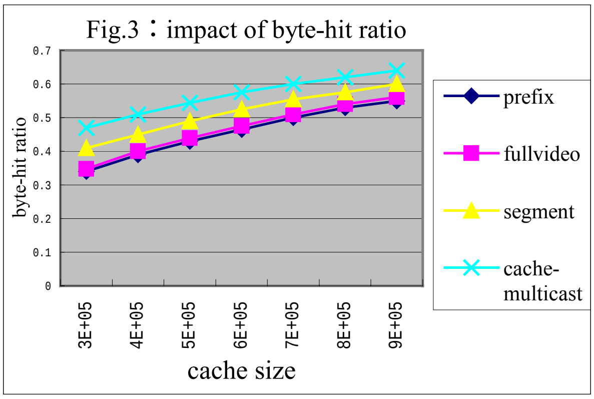


Fig.3：impact of byte-hit ratio

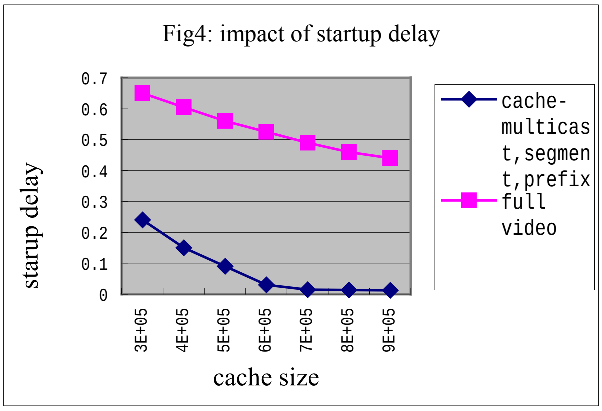
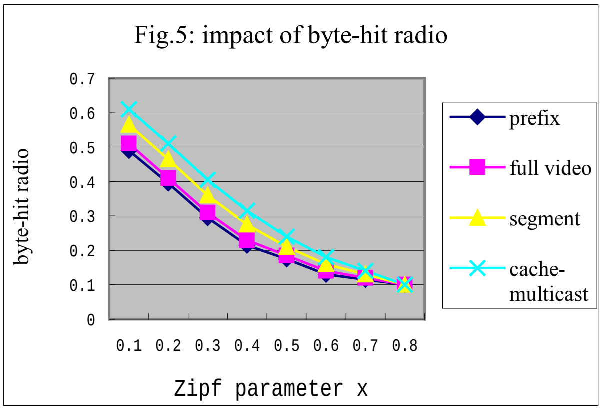
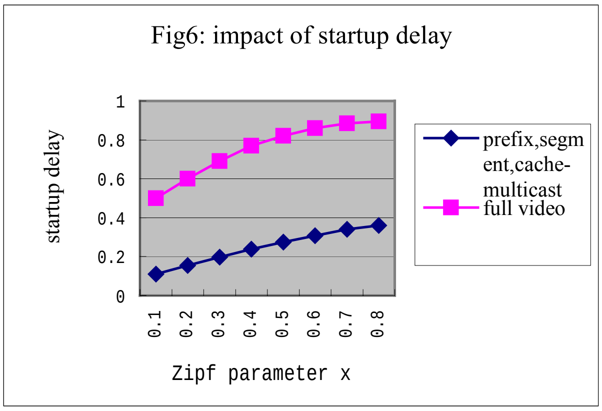


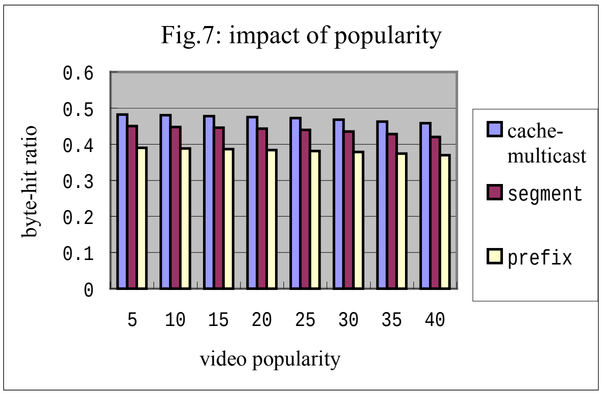
Fig4: impact of startup delay

4.3 Impact of Video Popularity

Let us examine the impact that the video popularity imposes on the byte-hit ratio and start delay. The dynamic cache-multicast strategy has the highest byte-hit ratio when the video popularity makes changes of wide scope. The dynamic cache-multicast approach, the variable-sized segment and the prefix/suffix schemes all have the same least request time with start delay, which is superior to the whole video. Fig. 5 shows the impact of skew in video popularity on byte-hit ratio, while Fig. 6 shows its impact on the start delay. In addition to the parameter of Zipf, x, we also studied the changes of the popularity distribution and the impact of the maximum video shifting position k. The request R of the video shift was set to be 200. Fig. 7 shows the impact of the maximum shifting position of a video. When the maximum shifting distance increases, the byte-hit ratio of the dynamic cache-multicast, the variable-sized segment and the prefix approaches will fall, but only very slightly. The dynamic cache-multicast strategy is always better than the variable-sized segmentation and the prefix approach, which is closely related with the popularity distributions of the video titles and with the range of K, which is from 5 to 40. We also change the value of R, but the result is quite similar. Its byte-hit ratio has only very slight impact on R.



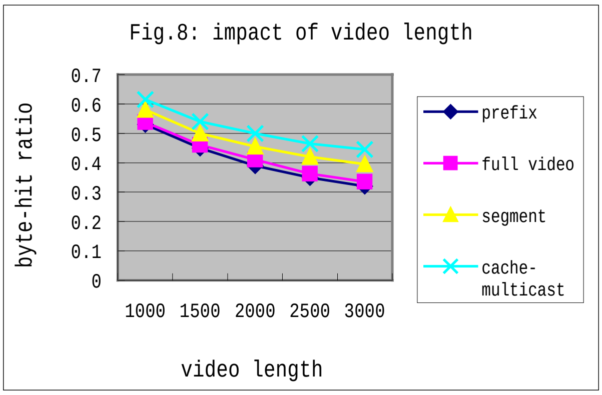


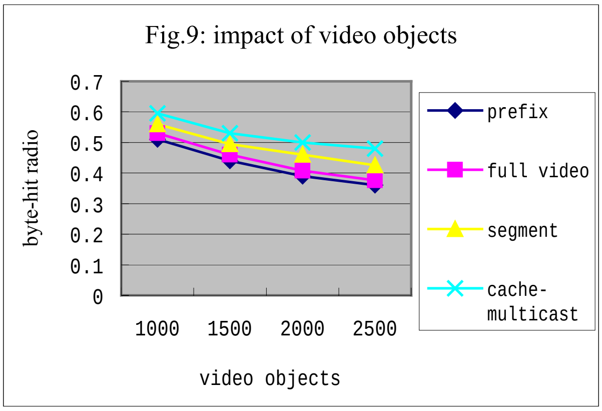


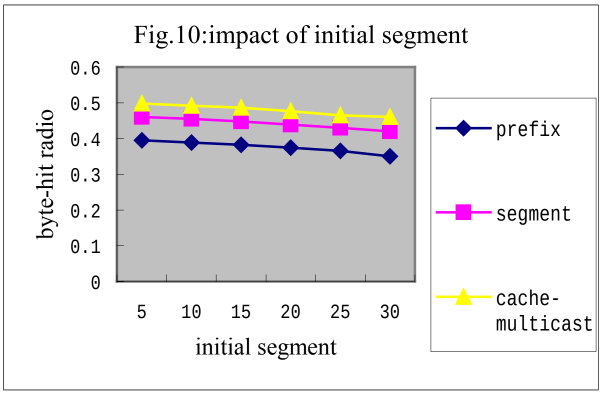
4.4 Impact of Other System Parameters

Fig. 8 shows the impact of video length imposes on the byte-hit ratio. In general, as the size of a media file increases, the byte-hit ratio will fall, this is true for all the four approaches. When the size of a media file is very large, the dynamic cache-multicast algorithm can ensure higher byte-hit ratio than the segmentation and other two approaches. As to a video with the length of 3000 blocks, the byte-hit ratios of dynamic cache strategy and variable-sized segment strategy are respectively 32% and 28%. If the length falls to 1000 blocks, the byte-hit ratios may reach 61% and 59% respectively. No matter which approach we use, dynamic cache strategy or variable-size segment approach, caching large media will cause the byte-hit ratio in proxy cache to fall. However, dynamic cache-multicast strategy is better than variable-sized segment strategy.

Besides the size of a media file, the distinct media objects can also affect the efficiency of proxy cache. On the Web there exist many distinct media objects. As the demand on such objects increases, caching becomes less effective. Fig. 9 shows the cases of applicants for distinct media objects from the angle of quantity. Once again, the dynamic cache-multicast strategy and the variable-sized segment approach have much more advantage over the other two approaches, even when the conditions for caching are less favorable for both of them. Comparatively speaking, the advantage of the cache-multicast strategy is more outstanding. Fig. 10 examines the percentage of caching dedication for storing the initial segments (prefixes). Because the cache for the suffixes is reduced, the byte-hit ratio falls with the increase in using initial segments. This slight decreasing in byte-hit ratio can be offset by increasing benefits substantially by the means of reducing start delay. For example, let us compare these two cases, 5% and 15%. The byte-hit ratio is barely decreased, but the fraction of delayed starts drops substantially. However, no more benefits can be derived once the percentage of the initial segments cached increases beyond 20%.







5. CONCLUSIONS

In this paper, we put forward the algorithm for dynamic cache-multicast, based on the variable-sized segmented approach. The segment-based approach groups media blocks into variable-sized segments. This method differs from the way we handle a web object, which is usually handled as a whole. The algorithm of dynamic cache-multicast considers adequately the users' request behavior. While maintaining the advantage of the variable-sized segmentation, it provides the multi-user within a period of time with the same media they request by using dynamic cache. The algorithm of cache-multicast greatly saves traffics resource on the backbone of network, and enhances the byte-hit ratio and efficiency of the proxy cache. Event-driven simulations evaluation was introduced to compare the dynamic cache-multicast with the variable-sized segment approach, the full video approach and the prefix caching approach from the following angles: the cache size, the skew of video of popularity, the length and the quantity of a certain video. Therefore, the algorithm for dynamic cache-multicast effectively saves the network resource and enhances the byte-hit ratio of proxy cache.

6. REFERENCES

[1] Z.Miao and A.Ortega, Proxy caching for efficient video services over the Internet, In Proc. Of Int, Web Caching Workshop, Apr. 1999.

[2] K.L.Wu and P.S.Yu, Segment-Based Proxy Caching of Multimedia Streams In Proc. Of IEEE INFOCOM ,May, 2001.

[3] O.Verscheure, C. Venkatramani, P. Frossard, and L. Amini, "Joint server scheduling and proxy caching for video delivery," *Computer Commun.*, vol. 25, no. 4, pp. 413–423, Mar. 2002.

[4] S.Sen, J.Reforrd, and D.Towsley, Proxy prefix caching for multimedia streaming, In Proc. Of IEEE INFOCOM, Mar.1999.

[5] Pascal Frossard and Olivier Verscheure, "Batched Patch Caching for Streaming Media", IEEE COMMUNICATIONS LETTERS, VOL. 6, NO. 4, APRIL 2002.

[6] B.Wang, S.Sen,M.Adler and D.Towsley. "Optimal Proxy Cache Allocation for Efficient Streaming Media Distribution", In Proceedings of IEEE INFORCOM, 2002.

[7] S. Ramesh, I. Rhee, and K. Guo, "Multicast with cache (mcache): Anadaptive zero-delay video-on-demand service", in *Proc. IEEE INFOCOM*,April 2001.

Xu Zhiwen is a assistant professor in Faculty of Computer Science and Technology of Ji Lin University. He graduated from Ji Lin University in 1987. His research is the network multimedia.

Guo Xiaoxin is a doctoral student in Faculty of Computer Science and Technology of Ji Lin University. His research is the computer graphics and image.

Pang Yunjie is a professor in Faculty of Computer Science and Technology of Ji Lin University. He is doctoral teacher. His research is the computer graphics.

Wang Zengxuan is a professor in Faculty of Computer Science and Technology of Ji Lin University. He is doctoral teacher. His research is the computer graphics.

The Application of Distributed Streaming Media Technology in CAI

Pan Wenhong, Zhang Jianhua
School of Information and Science & Engineering, Shenyang University of Technology,
Shenyang, Liaoning Province, 110023, P. R. China
E-mail: zhangjianhua2002@hotmail.com

ABSTRACT

This article introduces the application of streaming media technology in CAI. It concerns the basic communication protocol, the characteristics of applied software, distributed systems, parallel programming, the brief of introduction for main algorithm.

Keywords: streaming media, distributed systems, parallel programming, CAI

1. INTRODUCTION

Recent years, the layered software architecture, Client/Server, constructed Distributed Server System is one of the main service models on Internet. In the past, all most of data are text or Hypertext type on Internet. What people greatly care about is the precision of the data transmission. And then, there are media streaming data now. They are the common feature with large quantity, communication timely and more interactively. They are the special type multimedia data, such as the data being used Video-on-demand system. The specialist indicates that, the application and research is pay attention to industry and scientific organization. It’s an important application technology on Internet.

This technology contains encode/decode, agent server technology, communication technology base on end-to-end and service system. There are some typical applications, such as the Video-on-demand system, or the Video Conference system, the Distance-learning system, Digital Library system. On Internet, there are three common streaming media system, Microsoft Windows Media Player, Apple QuickTime, Real Real-Networks. They make it convenient for computer users to manage the streaming media technology, and they also make contributions to improve that technology in the previous time.

As a developing technology, it is difficult to find a mature product for computer users in the market. In addition, the fundamental research is in the deeper spreading phase. But people have noted its potential value in the market, so more input is given to develop its software and hardware. Some authoritative experts pointed out the strategic research and development direction of the information industry in the future, they had pondered the factor of this aspect.
In the following part, I will introduce the application of this technology for CAI.

2. DISTRIBUTED AND LAYERED SOFTWARE ARCHITECTURE SERVICE SYSTEM

According to our realistic situation, quantity of users is not quite much and ranges that are divided are clear, the model of CAI system is multi-grade of software architecture with two layers. The basic architecture is Client/Server model.

3. SYSTEM SERVICES

By using CAI system in the class, the students and teachers can do the following on Internet:

- Learning Forum-based on the News service
- Discussion of learning content-based on the Video Conference System
- Chat Room-based on the exchange service
- Up/Download for the multimedia files
- Sending/Receiving E-mails
- Video-on-demand for the content of teaching
- Inquire and respond system of teaching

4. BASIC COMMUNICATION PROTOCOL

According to the above demands, the transferred information of network can be divided into two types: one is the information in the text or Hypertext files, the other is the information in the streaming media data. We can analysis the two points from the basic protocol of network and the service offered by the demands to support the two types of information.

As for the first type of information, the basic protocol of network that the network operating system should support contains the family of TCP/IP (Transfer Control Protocol/Internet Protocol) and its sustained protocol suite, that is, HTTP (Hypertext Transfer Protocol), NNTP (Network News Transfer Protocol), SMTP (Single Mail Transfer Protocol), POP3 (Post Office Protocol Version 3), IMAP (Internet Message Access Protocol), FTP (File Transfer Protocol), etc. The second type is the main model of network transmission, for the function expansion of IPV6, the higher version of IP protocol, multicast technology can be used. It is bind to UDP (User Data-gram Protocol), and it is compatible to the traditional transferred mode of UDP.

In regard to the second type of information, the basic network protocol that the network operating system should support contains the family of RTP/RTCP (Real-time Transfer Protocol/Real-time Transfer Control Protocol) and related protocols.

RTP sustains the real-time transferred data in the point-to-point and point to multi-point broadcasts of the network service. These data consist of audio-on-demand, video-on-demand, Internet telephony, videoconferencing, and so on. It may also transfer different forms of data, such as the data in the form of WAV, GSM (global system for mobile communications), and MPEG (Moving Picture Experts Group). RTCP is used for count, manage, and control the transference of RTP packet. RTP\RTCP provide the load platform of network for streaming media technology.

5. TECHNOLOGY OF SYSTEM SOFTWARE

Windows Server 2003 that was issued in 2003 has been greatly improved in security, reliability, efficient exploitation and concentrated management. If that operating system is used, we will get many profits.

Active Directory Service:
There are many problems to deal with, if we want to establish an efficient and distributed environment. For example, resources protection of computer system, it only can be employed and dominated by legal users, and communication of distributed software system, the application of establishing and sustaining more complicated users, the application based on the Web, setting up the programs for users and application under the efficient, administrative and distributed environment; etc. Active Directory supplies us with the directory service, the secure service, the applied and the sustaining service, so that it can solve the above problems.

Enhanced and Distributed File System:
Distributed File System is able to provide the service for supervising the distributed resources on network. DFS is composed of some software that is stationed in the server and the client computer on network. It connects the shared files stored on the different file servers (or users' computers) to a namespace with a single name, and come into being a tree topology. Therefore, it is much easier for users to find a file on the Internet.

VDS and SAN:
VDS (Virtual Disk Service) is the only interface of storage block for storage system in the software, hardware and interconnected device and SCSI, etc. It can be realized to interconnect operate between the storage devices supplied in the operating system.

SAN (Storage Area Network), a storage sub-network, is made up of many servers, bridges and storage devices. Every server can regard the storage devices of SAN as the hardware of his own computer. Hence the data storage is very elastic. Presently, its storage capacity may reach 400 TB. Its service function compels the storage and transference of the information in the streaming media type and the extensibility of the whole system to provide us a completely new pattern.

Web Service: IIS (Internet Information Service Version6.0) has developed into a reliable web server. It not only offers the global web service, but also process FTP server, SMTP server, NNTP server. It has two important features. First, it support the technology of ASP.NET that established on the Microsoft.NET Framework, use to applied programs for setup users' web site, makes use of the Common Runtime to provide users for the complier framework which is able to set up the applied server with powerful function at the server (back end). Second, it employs the fresh file storage mode XML and sustenance of Internet standard, such as SOAP and WSDL. Consequently, it makes the system greatly improved in the efficient running, extensibility, secrecy and so on.

Web Service is the basic service when it is used for doing the distributed computing on Internet. Since XML was issued in 1998, Web Service has become the integrated platform of the applied programs for users, communication and cooperation between applied programs. UDDI (Universal Description Discovery, and Integration), the standard allows other users to discover them easier on Internet, they has registered their Web Service information. Moreover, Windows Server 2003 enables clients to use any programming language such as VB.NET, C++.NET, C#.NET, so that they can finish their tasks of programming more efficiently and quickly, because it has had the Microsoft.NET integrated.

Cluster and Network Load Balancing:
Cluster technology is a common scheme for user at the client on network, redundant system at a low cost, applicability and reliability.

The technology of Network Load Balancing can allocate system resources dynamic between servers, it improve the efficiency of computer system.

Application of Exchange Server:
Exchange Service supply transmission of real-time data, Video Conference, real-time communication, the amplify session and so on. The service is indispensable for the CAI.
We designed a suitable interface of windows for the CAI in our application system.

6. MAIN OF PROGRAMMING

Process Communication:
IPC between process of a computer and RPC between the application programs of computers on network, they must be considered in programming.

Thread Communication:
In parallel programming of multi-thread, using the interface of Socket is a good idea for control the input/output system. Microsoft Windows supply the interface, Winsock, so AETT (asynchronous event triggered task) model can be used in our application system.

Main Algorithm:
In programming there are three important algorithm had been used. One is Unblocking of Multi-thread Communication for Parallel Process Multi-requests. Second is Digital Signature, value computed with a cryptographic algorithm and appended to a data object in such a way that any recipient of the data can use the signature to verify the data's origin and integrity. Third is Video-on-demand and Scheduling of Streaming.

7. PROSPECT

Currently, the focus of researching of the streaming media application technology, that are Streaming Scheduling, Multimedia Proxy and Caching, Streaming Application Level Multicast.

In universities the opening of courses is important, that include Intelligent Optimization Algorithm and Soft Algorithm. The teaching content should enrich gradually with the rapid progress of this technology.

8. REFERENCES

[1] Barry Wilkinson, Michael Allen. Parallel Programming. China Machine Press, Peking, 2002

[2] Fang Shucheng, Wang Dingwei, Fuzzy Mathematical Theory And Optimization Algorithm, Science Press, Peking, 1997

[3] David E.Culler, Jaswinder Pal Singh, Anoop Gupta. Parallel Computing and System Structure. China Machine Press, Peking, 2000

[4] Steve Mack, Streaming Media Bible, China Electronics Industry, Beijig, 2003.1

Zhang Jianhua is a professor and a head of Interface and Control Lab., School of Information & Science and Engineering, Shenyang University of Technology, P. R. China.

Pan Wenhong is a lecturer of Shenyang University of Technology. She works in college of China-Germany, Shandong Agricultural University, P. R. China.

A Parallel Image Processing System Based on DSP Arrays

Liu Zhi Zhu Wei Chen Shu
School of Information Science & Engineering, Shandong UniversityJinan, 250100, China
Email: liuzhi@sdu.edu.cn

ABSTRACT

Parallel image processing is very important for high-speed vehicle detection in ITS (Intelligent Transport System). In this paper, a parallel image processing system that is based on one-dimensional DSP arrays is presented. This system integrates parallel memory access and parallel processing technologies to suit for the parallel all kinds of parallel algorithms, so that it can get better preference of processing than other architectures.

Keywords: parallel processing; image; DSP; neighborhood frame memory, ITS

1. INTRODUCTION

In ITS (Intelligent Transport System), it is necessary to detect the high-speed vehicles with complex background. Using video and image processing technology to get the vehicles' information is popular because of its flexibility and easy realization. However, for a real-time system a tremendous mount of processing power is required for different type of image processing algorithms .To provide the high computational demands under real-time constraints, a highly parallel processing scheme has to be applied.

Usually, several general-purpose processors are combined to a multi-PC or multi-processor system. This leads to problems for on-board systems that are limited in volume and power consumption. Therefore, a parallel image processing system based on one-dimensional DSP arrays has been developed.

2. THE ARCHITECTURE OF THE SYSTEM

According to the ITS system's requirements for flexibility and scalability, the basic architecture of the system is briefly shown as the Figure 1.

In this system DSP are the core part of each processing module. DSP are special processors developed for real time signal processing. They can deliver high computing performance at a reasonable cost and their architecture incorporates various features to perform intensive numerical computations, such as hardware multiplication, capability of performing multiple instructions per cycle, and on chip memories to speed up repeating instructions. Traditional approach is to organize processors as two-dimensional (2-D) arrays. However, architecture based on one-dimensional (1-D) linear connectivity is another powerful alternative to 2-D mesh connected systems. In this system, we use the 1-D linear arrays. 1-D arrays and their variations have been studied in [3,4]. The major advantage of 1-D arrays is their simplicity and low I/O requirements. We choose TMS320VC5402 for it good performance and low price.
This 1-D arrays architecture mainly includes master part, slave part and data bus between them. The communication between the master processing part and slave processing part is controlled by FPGA. Each processing module consists of one TMS320VC5402 whose main frequency is 100 MHz, performs all main computation, and some memories to store input data and output data as well as intermediate results. Especially, the master-processing module affords to distribute the processed data additionally. Each module features its own local memory, thus avoiding the memory bus conflicts of a shared memory design and providing the highest possibly bandwidth between DSP and memory. The main type of memory is synchronous DRAM (SDRAM). The SDRAM memory capacity is sufficient for storage of the whole input image data after the video is digitized to 8 bits per pixel or the size of the data after processing. To get the maximum performance, designing RAM as neighborhood frame memory is better than traditional memory methods. What's more, the result of process can be transported to PC by PCI bus whose interface has been designed by FPGA.

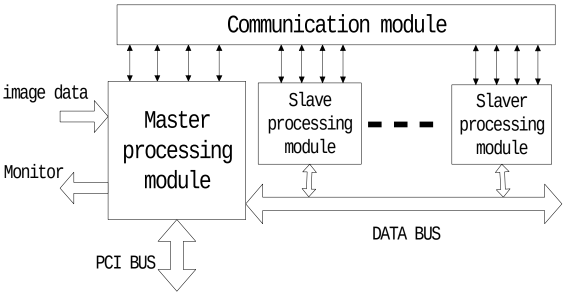


Figure 1 the diagram of the system

The FPGA's another purpose is controlling to send and receive data to and from the processing module. In this processing system, neighborhood frame memory that can realize parallel access for the image data take an important role [2], since the processing efficiency can get the maximum only if the architecture of the process coordinates the architecture of memory. In the processing for traffic image, 3×3 neighborhood images data are usually used for Sobel operator, cross median filter, 3×3 2-D convolution and so on, while, an average processor can only deal with a single point or a row but not neighborhood image data in real time. So the memory architecture to provide the neighborhood image data is necessary. In neighborhood frame memory, to complete parallel access function, the concept of alternation matrix, Namely, suppose matrixes $A=[a_{i,j}]_{m\times n}$ and $B=[b_{k,l}]_{p\times q}$ where $b_{k,l}=1$, and any sub-matrix of the Kronecker product of A and B is defined as the alternation matrix of the matrix of A. For example,

$$A=\begin{pmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{pmatrix} B=\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ and}$$

$$B \otimes A = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} & a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} & a_{20} & a_{21} & a_{22} \\ a_{00} & a_{01} & a_{02} & a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} & a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} & a_{20} & a_{21} & a_{22} \end{bmatrix}$$

The matrix's each 3×3 sub-matrix is its alternation matrix.

As an elementary transform, the meaning of alternation matrix lies at that it can descript the regularity of the architecture of neighborhood frame memory. Based on this concept, not only the neighborhood frame memory but also a complex address and sort circuit should be designed to complete the parallel access function.

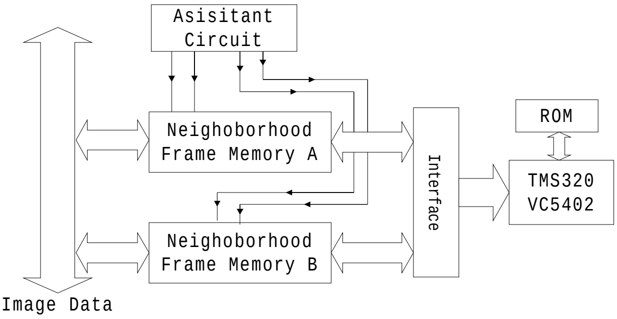


Figure 2. the diagram of processing module

Just as shown in the Figure 2, two neighborhood frame memories are necessary for each processing module. So they can complete the function that one is inputting data while the other is outputting data so that the dynamic image processing can be implemented. The assistant circuit includes address generator, address changer, memory clock and scan clock. The interface circuits include data exchange, data sort and data latch, and they just like tunnels for the image data to the DSP chip. The data bus that has operated as a DMA image communication bus, is designed for broadcast and reception of images with enough bandwidth for 3×3 neighborhood image data.So these highly structuralized image data can get parallel processing in the processing module and parallel data transforming.

3. EXPERIMENT WITH FFT

Since the traffic image processing, especially the vehicles image processing with complex background, anti-noise and image enhancement algorithms are used necessarily, and in these algorithms the fast Fourier transform (FFT) are usually used, so take the 2-D FFT for experiments. The 2-D can be defined by Eq. (1)

$$F(k_1,k_2)=\sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1} f(n_1,n_2)e^{-j(2\pi/N_1)K_1n_1}e^{-j(2\pi/N_2)K_2n_2} \quad (1)$$

Consider an image of size 512×512 pixels digitized to 8-bit accuracy.

In this system, the performance of the FFT can be determined by the execution timing and communication overhead. Total execution time can be derived considering the communication port bandwidth and the interconnection network, therefore the execution time is given by Eq. (2)

$$T = T_{FFT} + T_{DIST} + T_{REC} + T_{NFMP} \quad (2)$$

where T is total execution time. An image is distributed among the processors in time T_{DIST}. Initially, the neighborhood frame memory address time and the memory data sequence and exchange time are indicted by T_{NFMP}. The processed data are collected in time T_{REC}. T_{FFT} is the execution time of the 4 processors computing the 2-D FFT in parallelism without communication overhead. So if each processor's computing time is T_{FFTa}, T_{FFT} can be expressed as Eq. (3):

$$T_{FFT} = T_{FFTa} / 4 \quad (3)$$

The cost for neighborhood frame memory's parallel data is much smaller. Since all processing modules are interconnected as a complete network, image data can be transported between each node efficiently and quickly. In experiment, considering the both execution timing and communication overhead, this architecture of the system can get good performance for ITS real-time image processing.The system board only uses 3.3V and 5V input supply voltages, and the power consumption of the system is less than 20 W.

4. CONCLUSION

This system incorporates memory parallelism and processing parallelism, so it can get the maximum processing efficiency. So as a scalable, low power and compact image processing system, it can complete different kinds of image processing algorithm for ITS analysis in real-time, and the main criterions of performance excel other systems using pipeline process architectures.

5. REFERENCES

[1]. A.Y.H. Zomaya(Ed), Parallel and Distributed Computing Handbook McGraw-Hill, New York, 1996, pp. 1042-1070.

[2]. Su Guangda , The technology of parallel image processing , Tsinghua University Press, Beijing, 2002

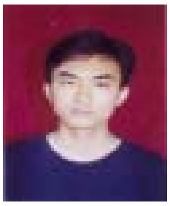
[3]. T.J.Fountain, Parallel Computing Principles and Practice. Cambridge University Press, Cambridge, 1994

[4]. D.W. Hammerstrom, D.P.Lulich, Image processing using one-dimensional processor arrays, Processing of IEEE, 84 (1996). 1005-1018.

[5]. M. Fikret Ercan, Yu-Fai Fung, M. Suleyman Demokan. Parallel image processing with one-dimensional DSP arrays. Future Generation Computer Systems 17 (2000) 197–214

[6]. David M.Harvey,Shirish P.Kshirsagar,C.Allan Hobson.Low cost scaleable parallel image processing system. Microprocessors and Microsysems 25(2001)143-157

Liu Zhi is a graduate student of the School of Information Science & Engineering Shandong University.His research interests are in high speed image processing and pattern recognition.



Zhu Wei is a graduate student in the School of Information Science & Engineering in Shandong University. He He has published several Journal papers during his study. His research interests are in Parallel Processing, the design of integrated circuit.Email: zhw81@sdu.edu.cn

The Hardware Designing for Real Time FPGA Based Image Processing

Tao Hongjiu Bao Yuliang Tong Xiaojun
Wuhan Polytechnic University, Wuhan 430023, P R China.
Wuhan University of Technology, Wuhan 430070, P.R.China.
Email: thjll@263.net

ABSTRACT

In this paper we present a high level software environment for FPGA-based image processing, which aims to hide hardware details as much as possible from the user. Our approach is to provide a very high level Image Processing Coprocessor with a core instruction set based on the operations of Image Algebra. The environment includes a generator which generates optimised architectures for specific user-defined operations.

Keywords: FPGA, Image, processing, DSP, Hardware.

1. INTRODUCTION

Image Processing application developers require high performance systems for computationally intensive Image Processing (IP) applications, often under real time requirements. In addition, developing an IP application tends to be experimental and interactive. This means the developer must be able to modify, tune or replace algorithms rapidly and conveniently.

Because of the local nature of many low level IP operations (e.g. neighbourhood operations), one way of obtaining high performance in image processing has been to use parallel computing. However, multiprocessor IP systems have generally speaking not yet fulfilled their promise. This is partly a matter of cost, lack of stability and software support for parallel machines; it is also a matter of communications overheads particularly if sequences of images are being captured and distributed across the processors in real time.

A second way of obtaining high performance in IP applications is to use Digital Signal Processing (DSP) processors [1,2]. DSP processors provide a performance improve-ment over standard microprocessors while still maintaining a high level programming model. However, because of the software based control, DSP processors have still difficulty in coping with real time video processing.

At the opposite end of the spectrum lie the dedicated hardware solutions. Application Specific Integrated Circuits (ASICs) offer a fully customised solution to a particular algorithm [3]. However, this solution suffers from a lack of flexibility, plus the high manufacturing cost and the relatively lengthy development cycle.

Reconfigurable hardware solutions in the form of FPGAs [4] offer high performance, with the ability to be electrically reprogrammed dynamically to perform other algorithms. Though the first FPGAs were only capable of modest integration levels and were thus used mainly for glue logic and system control, the latest devices [5] have crossed the Million gate barrier hence making it possible to implement an entire System On a Chip. Moreover, the introduction of the latest IC fabrication techniques has increased the maximum speed at which FPGAs can run. Design's performance exceeding 150MHz are no longer outside the realm of possibilities in the new FPGA parts, hence allowing FPGAs to address high bandwidth applications such as video processing.

A range of commercial FPGA based custom computing systems includes: the Splash-2 system [6]; the G-800 system [7] and VCC's HOTWorks HOTI & HOTII development. Though this solution seems to enjoy the advantages of both the dedicated solution and the software based one, many people are still reluctant to move toward this new technology because of the low level programming model offered by FPGAs. Although behavioural synthesis tools have made enormous progress, structural design techniques (including careful floorplanning) often still result in circuits that are substantially smaller and faster than those developed using only behavioural synthesis tools.

In order to bridge the gap between these two levels, this paper presents a high level software environment for an FPGA-based Image Processing machine, which aims to hide the hardware details from the user. The environment generates optimised architectures for specific user-defined operations, in the form of a low level netlist. Our system uses Prolog as the basic notation for describing and composing the basic building blocks. Our current implementation of the IPC is based on the Xilinx 4000 FPGA series.

The paper first outlines the programming environment at the user level (the programming model). This includes facilities for defining low level Image Processing algorithms based on the operators of Image Algebra [8], without any reference to hardware details. Next, the design of the basic building blocks necessary for implementing the IPC instruction set is presented. Then, we describe the runtime execution environment.

2. THE USER'S PROGRAMMING MODEL

At its most basic level, the programming model for our image processing machine is a host processor (typically a PC programmed in C++) and an FPGA-based Image Processing Coprocessor (IPC) which carries out complete image operations (such as convolution, erosion etc.) as a single coprocessor instruction. The instruction set of the IPC provides a core of instructions based on the operators of Image Algebra. The instruction set is also *extensible* in the sense that new compound instructions can be defined by the user, in terms of the primitive operations in the core instruction set. (Adding a new primitive instruction is a task for an architecture designer).

Many IP neighbourhood operations can be described by a template (a static window with user defined weights) and one

* This work is supported by Natural Science Foundation under Grant 2004 of Hubei Province and Foundation of Hubei Provincial Department of Education under Grant 2004Q001, China.

of a set of Image Algebra operators. Indeed, simple neighbourhood operations can be split in two stages:

A ‘local’ operator applied between an image pixel and the corresponding window coefficient.

A ‘global’ operator applied to the set of intermediate results of the local operation, to reduce this set to a single result pixel.
The set of local operators contains ‘Add’ (‘+’) and ‘multiplication’ (‘*’), whereas the global operator contains ‘Accumulation’ (‘Σ’), ‘Maximum’ (‘Max’) and ‘Minimum’ (‘Min’). With these local and global operators, the following neighbourhood operations can be built:

Neighbourhood Operation	*Local operation*	*Global operation*
Convolution	*	
Additive maximum	+	Max
Additive minimum	+	Min
Multiplicative maximum	*	Max
Multiplicative minimum	*	Min

For instance, a simple Laplace operation would be performed by doing convolution (i.e. *Local Operation* = ‘ · ’ and *Global operation* = ‘*’) with the following template:

~	-1	~
-1	4	-1
~	-1	~

The programmer interface to this instruction set is via a C++ class. First, the programmer *creates* the required instruction object (and its FPGA configuration), and subsequently *applies* it to an actual image. Creating an instruction object is generally in two phases: firstly build an object describing the operation, and then generate the configuration, in a file. For neighbourhood operations, these are carried out by two C++ object constructors:

image_operator (template & operator details)
image_instruction (operator object, filename)

For instructions with a single template operator, these can be conveniently combined in a single constructor:

Neighbourhood_instruction (template, operators, filename)

The details required when building a new image operator object include:

The dimension of the image (e.g. 256 × 256)
The pixels size (e.g. 16 bits).
The size of the window (e.g. 3 × 3).
The weights of the neighbourhood window.
The target position within the window, for aligning it with the image pixels (e.g. 1,1).
The ‘local’ and ‘global’ operations.

Later, to apply an instruction to an actual image, the *apply* method of the instruction object is used:

Result = instruction_object.*apply* (input image)

This will reconfigure the FPGA (if necessary), download the input pixel data and store the result pixels in the RAM of the IPC as they are generated.

3. ARCHITECTURES FROM OPERATIONS

When a new *Image_instruction* object (e.g. *Neighbourhood_instruction*) is created (by new), the corresponding FPGA configuration will be generated dynamically. In this section, we will present the structure of the FPGA configurations necessary to implement the high level instruction set for the neighbourhood operations described above. As a key example, the structure of a general 2-D convolver will be presented. Other neighbourhood operations are essentially variations of this, with different local and global operators sub-blocks.

(1) A general 2D convolver： As mentioned earlier, any neighbourhood image operation involves passing a 2-D window over an image, and carrying out a calculation at each window position.

To allow each pixel to be supplied only once to the FPGA, internal line delays are required. These synchronise the supply of input values to the processing elements, ensuring that all the pixel values involved in a particular neighbourhood operation are processed at the same instant[9, 10].

(2) Architecture of a Processing Element： Before deriving the architecture of a Processing Element, we first have to decide which type of arithmetic to be used- either bit parallel or bit serial processing.

While parallel designs process all data bits simultaneously, bit serial ones process input data one bit at a time. The required hardware for a parallel implementation is typically ‘n’ times the equivalent serial implementation (for an n-bit word). On the other hand, the bit serial approach requires ‘n‘ clock cycles to process an n-bit word while the equivalent parallel one needs only one clock cycle. However, bit serial architectures operates at a higher clock frequency due to their smaller combinatorial delays. Also, the resulting layout in a serial implementation is more regular than a parallel one, because of the reduced number of interconnections needed between PEs (i.e. less routing stress). This regularity feature means that FPGA architectures generated from a high level specification can have more predictable layout and performance. Moreover, a serial architecture is not tied to a particular processing word length. It is relatively straightforward to move from one word length to another with very little extra hardware (if any). For these reasons, we decided to implement the IPC hardware architectures using serial arithmetic.

Note, secondly, that the need to pipeline the bit serial Maximum and Minimum operations common in Image Algebra suggests we should process data Most Significant Bit first (MSBF). Following on from this choice, because of problems in doing addition MSBF in 2’s complement, there are certain advantages in using an alternative number representation to 2’s complement. For the purposes of the work described in this paper, we have chosen to use a redundant number representation in the form of a radix-2 Signed Digit Number system (SDNR) [11]. Because of the inherent carry-free property of SDNR add/subtract operations, the corresponding architectures can be clocked at high speed. There are of course several alternative representations which could have been chosen, each with their own advantages. However, the work presented in this paper is based on the following design choices:

Bit serial arithmetic
Most Significant Bit First processing
Radix-2 Signed Digit Number Representation (SDNR) rather than 2's complement.

Because image data may have to be occasionally processed on the host processor, the basic storage format for image data is still, however, 2's complement. Therefore, processing elements first convert their incoming image data to SDNR. This also reduces the chip area required for the line buffers (in which data is held in 2's complement). A final unit to convert a SDNR result into 2's complement will be needed before any results can be returned to the host system.

(3)Design of the Basic Building Blocks： In what follows, we will present the physical implementation of the five basic building blocks stated in section 2 (the adder, multiplier, accumulator and maximum/ minimum units). These basic components were carefully designed in order to fit together with as little wastage as possible.

(4)The 'multiplier' unit: The multiplier unit used is based on a hybrid serial-parallel multiplier outlined in [12]. It multiplies a serial SDNR input with a two's complement parallel coefficient B=$b_N b_{N-1} \ldots b_1$. The multiplier has a modular, scaleable design, and comprises four distinct basic building components [13]: *Type A*, *Type B*, *Type C* and *Type D*. An N bit coefficient multiplier is constructed by:

$$Type\ A \rightarrow Type\ B \rightarrow (N\text{-}3)*TypeC \rightarrow Type\ D$$

The coefficient word length may be varied by varying the number of type C units. On the Xilinx 4000 FPGA, *Type A*, *B* and *C* units occupy one CLB, and a *Type D* unit occupies 2 CLBs. Thus an N bit coefficient multiplier is 1 CLB wide and N+1 CLBs high. The online delay of the multiplier is 3.

(5)The 'accumulation' global operation unit: The accumulation unit is the global operation used in the case of a convolution. It adds two SDNR operands serially and outputs the result in SDNR format.

(6)The 'Addition' local operation unit: This unit is used in additive/maximum and additive/minimum operations. It takes a single SDNR input value and adds it to the corresponding window template coefficient. The coefficient is stored in 2's complement format into a RAM addressed by a counter whose period is the pixel word length. To keep the design compact, we have implemented the counter using Linear Feedback Shift Registers (LFSRs). The coefficient bits are preloaded into the appropriate RAM cells according to the counter output sequence. The input SDNR operand is added to the coefficient in bit serial MSBF.

(7)The Maximum/Minimum unit: The Maximum unit selects the maximum of two SDNR inputs presented to its input serially, most significant bit first.

4. THE COMPLETE ENVIRONMENT

In the front end, the user programs in a high level software environment (typically C++) or can interact with a Dialog-based graphical interface, specifying the IP operation to be carried out on the FPGA in terms of *Local* and *Global* operators, window template coefficients etc. The user can also specify:

The desired operating speed of the circuit.
The input pixel bit-length.
Whether he or she wants to use our floorplanner to place the circuit or leave this task to the FPGA vendor's Placement and Routing tools.

The system provides the option of two output circuit description formats: EDIF netlist (the normal), and VHDL at RTL level.

Behind the scenes, when the user gives all the parameters needed for the specific IP operation, the intermediate HIDE code is generated. Depending on the choice of the output netlist format, the HIDE code will go through either the EDIF generator tool to generate an EDIF netlist, or the VHDL generator tool to generate a VHDL netlist. In the latter case, the resulting VHDL netlist needs to be synthesised into an EDIF netlist by a VHDL synthesiser tool. Finally, the resulting EDIF netlist will go through the FPGA vendor's specific tools to generate the configuration bitstream file. The whole process is invisible to the user, thus making the FPGA completely hidden from the user's point of view. Note that the resulting configuration is stored in a library, so it will not be regenerated if exactly the same operation happens to be defined again.

Complete and efficient configurations have been produced from our high level instruction set for all the Image Algebra operations and for a variety of complex operations including 'Sobel', 'Open' and 'Close'. They have been successfully simulated using the Xilinx Foundation Project Manager CAD tools.

5. CONCLUSIONS

In this paper, we have presented the design of an FPGA-based Image Processing Coprocessor (IPC) along with its high level programming environment. The coprocessor instruction set is based on a core level containing the operations of Image Algebra. Architectures for user-defined compound operations can be added to the system. Possibly the most significant aspect of this work is that it opens the way to image processing application developers to exploit the high performance capability of a direct hardware solution, while programming in an application-oriented model. Figures presented for actual architectures show that real time video processing rates can be achieved when staring from a high level design.

The work presented in this paper is based specifically on Radix-2 SDNR, bit serial MSBF processing. In other situations, alternative number representations may be more appropriate. Sets of alternative representations are being added to the environment, including a full bit parallel implementation of the IPC [23]. This will give the user a choice when trying to satisfy competing constraints.

Although our basic approach is not tied to a particular FPGA, we have implemented our system on XC4000 FPGA series. However, the special facilities provided by the new Xilinx VIRTEX family (e.g. large on-chip synchronous memory, built in Delay Locked Loops etc.) make it very suitable target architecture for this type of application. Upgrading our system to operate on this new series of FPGA chips is underway.

6. REFERENCES

[1]. Rajan, K, Sangunni, K S and Ramakrishna, J, 'Dual-DSP systems for signal and image-processing', Microprocessing & Microsystems, Vol 17, No 9, pp 556-560, 1993.

[2]. Akiyama, T, Aono, H, Aoki, K, et al, 'MPEG2 video codec using Image compression DSP', IEEE Transactions on Consumer Electronics, Vol 40, No 3, pp 466-472, 1994.

[3]. L.A. Christopher, W.T. Mayweather and S.S. Perlman, 'VLSI median filter for impulse noise elimination in composite or component TV signals', IEEE Transactions on Consumer Electronics, Vol 34, no. 1, pp. 263-267, 1988.

[4]. J. Rose and A. Sangiovanni-Vincentelli, 'Architecture of Field Programmable Gate Arrays', Proceedings of the IEEE Volume 81, No7, pp 1013-1029, 1993.

[5]. Arnold, J M, Buell, D A and Davis, E G, 'Splash-2', Proceedings of the 4th Annual ACM Symposium on Parallel Algorithms and Architectures, ACM Press, pp 316-324, June 1992.

[6]. Gigaops Ltd., The G-800 System, 2374 Eunice St. Berkeley, CA 94708.

[7]. Chan, S C, Ngai, H O and Ho, K L, 'A programmable image processing system using FPGAs', International Journal of Electronics, Vol 75, No 4, pp 725-730, 1993.

[8]. Ritter G X, Wilson J N and Davidson J L, 'Image Algebra: an overview', Computer Vision, Graphics and Image Processing, No 49, pp 297-331, 1990.

[9]. Shoup, R G, 'Parameterised Convolution Filtering in an FPGA', More FPGAs, W Moore and W Luk (editors), Abington, EE&CS Books, pp 274, 1994.

[10]. Kamp, W, Kunemund, H, Soldner and Hofer, H, 'Programmable 2D linear filter for video applications', IEEE Journal of Solid State Circuits, pp 735-740, 1990.

[11]. Avizienis A, 'Signed Digit Number Representation for Fast Parallel Arithmetic", IRE Transactions on Electronic Computer, Vol. 10, pp 389-400, 1961.

[12]. Moran, J, Rios, I and Meneses, J, 'Signed Digit Arithmetic on FPGAs', More FPGAs, W Moore and W Luk (editors), Abington, EE&CS Books, pp 250, 1994.

[13]. Donachy, P, 'Design and implementation of a high level image processing machine using reconfigurable hardware', PhD Thesis, Department of Computer Science, The Queen's University of Belfast, 1996.

Tao Hongjiu is a full professor at the Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan, Hubei Province, China. He graduated from Tianjin University of Technology in 1980; and earned Engineering Doctor Degree in Pattern Recognition and Artificial Intelligence, at the Institute for Image Processing and Intelligent Control，from Huazhong University of Science and Technology in 2003, China. He has published one book, over 30 Journal papers. His research interests are in the computer image processing, pattern recognition, streaming media technology, digital signal processing etc.

The Ascertainment of Scale Sampling Step for Numerical Realization Adopting Binary Dot-and-Grid Sampling of the Continuous Wavelet Transform

Pu Yifei Yuan Xiao Liao Ke Zhou Jiliu
College of Electronics and Information Sichuan University, Sichuan University
Chengdu, Sichuan Province, China, 610064
E_mail: Puyifei_007@163.com **Tel:** 028-89869346

ABSTRACT

One of the key discrete approaches for the numerical realization of the continuous wavelet transform is to ascertain the scale sampling step, which is the requirement and basic for wavelet analytic engineering achievement and research in theory. This article disserts the basic theory for the numerical realization of the continuous wavelet transform. According to the highest numeric angle frequency of signal is or below π, in this two cases the article respectively deduces the best results of the scale sampling step of Morlet mother wave and Gauss function's each phases differential coefficient of even or odd symmetric mother wave in the numerical realization of the continuous wavelet transform when binary dot-and-grid sampling is adopted. In the end, it discusses the time shift required in getting the results of odd symmetric mother wave in the numerical realization of the continuous wavelet transform. These useful conclusions solve a fundamental issue of wavelet analyses in engineering practice and research in theory.

Keywords: the scale sampling step, binary dot-and-grid sampling, ripple coefficient, symmetry of wavelet, numerical wavelet filter, time shift

1. QUESTION TO PUT FORWARD

It is well known that the continuous wavelet transform equation is as (1) and (2), in which τ represents scaling time, a is scaling factor, ψ is mother wave, $\psi_{\tau,a}$ is basic wavelet [1]. The article takes the energy-constant wavelet as example to discuss in generic cases.

$$\psi_{\tau,a}(t) = 1/\sqrt{a}\,\psi(\frac{t-\tau}{a}) \Leftrightarrow \hat{\psi}_{\tau,a}(\Omega) = \sqrt{a}\,\psi(\hat{a}\Omega)\,e^{-j\Omega\tau} \tag{1}$$

$$s(t) \Leftrightarrow W_s(\tau, a) = \int_t s(t)\psi^*_{\tau,a}(t)dt = 1/\sqrt{a}\int_t s(t)\psi^*(\frac{t-\tau}{a})dt \tag{2}$$

The results of the continuous wavelet transform $W_s(\tau,a)$can be approximately worked out through numerical integral calculation by computer in the most cases, except few signals that can be analytically expressed might calculated through analytic integral calculation [2][3]. At the same time, the signal we got usually is numerical signal $s[k],\, k\in\mathbf{Z}$ or the serial signal sampling from continuous time signal $s(t), t\in\boldsymbol{R}$ using recurrent Dirac function $\delta_{T_s}(t)$. So, the approximate equation of $W_s(\tau,a)$ is as (3).

$$W_s(\tau, a) \approx \frac{T_S}{\sqrt{a}}\sum_k s[kT_S]\psi^*\left(\frac{kT_S-\tau}{a}\right) \tag{3}$$

$W_s(\tau, a)$ includes super information redundancy , and the simplest way to get rid of them is discretion [4]. Similarly, scale time τ and step a will also be discrete. One of discretion way for the phase plane of time-frequency is binary dot-and-grid sampling, see as equation (4).

$$\begin{cases} a = 2^n a_0 \\ \tau = m\tau_0 \end{cases} \quad m\in\mathbf{Z}, n\in\mathbf{N} \tag{4}$$

For length limited, the article only deduces the ascertainment of scale sampling step a_0 when binary dot-and-grid sampling is adopted. Take (4) into (3), and make it unit ($\tau_0 = T_S = 1$), then get equation (5).

$$\tilde{\tilde{W}}_s(m, n) = \frac{1}{\sqrt{2^n a_0}}\sum_k s[k]\psi^*\left(\frac{k-m}{2^n a_0}\right) \tag{5}$$

Predigest equation (5), then get equation (6). g_n of equation (6) is shown as equation (7) which is called numerical wavelet filter. In which, n is scale, m is scan time and a_0 is scale sampling step.

$$\tilde{\tilde{W}}_s(m, n) = s[m] * g_n[m] = \sum_l s[l]g_n[m-l] \tag{6}$$

$$g_n:\ g_n[m] = \frac{1}{\sqrt{2^n a_0}}\psi^*\left(-\frac{m}{2^n a_0}\right) \quad m\in\mathbf{Z},\ n = 1,2,3,\Lambda \tag{7}$$

The utmost issue of approximately calculation for $\tilde{\tilde{W}}_s(m, n)$ by the way of numerical realization is to ascertain scale- sampling step.

2. CHOOSE CONTINUOUS WAVELET

Symmetry is necessary for the engineering practice. According to the mother wave's symmetry, it can be divided as complex analytic, even symmetric and odd symmetric mother wave [1][3]. As to different mother wave ψ, we will start from the simplest case when $n=1$. See as equation (5). When $\tau_0=1$, the frequency spectrum density $S(e^{j\omega})$ of limited discrete signal $s(k)$ is uninterrupted and periodic function. The period of $S(e^{j\omega})$ is $\omega_T = 2\pi$ [5]. We form mother wave as follows: $\psi(t)=g(t)e^{j\Omega_0 t} \Leftrightarrow \hat{\psi}(\Omega)=g(\Omega-\Omega_0)$, in which $g(t)$ is window function of Gabor transform [1][6]. Therefore, in order to exactly calculate the $\tilde{\tilde{W}}_s(m, 1)$ that is

$W_s(\tau,a)$ when $(m\tau_0,\, 2^n a_0)\big|_{\tau_0=1,n=1,} \Rightarrow (m,2a_0) \Rightarrow (m,2\times 1)$, it must make the corresponding numerical filter function $G_1(\omega)=\sum_m g_1[m]\mathrm{e}^{-\mathrm{j}\omega m}$ equal to the repeated sum with 2π period of wavelet filter function $\hat{\psi}^*_{2a_0}(\Omega)=\sqrt{2a_0}\hat{\psi}^*(2a_0\Omega)=\sqrt{2a_0}\hat{g}^*(2a_0\Omega-\Omega_0),\, a_0>0$ when scale $n=1 \Rightarrow a_{n=1}=2a_0$. See as equation (8).

$$G_1(\omega)=\sum_l \sqrt{2a_0}\hat{\psi}^*[2a_0(\omega-2\pi l)]=\sum_l \sqrt{2a_0}\,\hat{g}{}^*[2a_0(\omega-2\pi l)-\Omega_0] \tag{8}$$

3. DEDUCE OF SCALE SAMPLING STEP OF COMPLEX ANALYTIC MOTHER WAVE

Academic Analysis

The spectrum of complex analytic mother wave $\hat{\psi}_A(\Omega)$ usually is single-apex and fast-descending function [6][7], seen as Figure 1. In which Ω_c is the frequency at the apex, Ω_m is the highest frequency. When scale $a=2a_0$, namely $n=1$, the corresponding numerical filter function $G_1(\omega)$ is seen as Figure 2.

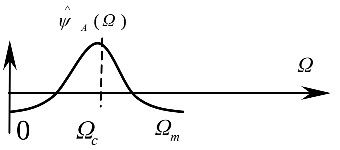


Figure 1 Frequency Spectrum of Complex Analytic Mother Wave

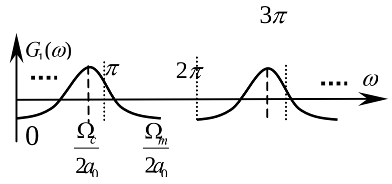


Figure 2 Corresponding Frequency Spectrum of Numerical Filter of Complex Analytic Mother Wave

From Figure 2 we can see that the frequency at the apex of $G_1(\omega)$ is $\omega_c=\frac{\Omega_c}{2a_0}$ and the highest frequency is $\omega_m=\frac{\Omega_m}{2a_0}$ [3][6]. It is well-known that the highest numerical angle frequency of numerical signal is $\omega=\pi$[2]. So, to analyze the ingredient of the highest frequency of signal $s[k]$, it is usually using the numerical filter's apex to analyze the highest frequency of signal, that is $\omega_c \ge \pi$, then $a_0 \le \Omega_c/2\pi$. However, to avoid over-reducing a_0 and make the highest frequency ω_m of $G_1(\omega)$ above 2π and making spectrum superposed and intermixed, it must $\omega_m \le 2\pi$, then $a_0 \ge \frac{\Omega_m}{4\pi}$. And gets equation (9).

$$\frac{\Omega_m}{4\pi} \le a_0 \le \frac{\Omega_c}{2\pi} \tag{9}$$

Only when $\Omega_m \le 2\Omega_c$, equation (9) is correct. Because $\hat{\psi}_A(\Omega)$ is single-apex and fast-descending function, only when Ω_m is the biggest in its range can ensure $\frac{\overset{\wedge}{\psi}_A(\Omega_m)}{\overset{\wedge}{\psi}_A(\Omega_c)}$ is the smallest [8]. To do this can reduce spectrum superposition as much as possible. So, when $\Omega_m=2\Omega_c$, then gets the best a_0, seen as equation (10).

$$a_0=\frac{\Omega_c}{2\pi} \tag{10}$$

In addition, if $s[k], k\in \boldsymbol{Z}$ is the sequence sampled from continuous time signal $s(t), t\in \boldsymbol{R}$, suppose $\Omega_{s\max}$ is the highest frequency of $s(t)$ and sampling frequency f_S, then get the highest numeric frequency of $s[k]$ is $\omega_{\mathrm{sm}}=\Omega_{s\max}/f_s$ [2].

Obviously it might have the case $\omega_{\mathrm{sm}}<\pi$[2]. Only when $\omega_c \ge \omega_{sm}$ can analyze the ingredient of the highest frequency, then get $a_0 \le \Omega_c f_s/2\Omega_{s\max}$. And for $a_0 \ge \frac{\Omega_m}{4\pi}$, then get equation (11).

$$\frac{\Omega_m}{4\pi} \le a_0 \le \frac{\Omega_c f_s}{2\Omega_{s\max}} \tag{11}$$

Only when $\Omega_m \le 2\pi\Omega_c f_s/\Omega_{s\max}$, equation (11) is correct. If $\Omega_m=2\pi\Omega_c f_s/\Omega_{s\max}$, then get the best a_0, see as equation (12).

$$a_0=\frac{\Omega_c f_s}{2\Omega_{s\max}} \tag{12}$$

Ascertain The Scale Sampling Step of Morlet Mother Wave

Morlet mother wave is $\psi_m(t)=1/\sqrt{2\pi}e^{-\frac{1}{2}t^2}e^{j\Omega_0 t} \Leftrightarrow \overset{\wedge}{\psi}_m(\Omega)=e^{-\frac{1}{2}(\Omega-\Omega_0)^2}$, when $\Omega_0 \ge 5$, Morelet mother wave approximately satisfied the fluctuation. $\hat{\psi}_m(\Omega)$ is single-apex and fast-descending function. Let $\frac{d}{d\Omega}\hat{\psi}_m(\Omega)=0$, then get the frequency at apex is $\Omega_c=\Omega_0$. According equation (10) and (12), gets table 1.

Table 1 The best scale-sampling step of Morlet mother wave

$\psi_m(t)$	when $\omega_{\mathrm{sm}}=\pi$	when $\omega_{\mathrm{sm}}<\pi$
$1/\sqrt{2\pi}\,e^{-\frac{1}{2}t^2}e^{j\Omega_0 t}$	$a_0=\frac{\Omega_0}{2\pi}$	$a_0=\frac{\Omega_0 f_s}{2\Omega_{s\max}}$

4. DEDUCE THE SCALE SAMPLING STEP OF REAL EVEN SYMMETRIC MOTHER WAVE

Academic Analysis

The time domain of real even symmetric mother wave is $\psi_{\mathrm{E}}(-t)=\psi_{\mathrm{E}}(t)$, the spectrum $\hat{\psi}_{\mathrm{E}}(\Omega)$ is double-apex and fast-descending even function. See as Figure 3.

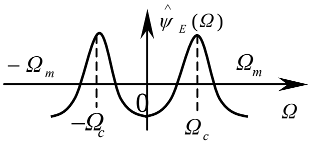


Figure 3 Frequency Spectrum

When the highest numeric angle frequency $\omega=\pi$, if $\omega_m=\frac{\Omega_{\mathrm{m}}}{2a_0}\le\pi\Rightarrow a_0\ge\frac{\Omega_{\mathrm{m}}}{2\pi}$, the corresponding numerical filter function $G_1(\omega)$ see as Figure 4. It doesn't intermix. The coefficient of filter g_1 is the over-sampling or critical sampling,

$$g_1:g_1[m]=\psi_E^*(-m/2a_0)\big/\sqrt{2a_0}=\psi_{\mathrm{E}}(m/2a_0)\big/\sqrt{2a_0}\ .$$

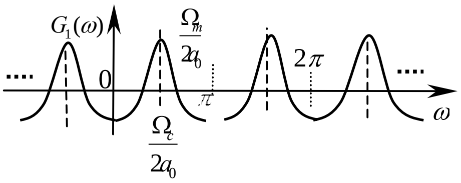


Figure 4 Frequency Spectrum of Numerical Filter of Even Symmetric Mother Wave when a_0 on the Large Side.

Just like the analysis of complex analytic mother wave, we can get the best a_0 as equation (13).

$$a_0=\frac{\Omega_{\mathrm{c}}}{2\pi} \tag{13}$$

In this case, the frequency spectrum $G_1(\omega)$ of numerical filter is fall in line with itself at the apex. Though there is superposed, it can analyze the ingredient of the signal's highest angle frequency, see as Figure 5.

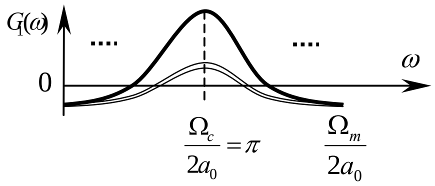


Figure 5 Frequency Spectrum of Numerical Filter of Real Even Symmetric Mother Wave when a_0 on the Small Side

In addition, when the highest numeric frequency $\omega_{\mathrm{sm}}=\Omega_{s\max}/f_s<\pi$, just like the analysis of complex analytic mother wave, then get the best a_0, see as equation (14).

$$a_0=\frac{\Omega_{\mathrm{c}}f_s}{2\Omega_{s\max}} \tag{14}$$

Ascertain Scale Sampling Step Of Analytic Mother Wave of Each Phases Differential Coefficient of Even Symmetric Gauss Function

Because Gauss signal is the best signal according with Uncertainty principle [3], to deduce scale sampling step a_0 of analytic mother wave of each phases differential coefficient of Gauss function has been widely needed in engineering. Complex analytic mother wave ${}_k\psi$ formed by phases differential coefficient of Gauss function, whose spectrum see as equation (15). In which, $A(k)$ is a energy-unit constant and depends on the variable k[7][9].

$${}_k\hat{\psi}(\Omega)=\begin{cases}A(k)\Omega^k\exp\left(-\frac{1}{2}\Omega^2\right) & \Omega\ge0\\ 0 & \Omega<0\end{cases} \tag{15}$$

Suppose $\frac{d}{d\Omega^k}{}_k\hat{\psi}(\Omega)=0$, then gets frequency at apex is $\Omega_c=\sqrt{k}$.

When $\sqrt{k}=\Omega_c<\Omega_m\le2\Omega_c=2\sqrt{k}$. If $\frac{{}_k\hat{\psi}(\Omega_m)}{{}_k\hat{\psi}(\Omega_c)}=(\frac{e}{k})^{\frac{k}{2}}\cdot\frac{\Omega_m^k}{e^{\frac{1}{2}\Omega_m^2}}=\frac{1}{m}$, m must be a constant with module over than 1. The equation can be worked out only by Newton method or fast-descending method [10] or Figureic method. For the equation includes linear part and exponent part [11]. Calculating as the above method will make Ω_{m} varied with m. So Ω_{m} is not always in the domain of Ω_{m} . If $\Omega_{\mathrm{m}}=2\Omega_c=2\sqrt{k}$, $\frac{{}_k\hat{\psi}(\Omega_m)}{{}_k\hat{\psi}(\Omega_c)}$ will be the smallest [8]. And try to reduce intermix of frequency spectrum [12]. When the highest frequency $\omega=\pi$, from equation (13) can get equation (16).

$$a_0=\frac{\Omega_c}{2\pi}=\frac{\sqrt{k}}{2\pi}\qquad k=2q,q\in N \tag{16}$$

For after Fourier transform Gauss signal is unchanged, the time domain of mother wave $\psi(t)$ can be expressed as equation (17). In which, j is a unit of unreliable figure and k is even

$$\psi(t)=(-1)^{k-1}j^k(te^{-\frac{1}{2}t^2})^{(k-1)} \tag{17}$$

Besides, if the highest numeric frequency of $s[k]$ $\omega_{\mathrm{sm}}=\Omega_{s\max}/f_s<\pi$, from equation (14) get equation (18). See as table 2.

$$a_0=\frac{\Omega_{\mathrm{c}}f_s}{2\Omega_{s\max}}=\frac{\sqrt{k}f_s}{2\Omega_{s\max}}\qquad k=2q,q\in N \tag{18}$$

Table 2 The best a_0 of even analytic mother wave of each phases differential coefficient of Gauss function

k	$\psi_E(t)$	a_0 when $\omega_{\mathrm{sm}}=\pi$	a_0 when $\omega_{\mathrm{sm}}<\pi$
2	$(1-t^2)e^{-\frac{1}{2}t^2}$	$\frac{\sqrt{2}}{2\pi}$	$\sqrt{2}f_s/2\Omega_{s\max}$
4	$(3-6t^2+t^4)e^{-\frac{1}{2}t^2}$	$\frac{1}{\pi}$	$f_s/\Omega_{s\max}$
6	$(15-45t^2+15t^4-t^6)e^{-\frac{1}{2}t^2}$	$\frac{\sqrt{6}}{2\pi}$	$\sqrt{6}f_s/2\Omega_{s\max}$
even	$(-1)^{k-1}j^k(te^{-\frac{1}{2}t^2})^{(k-1)}$	$\frac{\sqrt{k}}{2\pi}$	$\sqrt{k}f_s/2\Omega_{s\max}$

Ascertain The Scale Sampling Step Of Mexico Hat Wave:

When k = 2， the mother wave is Mexico hat mother wave. Suppose, $\frac{d}{d\Omega}{}_{hat}\hat{\psi}(\Omega)=0$, then get $\Omega_c=\sqrt{k}=\sqrt{2}$. From equation (16), get $a_0=\frac{\sqrt{2}}{2\pi}$. From equation (17), get its expression in time domain of mother wave $\psi_{hat}(t)$ is as equation (19).

$$\psi_{\mathrm{hat}}(t)=(1-t^2)\mathrm{e}^{-\frac{1}{2}t^2}\Leftrightarrow\hat{\psi}_{\mathrm{hat}}(\Omega)=\Omega^2\mathrm{e}^{-\frac{1}{2}\Omega^2} \qquad (19)$$

The expression in time domain of corresponding wavelet $\psi_{\mathrm{hat}a_0}(t)$ and numerical filter g_1 is seen as Figure 6.

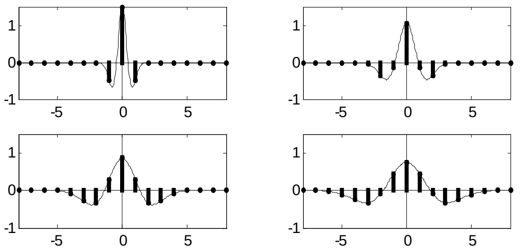


Figure 6 Mexico Hat Wave and Numerical Filter

Cut down g_1 and make it energy-unit, get equation (20). From equation (7), get corresponding numerical filter, see as equation (21).

$$g_1=\left\{\frac{\pi^2-8}{32-16\pi^2}e^{3\pi^2/16},\frac{1}{2},\frac{\pi^2-8}{32-16\pi^2}e^{3\pi^2/16}\right\} \qquad (20)$$

$$g_n:\ g_n[m]=\frac{1}{\sqrt{2^n\sqrt{2}/\pi}}\left[1-\left(\frac{m}{2^n\sqrt{2}/\pi}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{m}{2^n\sqrt{2}/\pi}\right)^2\right]\ n=2,3,4,\Lambda \qquad (21)$$

5. DEDUCE THE SCALE SAMPLING STEP OF REAL ODD SYMMETRIC MOTHER WAVE

Academic Analysis

The time domain of odd symmetric wave is $\psi_{\mathrm{O}}(-t)=-\psi_{\mathrm{O}}(t)$. Since the time domain of signal is odd symmetry, its frequency domain must be unreliable figure odd symmetric function, the frequency spectrum $\hat{\psi}_{\mathrm{E}}(\Omega)$ is double-apex unreliable figure odd function, see as Figure 7.

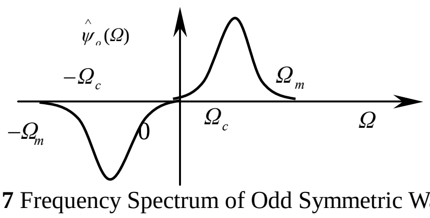
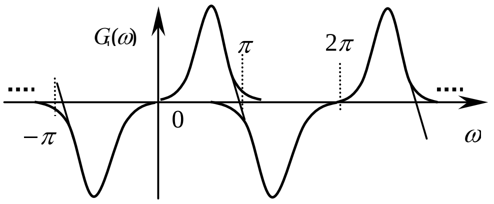


Figure 7 Frequency Spectrum of Odd Symmetric Wavelet

When $\omega_m=\frac{\Omega_m}{2a_0}\le\pi\Rightarrow a_0\ge\frac{\Omega_m}{2\pi}$, the corresponding numerical filter function $G_1(\omega)$ doesn't intermix, but can't analyze the ingredient of the highest frequency. Reduce a_0 and lack-sample the odd wavelet and make $a_0<\Omega_m/2\pi$. Though there will be intermixed, it's positive and negative apex counteracts in the mixed part. See as Figure 8.

Figure 8 Frequency Spectrum of Numerical Filter of Odd Symmetric Mother Wave



So it still can't do analyze the ingredient of highest numeric angle frequency.

Let's begin with an example. The simplest first-phase Gauss mother wave is ${}_1\psi(t)=-t\mathrm{e}^{-\frac{1}{2}t^2}\Leftrightarrow{}_1\hat{\psi}(\Omega)=\mathrm{j}\Omega\mathrm{e}^{-\frac{1}{2}\Omega^2}$. Suppose $\frac{d}{d\Omega}{}_1\hat{\psi}(\Omega)=0$, then get the frequency at apex is $\Omega_c=1$. At first, according to the above-mentioned methods to get a_0 of complex analytic and odd symmetric mother wave, see as equation (10) and (16). Then $a_0=\frac{\Omega_c}{2\pi}$, get equation (22).

$$a_0=1/\pi\Rightarrow{}_1\psi_{a_0}(t)=-\pi^{3/2}t\mathrm{e}^{-\frac{1}{2}(\pi t)^2}\Rightarrow h_1:h_1[i]=-\pi^{3/2}ie^{-\frac{1}{2}(\pi i)^2} \qquad (22)$$

Cut down g_1 and make energy-unit, get figure of ${}_1\psi_{a_0}(t)$ and g_1 see as Figure 8. Its change frequency is lower than π . Therefore, g_1 cann't be used to analyze the ingredient of the highest numeric angle frequency （$\omega=\pi$） . Imitate the type of haar mother wave [13] [14] [15], to let mother wave ${}_1\psi_{a_0}(t)$ shift to the right or left $\Omega_c/2=1/2$, then get ${}_1\psi_{a_0}(t\mu 1/2)$ and do binary dot-and-grid sampling. So its numeric angle frequency is π , it can used to analyze the ingredient of signal's highest numeric angle frequency. See as Figure 9&10

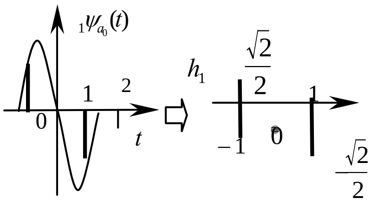


Figure 9 Direct sampling to even mother wave

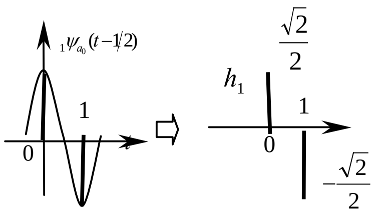


Figure 10 After moving $1/2$, sampling

Generalize the result, as to any odd symmetric mother wave $\psi_o(t)$, it needs shift to the left or right $\Omega_c/2$. Then get $\psi_{oa_0}(t\,\mu\Omega_c/2)$ and do binary dot-and-grid sampling the same as that of complex analytic and even symmetric mother wave to get equation（23）,（24）and（25）. In which, τ is time shift.

$$a_0 = \frac{\Omega_c}{2\pi} \tag{23}$$

$$a_0 = \frac{\Omega_c f_s}{2\Omega_{s\max}} \tag{24}$$

$$\tau = \frac{\Omega_c}{2} \tag{25}$$

Ascertain the Scale Sampling Step and Time Moving of Odd Analytic Mother Wave of Each Phases Differential Coefficient of Gauss Function

According to equation (17), we can form complex analytic mother wave $_k\psi$. From equation (23),（24） and (25)can get equation（26）,（27） and（28）.

$$a_0 = \frac{\Omega_c}{2\pi} = \frac{\sqrt{k}}{2\pi} \qquad k = 2q-1, q \in N \tag{26}$$

$$a_0 = \frac{\Omega_c f_s}{2\Omega_{s\max}} = \frac{\sqrt{k} f_s}{2\Omega_{s\max}} \qquad k = 2q-1, q \in N \tag{27}$$

$$\tau = \frac{\Omega_c}{2} = \frac{\sqrt{k}}{2} \qquad k = 2q-1, q \in N \tag{28}$$

See as table 3. In which, $\psi_o(t)$ gets rid of the unrelible figure unite j got from Fourier inverse transform to make sure it is the real function.

Table 3 Scale sampling step a_0 and time moving τ of analytic mother wave of each phase's differential coefficient of odd Gauss function

k	$\psi_o(t)$	a_0 When $\omega=\pi$	a_0 When $\omega<\pi$	τ
1	$t\,\mathrm{e}^{-\frac{1}{2}t^2}$	$\frac{1}{2\pi}$	$f_s/2\Omega_{s\max}$	$\frac{1}{2}$
3	$(3-t^2)t\mathrm{e}^{-\frac{1}{2}t^2}$	$\frac{\sqrt{3}}{2\pi}$	$\sqrt{3}f_s/2\Omega_{s\max}$	$\frac{\sqrt{3}}{2}$
5	$(15-10t^2+t^4)t\mathrm{e}^{-\frac{1}{2}t^2}$	$\frac{\sqrt{5}}{2\pi}$	$\sqrt{5}f_s/2\Omega_{s\max}$	$\frac{\sqrt{5}}{2}$
odd	$(-1)^{k-1}j^{k-1}(t\mathrm{e}^{-\frac{1}{2}t^2})^{(k-1)}$	$\frac{\sqrt{k}}{2\pi}$	$\sqrt{k}f_s/2\Omega_{s\max}$	$\frac{\sqrt{k}}{2}$

6. CONCLUSIONS

One of the key discrete approach for the numerical realization of the continuous wavelet transform is to ascertain the scale sampling step, which is the requirement and basic for wavelet analytic engineering achievement and research in theory. This article disserts the basic theory for the numerical realization of the continuous wavelet transform. These useful conclusions solve a fundamental issue of wavelet analyses in engineering practice and research in theory.

7. REFERENCES

[1] （France） Wrote by Ste'phane Mallat , Translated by Yang Lihua, Dai Daoqing, Huang Wenliang, Zan Qiuhui,A wavelet tour of signal processing, second edition, China Machine Press，2002.9

[2] （U.S.A） Alan V.Oppenheim, Signals and Systems, second edition, Publishing house of electronics Industry，2002.8

[3] （U.S.A） Wrote by CHARLES K.CHUI, Translated by Cheng Zhengxing, Audited by Bai Juxian, An Introduction to Wavelets, Xi'an Jiaotong University Press，1995.1

[4] (China) YuanXiao, Yu Juebang, On Orthogonality condition of Bubble Wavelet, Journal of UEST of China， Vol.27, No.1,Feb,1998.2,p25-28.

[5] (China) Chen Yuanheng, Information and signal basic , Higher Education Press，1989.5

[6] (China) YuanXiao,. On a novel class of complex analytic wavelet, ACTA ELECTRONICA SINCA ， Vol.28,No.4 ,April,2000.4, p123-126.

[7] (China) YuanXiao, Yu Juebang, Chen Xiangdong, Yang Jiade, Super-Gaussian Spectrum Functions and Their Time-frequency Localization Characteristics, ACTA ELECTRONICA SINCA，Vol.29,No.1,Jan,2001,p80-83.

[8] （France）wrote by Y.Mayer, translated by You Zhong, Wavelet and arithmetic operators, World Books Press company，1992.6

[9] (China) YuanXiao, Yu Juebang, Complex Analytical Wavelet Transform for the Extraction and Analysis of Speech Signal Envelops, ACTA ELECTRONICA SINCA ， Vol.27,No.5,May,1999,p142-144.

[10] （U.S.A）Wrote by B.Widrow, Tranlated by Wang Yongde, Long Xianhui, Self-adoption Signal Processing, Sichuan University Press ,1989.11

[11] （U.S.A）Wrote by D.Hughes-Hallett，A.M.Gleason,etc., Translated by Hu Naijiong,Sao Yong, etc., Calculus, Higher Education Press，2002.6

[12] （U.S.A） Wrote by L.Korn, Translated Bai Juxian, Time-Frequency Analysis, Xi'an Jiaotong University Press，1998

[13] (China) Zhang Gongqing, Lin Yuanqu, Teaching Materials of Functional Analysis, Peking University Press，2003.1

[14] (China) Zhang Xianda, Modern Signal Processing, second edition, Qtsinghua University Press，2002.10

[15] (China) Tao Deyuan, Yuan Xiao, He Xiaohai, Time-frequency Localization Characteristics of A Novel Class of Complex Wavelet, Journal of UEST of China，Vol.30, No.1, Feb, 2001.2, p21-25.

Pu Yifei is of College Electronics and Information Sichuan University, Chengdu, Sichuan Province, China. His birthday is 1975.2.23. His Major is Communication and Information system. His main research field is image processing and network communication, communication and information processing, wavelet transform, fraction calculation, ANN, GA.

Application of Parallel Distributed Technology in Simulation Engineering

Wang Xuehui[1], Zhang Lei[2]
School of Mechatronics Engineering and Automation[1] & Department of Computer Science[2],
National University of Defense Technology
Changsha, Hunan Province, P.R.China
Email: yzmailbox2003@163.com, zlmailbox2000@163.com **Tel.:** 13617496807, 13574838837

ABSTRACT

Parallel distributed technology is a widely accepted and extensively used method for various aspects of engineering and science field, especially in simulation study. An overview of technologies concerned with distributing the execution of simulation programs across multiple processors is presented. First of all, we introduce what is distributed simulation and how it develop today; Following, we provides an emphatic introduction to High Level Architecture (HLA); The remainder of this paper is focused on time management, which is a central issue concerning the synchronization of distributed simulation.

Keywords: Parallel distributed technology, distributed simulation, HLA, time management, synchronization

1. INTRODUCTION

1.1. What is Distributed Simulation

Simulation is a powerful tool for the analysis of new system designs, retrofits to existing systems and proposed changes to operating rules [1]. Manipulating a high-efficient simulation is not any an art but also a science. There are many types and kinds of simulation. In this paper we limit ourselves to distributed simulation [1]. Here, the term distributed simulation refers to distributing the execution of a single “run” of a simulation program across multiple processors. This encompasses several different domains. One domain concerns the motivation for distributing the execution. One paradigm, often referred to as parallel simulation [2], concerns the execution of the simulation on a tightly coupled computer system, e.g., a supercomputer or a shared memory multi-processor. Here, the principal reason for distributing the execution is to reduce the length of time to execute the simulation. In principal, by distributing the execution of a computation across N processors, one can complete the computation up to N times faster than if it were executed on a single processor. Another reason for distributing the execution in this fashion is to enable larger simulations to be executed than could be executed on a single computer. When confined to a single computer system, there may not be enough memory to perform the simulation. Distributing the execution across multiple machines allows the memory of many computer systems to be utilized [3].

Another dimension that differentiates distributed simulation paradigms is the geographical extent [3] over which the simulation executes. Often distributed simulations are executed over broad geographic areas. This is particularly useful when personnel and/or resources (e.g., databases or specialized facilities) are included in the distributed simulation exercise. Distributed execution eliminates the need these personnel and resources to be physically collocated, making for an enormous cost savings. Distributed simulations operating over the Internet have created an enormous market for the electronic industry.

Historically, the term distributed simulation has often been used to refer to geographically distributed simulations [1], while parallel simulation traditionally referred to simulations executed on a tightly coupled parallel computer, however, with new computing paradigms such as clusters of workstations and grid computing, this distinction has become less clear, so we use the single term distributed simulation here to refer to all categories of distributed execution.

1.2. Architectures for Distributed Simulation

The client-server [4] and the peer-to-peer [5] approaches are two widely-used architectures for distributed simulation. As its name implies, the client-server approach involves executing the distributed simulation on one or more server computers (which may be several computers connected by a local area network) to which clients (e.g., users) can “log in” from remote sites. The bulk of the simulation computation is executed on the server machines. This approach is typically used in distributed simulations used for multiplayer gaming. Centralized management of the simulation computation greatly simplifies management of the distributed simulation system, and facilitates monitoring of the system, e.g., to detect cheating. On the other hand, peer-to-peer systems have no such servers, and the simulation is distributed across many machines, perhaps interconnected by a wide area network. The peer-to-peer approach is often used in distributed simulations used for defense.

1.3. Developmental Trend and Significance

Here, increasingly important motivation for distributed simulation concerns the desire to integrate several different simulators into a single simulation environment. One example where this paradigm is frequently used is in military training. Tank simulators, flight simulators, computer generated forces, and a variety of other models may be used to create a distributed virtual environment into which personnel are embedded to train for hypothetical scenarios and situations.

Another emerging area of increasing importance is infrastructure simulations where simulators of different subsystems in a modern society are combined to explore dependencies among subsystems. In both these domains (military and infrastructure simulations) it is far more economical to link exiting simulators to create distributed simulation environments than to create new models within the context of a single tool or piece of software.

2. TIME MANAGEMENT

Time management is concerned with ensuring that the execution of the distributed simulation is properly synchronized. This is particularly important for simulations used for analysis Time management not only ensures that events are processed in a correct order, but also helps to ensure that repeated executions of a simulation with the same inputs produce exactly the same results[6].

Time management algorithms usually assume the simulation consists of a collection of logical processes LPs) that communicate by exchanging time-stamp messages or events. Each federate can be viewed as a single LP. The goal of the synchronization mechanism is to ensure that each LP processes events in time-stamp order. This requirement is referred to as the local causality constraint (LCC).

Synchronization [7] means each LP maintains local state information corresponding to the entities it is simulating and a list of time stamped events that have been scheduled for this LP, but have not yet been processed. This pending event list includes local events that the LP has scheduled for itself as well as events that have been scheduled for this LP by other LPs. The main processing loop of the LP repeatedly removes the smallest time stamped event from the pending event list and processes it. Thus, the computation performed by an LP can be viewed as a sequence of event computations. Processing an event means zero or more state variables within the LP may be modified, and the LP may schedule additional events for itself or other LPs. Each LP maintains a simulation time clock that indicates the time stamp of the most recent event processed by the LP. Any event scheduled by an LP must have a time stamp at least as large as the LP's simulation time clock when the event was scheduled. This property also helps to ensure that the execution of the simulation is repeatable; one need only ensure the computation associated with each event is repeatable.

Time management algorithms can be classified as being either conservative or optimistic [8]. Briefly, conservative algorithms take precautions to avoid the possibility of processing events out of time stamp order, i.e., the execution mechanism avoids synchronization errors. On the other hand, optimistic algorithms use a detection and recovery approach. Events are allowed to be processed out of time stamp order, however, a separate mechanism is provided to recover from such errors.

2.1. Conservative Time Management

The principal task of any conservative protocol is to determine when it is "safe" to process an event. An event is said to be safe when one can guarantee no event containing a smaller time stamp will be later received by this LP. Conservative approaches do not allow an LP to process an event until it has been guaranteed to be safe. [12]

At the heart of most conservative synchronization algorithms is the computation for each LP of a Lower Bound the Time Stamp (LBTS)[9] of future messages that may later be received by that LP. This allows the mechanism to determine which events are safe to process. For example, the synchronization algorithm has determined that the LBTS value for an LP is 8, then all events with time stamp less than 8 are safe, and may be processed. Conversely, all events with time stamp larger than 8 cannot be safely processed. Whether or not events with time stamp equal to 8 can be safely processed depends on specifics of algorithm, and the rules concerning the order that events with the same time stamp (called simultaneous events) are processed. Processing of simultaneous events complex subject matter that is beyond the scope of the current discussion .The discussion here assumes that each event has a unique time stamp.

2.2. Optimistic Time Management

Optimistic approaches offer two important advantages over conservative techniques. First, they can exploit greater degrees of parallelism. If two events *might* affect each other, but the computations are such that they actually don't, optimistic mechanisms can process the events concurrently, while conservative methods must orderly execute. Second, conservative always mechanically rely on application specific information in order to determine which events are safe to process. While optimistic mechanisms can execute more efficiently if they exploit such information, they are less reliant on such information for correct execution. This allows the synchronization mechanism to be more transparent to the application program than conservative approaches, simplifying software development. On the other hand, optimistic methods may require more overhead computations than conservative approaches, leading to certain performance degradations.

The Time Warp mechanism [8] is well known optimistic method. When an LP receives an event with timestamp smaller than one or more events has already processed, it rolls back and reprocesses events in timestamp order. Rolling back an event involves restoring the state of the LP to that which existed prior processing the event.

3. HIGH LEVEL ARCHITECTURE

3.1. Concepts and Comprehension

The High Level Architecture (HLA)[9][10] developed by the Department of Defense in the United States is first described to provide a concrete example of a contemporary approach to integrate, or federate, separate, autonomous simulators into a single, distributed simulation system. It aims to promote reuse and interoperation of simulations. The intent of the HLA is to provide a structure that supports reuse of different simulations, ultimately reducing the cost and time required to create a synthetic environment for a new purpose. Though this concept of HLA first has been mentioned in the context of defense application, the HLA was intended to have applicability across a broad range of simulation application areas, including education and training, analysis, engineering and even entertainment, at a variety of levels of resolution. These widely diverse application areas indicate the variety of requirements that were considered in the development and evolution of the HLA.

3.2. Support Software RTI

The HLA does not prescribe a specific implementation, nor does it mandate the use of any particular set of software or programming language. It was assumed that new technological advances become available, and different implementations would be possible within the framework of the HLA [10] [11]. A federation of HLA consists of a collection of interacting simulations, termed federates. It may be computer simulation, a manned simulator, a supporting sensor, or an interface a live player. All object representation state within the federates. The HLA require all federates incorporate specified capabilities to allow the objects in the simulation to interact with else objects in other simulations in virtue of exchange of data [10].

Data exchange and a variety of other services are realized by software called the Runtime Infrastructure (RTI)[11]. The RTI is, in effect, a distributed operating system for the federation. The RTI provides a general set of services that support the simulations in carrying out these federate-to-federate interactions and federation management support functions. All interactions among the federates go through the RTI.

3.3. Time Management in the HLA

It was envision that some federates may be executing on a parallel processor, and may be using conservative or optimistic synchronization mechanisms within their federate. The HLA time management services were designed to accommodate this wide variety of applications.

The RTI provides an optional time management service to coordinate the exchange of events between federates. Events can be associated with a point in time and the RTI can assist in ensuring causal behavior. It is also possible for one or more federates in a federation to fully ignore time. By default, the RTI does not attempt to coordinate time between federates. In addition, the HLA not only supports a variety of time management policies, but also facilitates interoperability between federates with different policies. Even if the optional time management services are ignored, it pays to understand available time management schemes.

3.3.1. Regulating and Constrained

In a federation, time always moves forward. However, the perception of the current time may differ among participating federates. Time management is concerned with the mechanisms for controlling the advancement of each federate along the federation time axis. In general, time advances must be coordinated with object management services so that information is delivered to each federate in a causally correct and ordered fashion.

In some situations, it is appropriate to constrain the progress of one federate based on the progress of another. In fact, any federate may be designated a regulating federate. Regulating federates regulate the progress in time of federates that are designated as constrained. In general, a federate may be "regulating," "constrained," "regulating and constrained," or "neither regulating nor constrained." By default, federates are neither regulating nor constrained.

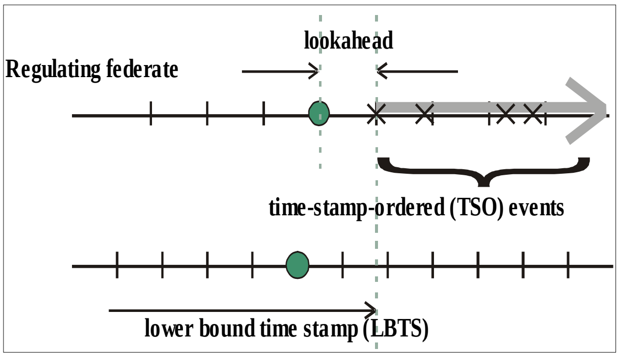

lookahead
Regulating federate
time-stamp-ordered (TSO) events
lower bound time stamp (LBTS)

Figure 1 Two-Axis Diagram

Figure 1, known as the "two-axis diagram," introduces the definitions of "regulating," "look-ahead," "TSO event," "constrained," and "lower bound time stamp (LBTS)." The little dot in time-axis indicates the current time apperceived by federate; 'x' symbol shows the location of time-stamp-ordered events in time-axis. There are four time-stamp-ordered events of regulating federate. Among them, the least timestamp is LBTS (lower bound time stamp).Here we presume only two federate in this federation.

A federate that declares itself to be "regulating" is capable of generating time-stamp-ordered (TSO) events. TSO events would occur at a specific point in time. Federates that are not regulating can generate events, however these events without time-stamp, namely RO (Receive-Ordered) events. A regulating federate coordinates time advances with the local RTI component (LRC). The regulating federate perceives the current time to be "*T current.*" Federates can dynamically alter their status becoming regulating or non-regulating.

A federate that declares itself to be "constrained" is capable of receiving TSO events. Federates that are not constrained still learn of TSO events, but absent the time-stamp information.

3.3.2. Look-ahead

Each regulating federate establishes a "look-ahead" value. The regulating federate promises that any TSO events it generates will occur if satisfy below Eq (1)

$$T_{tso} \geq T_{\mathrm{current}} + T_{lookahead} \quad \ldots\ldots\ldots\ldots\ldots \text{(1)}$$

$T_{current}$ **:** The current time

$T_{lookahead}$ **:** Represents a contract between the regulating federate and the federation.

It establishes the earliest possible TSO event the federate can generate relative to the current time, Regulating federates must specify a look-ahead value at the time they become regulating. Facilities exist to alter the look-ahead value dynamically.

3.3.3. Lower Bound Time Stamp (LBTS)

Constrained federates have an associated LBTS. The LBTS specifies the time of the earliest possible time-stamp-ordered event the federate can receive. The LBTS is determined by looking at the earliest possible message that might be generated by all other regulating federates. It changes as the regulating federates advance in time. A constrained federate cannot advance beyond its LBTS, because the RTI can only guarantee there will be no more packets received prior to the LBTS.

3.4. Advancing Time

In Figure 2, five of six federates have joined an established federation. One of the federate has not shown up yet – it is said to be *late arriving*. The small, solid circles represent the federation time as perceived by each federate. It is extremely important to understand that there is no universal "federation time" (at any given point each federate could have different “current times.” Each federate is free to increment time independently. Some federates will apply the same time increment repeatedly. Other federates may jump through time based on the next available TSO event or some other criteria.

The thick, shaded regions in the diagram represent the look-ahead values specified by regulating federates. Federate 1's look-ahead is twice its time step interval. Federate 3 and 5 have look-ahead values that appear to be one time interval ahead. The look-ahead values need not be related to a federates time interval (as we will see when Federate 2 arrives).

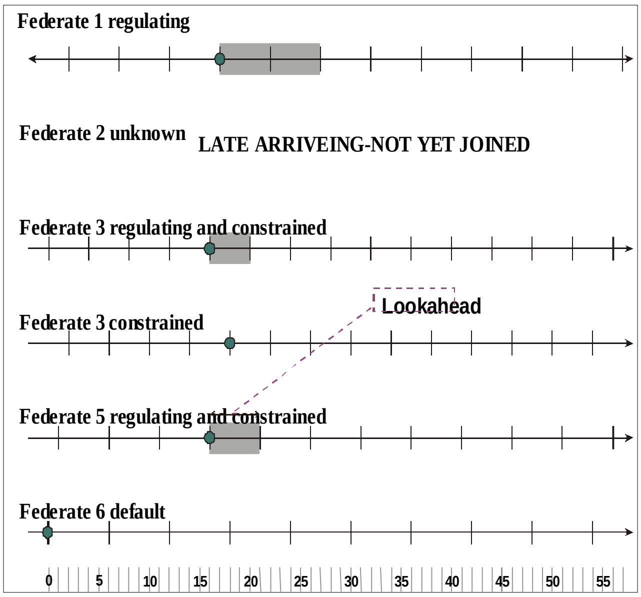


Figure 2 Six-axis Diagram – Late Arrival

Clearly, each federate in this federation has a unique perspective on the current time.

Federate 1 t = 17 seconds
Federate 2 not applicable
Federate 3 t = 16 seconds
Federate 4 t = 18 seconds
Federate 5 t = 16 seconds
Federate 6 t = 0 seconds

In general, unconstrained federates are free to progress through time. An unconstrained federate has no requirement to request time advance grants through the RTI. For example, Federate 1 and Federate 6 can advance in time as fast as they want (or at least as fast as their simulation model can run). Should these unconstrained federates request permission to advance in time, their LRC realizes that they are unconstrained and grants permission to advance as a matter of course.

3.4.1. LBTS Constraint Advancing Time

Constrained federates cannot proceed beyond their current LBTS. The LBTS for a given federate is determined by calculating the earliest possible message a federate might receive from other regulating federates. Enforcing the LBTS constraint requires coordination between federate LRCs. As regulating federates advance, the LBTS of constrained federates increases. Figure 3 illustrates the LBTS for constrained federates [10] [11].

The vertical dashed lines in Figure 3 represent the earliest possible TSO message that can be produced by each of the regulating federates – given their current time and their promised look-ahead values. Below each constrained federate, a horizontal line is extended from "t = 0" to the federate's LBTS. In Figure 3, it is clear that the current time as perceived by each of the constrained federates is within their respective LBTS windows. Therefore, the "combination of perceived times" for each federate shown is legitimate.

Constrained federates are free to advance in time to their LBTS, but no further. In Figure 3, Federate 3 could increment to the next "tick mark" since the resulting time would be within its LBTS. However, Federate 4 and Federate 5 cannot proceed to their next "tick mark," as each would have to move beyond its respective LBTS values.

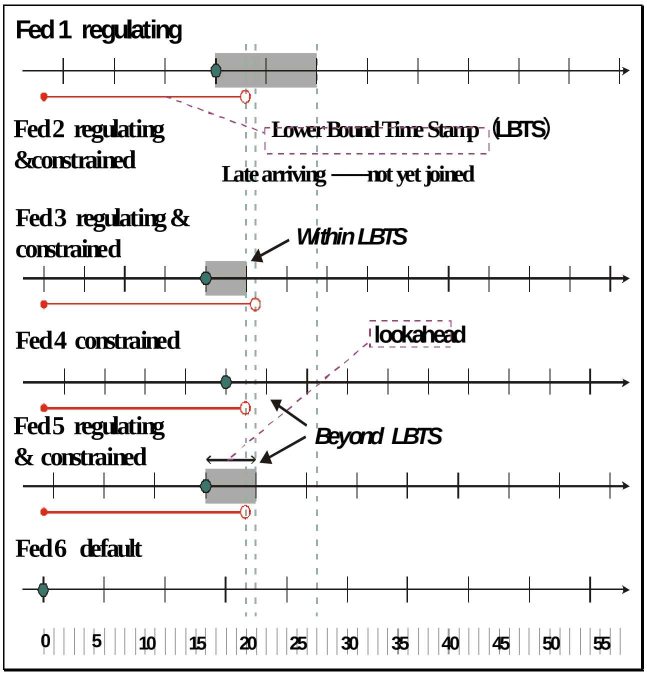
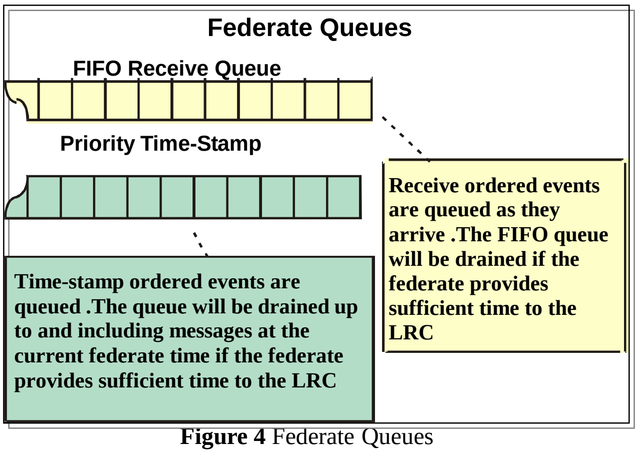

Fed1 regulating
Fed2 regulating &constrained
Lower Bound Time Stamp (LBTS)
Late arriving —— not yet joined
Fed3 regulating & constrained
Within LBTS
Fed4 constrained
lookahead
Fed5 regulating & constrained
Beyond LBTS
Fed6 default
0
5
10
15
20
25
30
35
40
45
50
55

Figure 3 LBTS for Constrained Federates

3.4.2. Federate Queues

As illustrated in Figure 4 each LRC maintains two queues. Events that meet the TSO criteria are placed in the time-stamp queue. The time-stamp queue orders incoming events based on the time stamp. Events that fail to meet the TSO criteria are placed in the receive queue in the order in which they arrive. Information in the receive-order queue is immediately available to the federate. The federate has access to all events in the TSO queue with time stamps less than or equal to the federate's perceived time [13].



Receive ordered events are queued as they arrive .The FIFO queue will be drained if the federate provides sufficient time to the LRC

Time-stamp ordered events are queued .The queue will be drained up to and including messages at the current federate time if the federate provides sufficient time to the LRC

Figure 4 Federate Queues

4. CONCLUSIONS

Beginning with research and development efforts in the 1970’s, research in distributed simulation systems has matured over the years. For many problems such as simulation of large-scale networks such as the Internet, performance remains a principal motivating objective, however, much interest in this technology today stems from the promises of cost savings resulting from model reuse. Standards such as IEEE 1516 for the High Level Architecture demonstrate the widespread interest in use of distributed simulation technology for this purpose.

What is the future for the technology? It is interesting speculate. One potential path is to focus on applications. High performance computing remains a market that targets a handful of important, computation intensive applications, for broader impacts in society, where distributed simulation technology has seen the most widespread deployment. Another view is observe that software is often driven by advances in hardware technology, and look to emerging computing platforms to define the direction the technology will turn. this light, ubiquitous computing stands out as an emerging area where distributed simulation may be headed. For example, execution of distributed simulations on handheld computers necessitates examination of power consumption because battery life is a major constraint in such systems. Grid computing is still another emerging approach where distributed simulations may emerge and have an impact.

5. REFERENCES

[1]. Huang Kedi. et al. write and compile 《The Technology of Simulation System》, Publishing Company of National University of Defense Technology ,1998.in Chinese

[2]. Beraldi, R. and L. Nigro (2000). Exploiting Temporal Uncertainty in Time Warp Simulations. Proceedings of the 4th Workshop on Distributed Simulation and Real-Time Applications: 39-46.

[3]. Richard M. Fujimoto. Distributed Simulation Systems. Proceedings of the 2003 Winter Simulation Conference:124-134

[4]. Bononi, L., G. D’Angelo, et al. (2003). HLA-Based Adaptive Distributed Simulation of Wireless Mobile Systems. *Proceedings of the 17th Workshop on Parallel and Distributed Simulation*: 40-49.

[5]. Bodoh, D. J. and F. Wieland (2003). Performance Experiments with the High Level Architecture and the Total Airport and Airspace Model (TAAM). *Proceedings of the 17th Workshop on Parallel and Distributed Simulation*: 31-39.

[6]. Carothers, C. D., K. Perumalla, et al. (1999). “Efficient Optimistic Parallel simulation Using Reverse Computation.” ACM Transactions on Modeling and Computer Simulation 9(3).

[7]. Chen, G. and B. K. Szymanski (2002). Lookback: A New Way of Exploiting Parallelism in Discrete Event Simulation. Proceedings of the 16th Workshop on Parallel and istributed Simulation: 153-162.

[8]. Ferscha, A. (1995). Probabilistic Adaptive Direct Optimism Control in Time Warp. Proceedings of the 9th Workshop on Parallel and Distributed Simulation: 120-129.

[9]. High Level Architecture Run-Time Infrastructure Programmer's guide 1.3 version 6, DMSO, Mar. 12, 1999

[10]. IEEE P1516.1 Standard [for] Modeling and Simulation (M&S) High Level Architecture (HLA) – Federate Interface Specification，Draft 1，DMSO，April 20, 1998

[11]. IEEE Std 1516-2000 (2000). IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-- Framework and Rules. New York, NY, Institute of Electrical and Electronics Engineers, Inc.

[12]. Sokol, L. M. and B. K. Stucky (1990). MTW: Experimental Results for a Constrained Optimistic Scheduling Paradigm. Proceedings of the SCS Multiconference on Distributed Simulation. 22: 169-173.

[13]. DMSO, Federation Development and Execution Process(FEDEP) Model, Version 1.4, June 9, 1999

Intelligent Grading Based on Image Recognition

Zheng Guang Kong Meijing Fu Dong Zhang Xiaoming Wang Jianxia
College of Information Science & Engineering, Hebei University of Science and Technology
Shijiazhuang Hebei 050054, China
Email: Zheng_g@hebust.edu.cn **Tel.:** 0311－8613336

ABSTRACT

There are a lot of methods to realize Machine Grading. In this article we introduce an Intelligent Grading method based on Image Recognition and ADO technique, and illustrate the key technology, such as calculating gradient, adjusting image, area division and location, extracting the character, manage repeat or omit the answer, etc. This software has obtained patent.

Keywords: Image Recognition, Machine Grading, ADO, Character Extracting

1. INTRODUCTION

In recent years many universities are enlarging the number of students. The workload of teachers is increased. Currently there are many impersonal test questions in test paper, especially in English examination, thus grading is difficult. This system's design is based on this fact.

There are a lot of methods to realize Machine Grading. For reasons of price and technique, those productions are expensive and are not extensive application. This system adopts a common computer and image capture equipment (scanner, digital camera) to realize Machine Grading. This is cheap and convenient way. It can be accepted by a great number of schools in china.

2. KEY TECHNOLOGY

Computer has good image process ability. Combined with other techniques Image Recognition can solve many factual problems. This system obtains answer card image by capture equipment. It gets useful image characters by analyzing and processing the image. These image characters are corresponding with the information of student (class, answer, etc). We can save this information to database and manage the information by database technology. Finally we get an integrated Graded system. The key technology in this system is as follows.

2.1 Extracting the Image Information

There are a lot of methods to get the image, for example we can use scanner, digital camera, etc. Think of the price and operability, we adopt scanner in this system. We scan answer card and switch the result to gray image directly by scanner. In this way we can reduce the calculation times and eliminate interferential ingredient. In order to prepare for extracting the image characters, we can preprocess (for example image adjusting) if the scanner supports.

2.2 Image Preprocessing

2.2.1 Calculating Grad

When we scan the answer card, the image may incline because of various reasons. If the gradient is not big enough, we can extract the information from answer card correctly. Conversely we can not get useful information. We can calculate gradient by two vertical lines. Show as the Fig1.

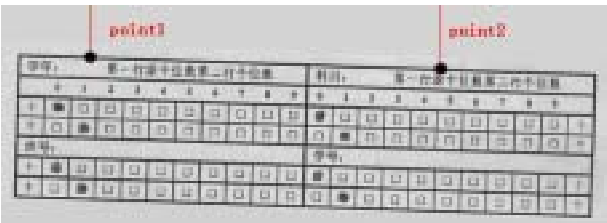


Fig. 1 calculate gradient

If the height of point1 and point2 is the same, the image is upright. We need not adjust the image. Conversely we must calculate the gradient in order to adjust the image and extract characters. The formula is as follows.

pi ＝3.14159265358979
y1: point2
y2: point1
D1: Distance of two points
gradient = arctan （(y1-y2）/D1）* 180 / pi

We get two experiential values "n1"=1.6 and "n2"=2 by large numbers of test. When gradient is smaller than "n1", we can get correct information from image directly. So we need not adjust the image. When gradient is between "n1" and "n2", we must adjust the image in order to get information. When gradient is bigger than "n2", we can not get useful information in any case. So we should stop image processing and modify image or scan the card again.

2.2.2 Adjusting Image

Image adjusting is mainly to rotate the image in this system in order to get an upright image. The circumrotation arithmetic in this system is as follows.

1) Loading the image, working out the center of image;
2) Division the image based on that point into four parts (if the image is too big you can division the image again)
3) Working out the gradient, then rotating every part base on the angle
4) Loading the result.

2.2.3 Area Division and Location

The useful information in the answer card locates in specific area. The margin and explanation part is not useful for us. That is to say we must divide and locate the information area. After we locate the area, we can extract the image character. Show as Fig2.

Student number area, test paper type area and answer area are very important for extracting the image characters. If we locate accurately we can continue the other works.

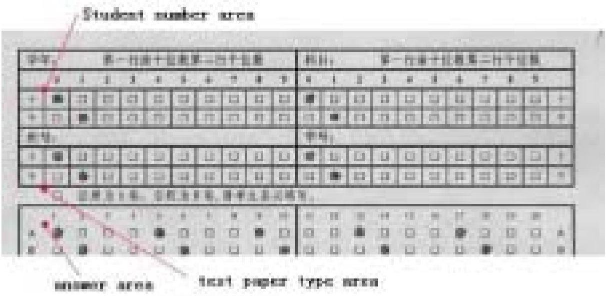


Fig. 2 area location

2.2.4 Extracting the Character

Extracting the image character in this system is based on pixels value of the image. We judge whether student black one area based on the pixels average of specific area. The formula is as follows.

Select: the length of average

Pcolor[i]: the pixels value of one point in a specific area

n: point number of specific area

Select = Length (Int(∑Pcolor[i]/n))

"Length" is getting the length of the average.

The specific area of most length average is selected in a group of option. Different option has different meaning. Then we can get all useful information in the card, such as subject, test type, student number, answer, etc.

2.2.5 Manage Repeat or Omit the Answer

Students may repeat or omit some part when they fill the answer card. It can be sorted in two types to process.

1) When the student information is repeated or missing, the information is wrong. We can not query and save this information in database. In this case, this system should stop image processing.

2) When the answer is repeated or missing, the answer is wrong. We sign the option with "M" when the answer is repeated. We sign the option with "N" when the answer is missing. When we statistic the answer, "M" and "N" are signed 0 score.

2.3 Data Processing

This system adopts popular ADO technology and convenient Access database. It provides input, modification, query, statistic, etc. Its function is as follows

1) Student information input, modification, query.
2) Standard answer input, modification.
3) Student answer query, scan.
4) Examination grade statistical.
5) Student grade query, print.

We must store the basal student information, standard answer, every student's answer, the statistical result of each class, the statistical result of each test type in this system. So we design the database as follows:

Infor: store student information;
Standard: store standard answer;
Student: store student's answer;
Fenxi: store class grade statistical information
Tongji: store test paper statistical information

3. PROGRAM FLOW

The core of this system is grade process of answer card. There are a lot of technologies in this part. Show as Fig.3

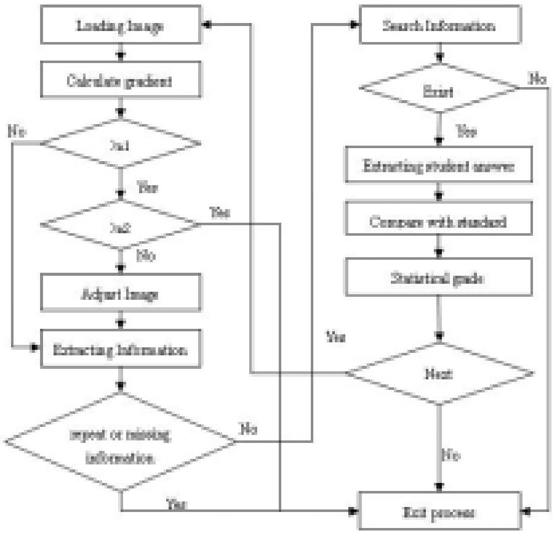


Fig. 3 Flow chart of program

4. EXPERIMENT RESOLUTION

In this test we use high-powered computer. The CPU is P4 1.8G. The memory is 265 MB. The operating system is windows XP. We adopt a high-speed scanner (8 piece of paper per minute). In this test we process four classes, 132 students. Each answer card has 100 options.

The test result is as follow：

1) Scan speed: 480 piece of paper per hour (8*60)
2) Answer card process speed: 1.1 second
3) If fill the answer card is exact, the accuracy is 99.6%
4) The accuracy is 98% if answer card need adjust.

5. CONCLUSIONS

This system makes answer card digital by common computer and image capture equipment. It achieves information extracting by Image Recognition and Character Extracting. It manages database by ADO technology. This system provides a cheap and convenient tool for teachers. This system has widely application.

6. REFERENCES

[1] Castleman K R. Digital Image Processing. Beijing: Publishing House of Electronics Industry, 1998

[2] Zheng Nan-Ning. Computer Visual and Pattern Recognition. Beijing: National Defence Industry Press, 1998

[3] Ma W Y, Zhang H J. Content-based image indexing and retrieval. In: Fuhrt B ed. Handbook of Multimedia Computing. Boca Raton, Florida: CRC Press, 1999. 227-254

[4] Carson C.,Thomas M.,Belongie S.,et al.Blobworld: A

System for Region-based Image Indexing and Retrieval[A].Third Int.Conf.on Visual Information Systems[C],June 1999.

Zheng Guang (1972-), male, is a lecturer of Hebei University of Science and Technology. He graduated from Hebei University of Science and Technology in 1996. He has published two books, over 9 Journal papers. His research interests are in Image Recognition technology.

Fu Dong (1976-), male, is a lecturer of Hebei University of Science and Technology. He graduated from Hebei University of Science and Technology in 1999. He has published two books, over 3 Journal papers. His research interests are in Image Recognition technology.

The Recognition and Decomposition of Mixed Pixels in the Remotely Sensed Images Based on Gray System Theory[*]

Gui Yufeng, Tao Jianfeng
College of Science, Wuhan University of Technology, Wuhan, P.R. China 430063
Email:guiyufeng@sina.com

ABSTRACT

This paper analyses the existing method of recognition and decomposition of mixed pixels in the remotely sensed images, and presents a new method of recognition and decomposition of mixed pixels based on gray system theory. The experiments demonstrate the method and give satisfactory results.

Keywords Mixed pixels, Recognition , Decomposition, Gray system theory

1. INTRODUCTION

With the development of social information, the demand for the application of remote sensing images increases greatly. Land-Cover/land-Use has become crucial basis work to carry out the prediction to the dynamical change of land-use, prevention to natural disaster, environment protection, land management and planning. Automatic recognition of the multi spectral images has studied for 20 years and has made great progress, but the accuracy of recognition is not higher. In remote sensing satellite images, the size of pixels, in general, may include more than one type of terrain cover. When these sensors observe the earth, the measured radiance is the integration of the radiance of all the objects that are contained in the pixel, implying the existence of the so-called mixture problems.

2. ANALYSIS OF THE EXISTING METHODS

Some main approaches have been used in the technical literature to solve the mixture problem:

2.1 Linear mixture model classification

By adopting the linear mixture model, the pixel value in any spectral band is given by the linear combination of the spectral response of each component within the pixel. The mixture model was solved by applying a least-square estimator (Shimabukuro and Smith 1991),with an unconstrained solution (Schanzer 1993)

The advantage of this method is that it has a simple model, but when endmember can not be select accurately, there is bigger errors.

2.2 Nonlinear mixture model classification

For overcoming the disadvantage of linear mixture model classification , some scholars put forward the nonlinear mixture model classification .The results are often better, but the stability of the results are not good, and the process of calculation is complex.

2.3 Experience coefficient method

The number of mixed pixels is proportional in some region. The coefficient table can be built according experience in past years. When we add up area of every type terrain cover, we add or subtract the area in proportion. It is easy to carry out, but it results in bigger error.

Above methods have their advantages, but have not a statement concerning the selection of endmembers feasible, and don’t take full considerable of the information of neighbour pixels.

3. RECOGNITION AND DECOMPOSITION METHOD BASED ON GRAY SYSTEM THEORY

Gray system theory, founded by Professor Deng Julong, can handle undetermined problem .It is effective when the sample datum can not satisfy some distribution.

Based on gray system theory this paper gives a new method. As we know, a gray number in gray system theory is an interval and its scope is known approximately, but its accuracy value is unknown. We denote it by $\otimes$. In remote sensing satellite images, the Digital Number of end member is also a set of number whose value can’t be known accurately but scope can be known approximately, so it is a gray number.

If A is interval, $\alpha \in$ A, and the scope of gray number $\otimes$ is A, than α is called a white value of $\otimes$. So, we have symbols as follow:

$\otimes$ is a gray number in a general.

$\otimes(\alpha)$ is a gray number whose a white value is α.

The spectral response of the mixed pixel is related to the spectral response of end member of neighbour pixels. According correlative space distribution of pixels and the gray end member, mixed pixels can be recognized and decomposed .It can used fully the space information , increase the precision of classification and recognition.

The process of recognition and decomposition method is:

(1) import image
(2) select training sample by vision
(3) decide end members using gray correlation degree.

[*] This work is supported by National Lab. Foundation of P.R.C.

If E is a distance space, we definite the mapping on E
d: $E^2 \rightarrow R = (-\infty, +\infty)$, satisfying $\forall x, y, z \in E$,
1^0 d(x,y) ≥ 0, and d(x,y)=0 $\Leftrightarrow$ x=y;
2^0 d(x,y)=d(y,x);
3^0 d(x,z) $\leq$ d(x,y)+d(y,z)
$1^0 \sim 3^0$ be called the three axiom of distance space.

Given $x_0, x_i \in E^n$, i=1,2,……,m,
x_0=($x_0(1)$, $x_0(2)$, $x_0(3)$, ……, $x_0(n)$)
x_i =($x_i(1)$, $x_i(2)$, $x_i(3)$, ……, $x_i(n)$)
where $x_i(k) \in E$, i=0,1,2,……,m; k=1,2,……,n.
we definite:

$$\Delta_{0i}(k) = d(x_0(k), x_i(k)), \qquad \Delta_{\min} = \min_i \min_k \Delta_{0i}(k) \quad ,$$

$$\Delta_{\max} = \max_i \max_k \Delta_{0i}(k)$$

the correlation coefficient between basic sequence and correlation sequence is:

$$r(x_0(k), x_i(k)) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_{0i}(k) + \rho\Delta_{\max}}$$

$\rho \in (0,1)$

The general correlation degree is :

$$r(x_0,x_i)=\sum_{k=1}^{n} w_k r(x_0(k), x_i(k))$$

E is a linear norm space , if $\forall x \in E$, $\|x\|$ defined one real number satisfy ：$\forall x, y \in E$

1^0 $\|x\| \geq 0$, and $\|x\| = 0 \Leftrightarrow x = 0$;
2^0 $\forall \lambda \in R, \|\lambda x\| = |\lambda| \|x\|$
3^0 $\|x + y\| \leq \|x\| + \|y\|$.
where $x_0, x_i \in E^n$, i=1,2,……,m,， we definite

$$\Delta_{0i}(k) = \|x_0(k) - x_i(k)\|$$

we can definite the correlation coefficient and the correlation degree on linear norm space.

The correlation degree is given according to “normality”, “symmetry”, “entirety”.

(4) build decomposition model

Let the threshold is T. If t-th endmember with Digital Number d (t) exists for the object with Digital Number d , such that

$$d(t) - T < d < d(t) + T$$

then we consider the object with Digital Number d as an endmember .We noted it by $\otimes$ (d).

If there are k types endmembers with Digital Number $\otimes$ (d(1)), $\otimes$ (d(2)) ,…, $\otimes$ (d(k)), we select a list of white value of Digital Number of endmembers. For convenience, these list of white value are arranged from big to small and are still noted by $\otimes$ (d(1)), $\otimes$ (d(2)) ,…, $\otimes$ (d(k)), Digital Number of (i,j) mixed pixels is noted by D(i,j)

If two pixels is symmetry to some pixel, and close to the pixel, then they are called a neighbour couple of pixels. For example, a neighbour couple of pixels of pixel (i,j) is:

$$(S1, S2) \in \{((i, j-1), (i, j+1)), ((i-1, j), (i+1, j)),$$
$$((i-1, j-1), (i+1, j+1)), ((i-1,. j+1), (i+1, j-1))\}$$

Identify the types of a neighbouring couple of pixels by near principle :

$$\min \ |D(S1) - \otimes(d(t)| = |D(S1) - \otimes(d(m))|$$
$$\min \ |D(S2) - \otimes(d(t)| = |D(S2) - \otimes(d(n))|$$

If one of following conditions is satisfied:

(a) $D(i, j) > \otimes(d(m))$ and $D(i, j) > \otimes(d(n))$

(b) $D(i, j) < \otimes(d(m))$ and $D(i, j) < \otimes(d(n))$

then the neighbour couple of pixels can not produce the mixed pixel and the neighbour couple of pixels should be removed.

Find a satisfactory solution of

$$\begin{cases} k(m,t1) \otimes (d(m)) + k(n,t2) \otimes (d(n)) = D(i, j) \\ k(m,t1) + k(n,t2) = 1. \end{cases} \qquad (1)$$

where $k(m,t1)$ is the ratio between area of m-th end member and area of mixed pixel which t1-th time occurs and $k(n,t2)$ is the ratio between area of n-th end member and area of mixed pixel which t2-th time occurs .

The solution of equations (1) is

$$\begin{cases} k(m,t1) = \dfrac{\otimes((d(n)) - D(i, j)}{\otimes(d(n)) - \otimes(d(m))} \\ k(n,t2) = \dfrac{\otimes(d(m)) - D(i, j)}{\otimes(d(m)) - \otimes(d(n))} \end{cases}$$

Find the mean of the ratios of the same type of endmembers in mixed pixel , we have:

$$\begin{cases} \overline{k(m)} = \dfrac{\sum_{t=1}^{t2} k(m,t)}{t1} \\ \overline{k(n)} = \dfrac{\sum_{t=1}^{t2} k(n,t)}{t2} \end{cases}$$

4. EXPERIMENT AND ANALYSIS

A experiment have been carried out in our paper.

4.1 TM image experiment

Given a block of TM image, degrade the image from 30m resolution to 60m resolution.

① Classify the original TM image, obtain the area ratios of types of objects. Assume that mixed pixels in this image can be ignored

② Classify the degraded TM image , using gray correlation pixel decomposion method, obtain the area ratios of types of objects

③ Classify the degraded TM image, using traditional supervised classification method, obtain the area ratios of types of objects.

Table 1 shows the area ratios of types of objects using 3 methods.

Table 1. The area ratios of types of objects using 3 methods.

Ratio of area (%)	Urban	Forest	Water	Roads	Croplands	vegetable
Supervised classification of degraded image	21.041	5.932	3.758	6.894	41.116	21.251
Pixel decomposition classification of degraded image	21.440	8.895	4.160	7.821	41.011	17.077
original image classification	20.042	9.483	4.683	10.381	39.252	16.159

Table 2 Precision table of two classification methods

Precisions (%)	Urban	Forest	Water	Roads	Croplands	vegetable	mean
Supervised classification of degraded image	94.98	62.55	80.23	66.42	95.25	68.49	77.99
Pixel decomposition classification of degraded image	95.80	93.77	90.84	80.37	95.52	94.33	91.77

4.2 Analysis of the results

Table 2 shows the comparism of precision of two classification methods.

By decomposing mixed pixels, mean precision increase form 77.99% to 91.77%, and precision of every type of objects increase obviously

5. CONCLUSION

The experiment and analysis show that the new method of recognition and decomposion is effective.

The method has feature as follow:

① The method is simple and practice.
The method is based on Gray System Theory, and regards the Digital Number of the end member as “gray number”, avoid failure for the unfit selection of mixed pixels.

② Error samples have be deleted using gray correlation degree.

③ The method use fully the information of the space .
Because of using the neighbour pixels of mixed pixels, the results is credible。

6. REFERENCES

[1] .Suenjiabing, Shuning,Guanzequn. Principle ,Method and Application of Remote Sensing. Mapping & Survey Press .1996

[2] G.M.Floody. Oridinal-level classification of Subpixel Tropical Forest Cover. Photogrammetric Engineering and Remote Sensing,1994,60(1):61-65.

[3] M.A.Cochrange and C.M.Souza. Linear Mixture Model Classification of Burnel Forests In Eastern Amazon.Int.J.Romate Sensing,1998,19(17)

[4] Deng Julong,Gray System Basic Method. Huazhong Science & Technology University Press. 1987

Gui Yufeng is an associate Professor and the director of statistics department in the school of science, Wuhan University of Technology. He graduated from Wuhan University in 2003 and received his Ph.D; He has published one book, over 20 Journal papers. His research interests are in mathematics, distributed parallel processing, digital image processing.

A Distributed Tracking System for Indoor Augmented Reality Applications

Jian Mao , Yaolin Gu
School of Information engineering, Southern Yangtze University
Wuxi, 214036, P.R. China
Email: mj5605@hotmail.com **Tel:** 0510-5707351

ABSTRACT

The indoor distributed tracking system proposed in this paper is a client-server architecture with three basis subsystems: the site information server, mobile units, and the network infrastructure. The tracking process is divided into four procedures: camera pose computing, 2D features tracking, 2D-3D correspondence and natural feature calibration. To achieve real-time behavior required for tracking applications, dynamic interpretation tree is used to organize data in a multimedia object database. The computations of 2D image registration for pose estimation are done on basis of the Fast Fourier Transform because of its robustness and fast calculation. To track camera pose and dynamically estimate the 3D positions of natural feature, we use an auto-calibration approach based on an iterative Extended Kalman Filter. The result shows that our system can successfully provide indoors mobile users with 3D positions from 2D image points and information of augmented reality.

Keywords: distributed tracking system, Dynamic Interpretation Tree, iterative Extended Kalman Filter, augmented reality.

1. INTRODUCTION

Augmented reality (AR) applications use computer-generated virtual scenes to enhance (or augment) the actual scene viewed by the user with additional information. AR systems have been used in the medical [1], commercial [2], and archaeology fields [3], as well as in engineering design. They enrich human perception and facilitate the understanding of complex 3D scenarios.

Today outdoor AR applications have been successfully achieved by differential Global Positioning System (DGPS) and corresponding tracking system, for example Archeoguide used to guide for archeological sites [3]. Simultaneously consistent positioning of indoor mutual sites, however, remains a complex issue.

For AR to work properly in a large building, a lot of technical problems must be addressed in order to produce a distributed AR system.

AR applications require tracking system with high accuracy, low latency, low jitter and good mobility. The system must confirm the user's head pose – the position and orientation of the user's head in the scene coordinate system. Further problems are due to working in a large environment:

1) Marker-based approaches are not available to AR scenarios requiring users to operate in a very large environment, e.g., a big industrial plant. Markerless approaches which track naturally occurring landmarks in the scene are required.
3) Lighting conditions change more rapidly in a building than outdoors, and outdoor tracking system can exploit global rather than local image properties by DGPS. Different images of the same object under different lighting conditions bring about difficulties of precise calibration and registration. Thus, it is apparent to choose a proper camera CCD that can adjust exposure, brightness and contrast to help overcome the difficulties.
4) An augmented reality system should interactively provide users requested information. Since users are working in respective actual 3D environments, the system should receive information requests through conventional means, either by tracking the motions of users and interpreting their gestures, or through a speech recognition system.
5) Sites' dimensions and remoteness from wired communication and power networks call for a wireless solution. A wireless local area network (WLAN), nevertheless, is required for mobility of AR applications.

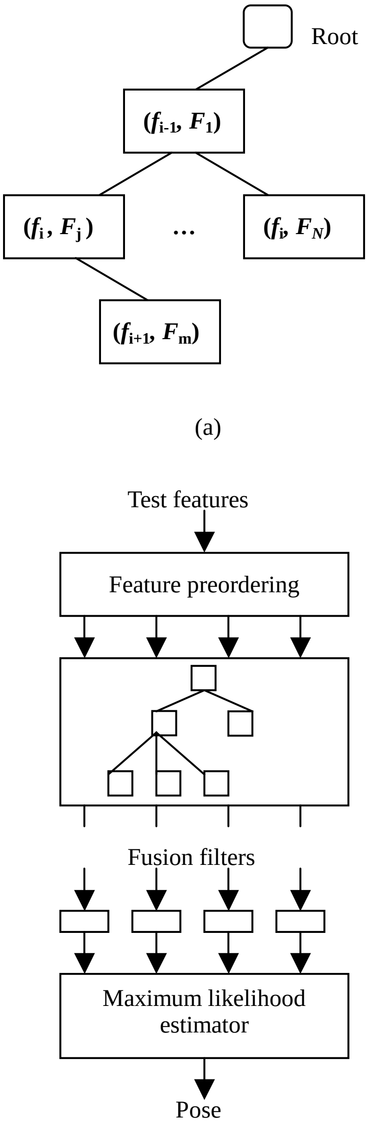
In this paper, we propose a distributed user tracking system, which provides indoors mobile users with accurate 3D positions from 2D image points. A detailed description of the tracking system is given in section II. Afterwards, an application reported in section III demonstrates the capability and potential benefit of our distributed tracking system.

2. SYSTEM ARCHITECTURE

The tracking system proposed in the paper uses client-server architecture with three basis subsystems: the site information server (SIS), mobile units, and the network infrastructure.
We built the server on a high-end PC with sufficient storage space to implement a multimedia database. It serves as the system's central repository for archiving the multimedia information used to construct augmented reality [4]. The SIS communicates this information to the clients via a WLAN. Users in different sites carry the mobile units, which are based on laptop and head-mounted display (HMD). The mobile units request multimedia information from the SIS based on user position and other parameters.

Site information server

We consider the SIS as the heart of the distributed system. It exploits the local multipoint distribution services (LMDS) technology supporting cooperative work of mobile users, and is based on a multimedia object database storing 2D images, 3D models, audio and video clips, and text objects on the specific site. These objects are organized in a dynamic interpretation tree (Ditree) [5] to achieve nearly real-time behavior required for tracking applications. Each feature used during correspondence search has an associated attribute vector $A=[a_1, a_2, \ldots, a_n]$. The elements of this vector denote geometric (such as location in the 3D scene) and other feature-dependent information (such as color). We denote the set of test features found in each tracker frame as $F=[F_1, F_2, \ldots, F_N]$ and the underlying model feature set as $f=[f_1, f_2, \ldots, f_m]$. The interpretation tree can help us establish a list of feasible



(a)

(b)

Figure 1: Interpretation trees.
(a) Basic structure of the interpretation tree. Each node corresponds to a valid pairing of a model and a test feature.
(b) Overview of the dynamic extensions to the interpretation tree.

interpretations given both F and f. To avoid unnecessary computation during tree initialization, we apply a number of constraints based on the feature attributes a, which are implemented as statistical hypothesis tests. Geometric constraints reflect the object's geometry and the imaging system (the perspective invariants, for example). In the final structure, each tree node reflects a valid pairing of a model and a test feature, as Figure 1a shows. Although efficient heuristics limit the time spent during the correspondence search, this static interpretation tree doesn't perform matching well enough to achieve real-time tracking.

The Ditree extends Grimson's static interpretation tree [6], as Figure 1b shows, by the following processing steps [7]:

1) Dynamic search ordering

For a valid interpretation, each test feature's expected location will vary only slightly. Thus, we keep the interpretation tree structure attributes. Test and model feature order is rearranged to reflect each interpretation's quality-of-fit. Thus, strong hypotheses will appear at the beginning of the interpretation tree processing. These, together with the search heuristic and the cut-off threshold, are powerful mechanisms to keep the residual tree complexity (the number of leaf nodes) small.

2) Node management

New feature F_i occurring in the image lead to the insertion of new nodes into the interpretation tree as in the static case. If required, the resultant tree can adapt to the scene and grow. Feature pairings that fail to conform to the set of constraints (for example, the attributes have changed too much between two consecutive tracker frames) must be removed from the tree. As in static trees, subtrees starting at an invalid feature pairing are removed. Thus, the tree is pruned whenever the system detects inconsistent features.

3) Feature recovery

Apart from delivering a list of valid feature parings for every feasible interpretation, the tree delivers the list of model features that haven't been assigned successfully to a test feature and it can be used to recover those features. Assuming that missing features result from partial occlusion, the system can use this information to selectively search the subsequent input image for the missing features.

The Ditree algorithm can deal with multiple hypotheses simultaneously through a set of Kalman filters [8]. It uses the five strongest hypotheses to repeatedly update the corresponding filters. In a final processing step, the system applies a maximum likelihood scheme to identify the object pose for each tracker frame.

The SIS also hosts a suite of authoring tools for creating and editing multimedia content and defining virtual and augmented scenes. The repository facilitates exchanging scientific findings and updating the stored information.

Mobile units

We built mobile units using a laptop and a Sony HMD with variable transparency upon which the AR worlds appear. As Figure 2 shows, the user wears a helmet with a USB video camera, a compass, tilt sensor module (Precision Navigation TCM2), and earphones. The camera is able to capture up to 30 frames per second at 640×480 and also adjust features such as exposure, brightness, contrast, etc. The compass and the module, which is specified to achieve approximately ±0.5 degree of error in yaw at a 16Hz update rate, provide the user's heading and two tilt angles in the local motion frame. In essence, mobile units identify their desire to view this specific scenery's augmentation. They transmit a request by a speech recognition system to the SIS, which mines the corresponding audiovisual data from its database and transmits it back to the mobile units. The system matches the reconstruction model to the live video stream from the camera, transforms it accordingly, and renders it.

Communication infrastructure

The communication infrastructure forms the backbone for the SIS and mobile units to exchange multimedia data and control information. We choose an IEEE 802.11b WLAN because of its good wall penetration and indoor range. We use a total of three access points to cover the whole building, and measurements show sufficient coverage for all areas accessible to visitors.

The access points exploit directional antennas to implement the WLAN and link to their neighbors in different floors using secondary point-to-point links.

The network runs at up to 11 Mbps and can support to 50 users at a time. As a user approaches a new viewpoint, the corresponding information is download, if not yet in the device's hard disk. To minimize wait time (in seconds), the system in the laptop downloads much information according to the current visiting site at the work's start and update it when required.

3. POSITION AND ORIENTATION TRACKING

The tracking system displays the position information on a structure map of the building. Figure 3 shows a sample map from the office building of the university. The colored compass indicates the user's position and viewing angle. The information gives an first estimation of the object the user is looking at, along with the viewing distance and angle. This is further refined by the vision based tracking method we describe next.

Camera tracking

Many tracking systems utilizing vision techniques have been developed to achieve accurate registration [9]. Such systems only work in prepared environments where the system designer has sufficient control over the environment to place calibrated fiducials in the regions of scene. However, calibrating fiducials in a large environment is very difficult. In this paper, we present a robust tracking method in which 2D features extracted from the camera images are used. To overcome difficulties of tracking in changing lighting conditions we track the feature of the threshold image rather than features of the actual image itself [9]. Robust and extendible tracking is achieved by dynamically calibrating the 3D positions of a prior uncalibrated natural feature [10]. Six main steps are included in our tracking method as Figure 4 shows.

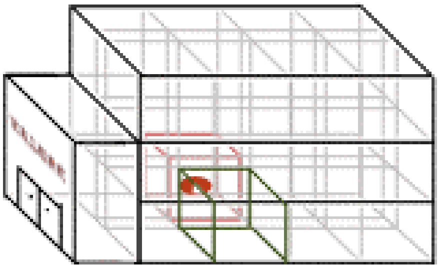
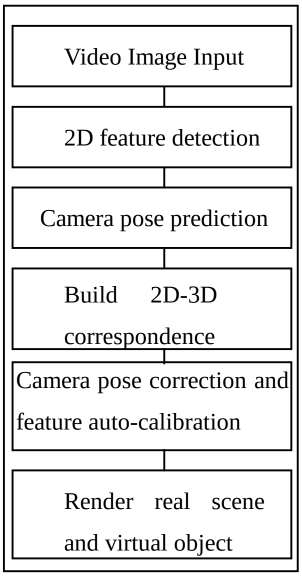


Figure 2: The structure map of our university's office building.
The green compass indicates major scenery's viewpoints. The twinkling red compass indicates the current user's position and orientation

In each new video image we detect 2D fiducials and natural features. Fiducials are used to initialize the camera pose. In subsequent frames, prior pose estimates are used to predict the current pose. More robust and accurate pose estimates are obtained by using the tracked 2D features whose 3D positions are auto-calibrated dynamically from multiple prior images. The result is refined iteratively until the estimated error converges. The final output of the tracker is an accurate estimate of camera pose that specifies a virtual camera used to project the augmented reality media into the scene. Besides the camera pose, our system also produces the set of auto-calibrated natural features that are useful for automatic scene modeling. Through autocalibration, the range of tracking is extended beyond the initially calibrated area into unprepared areas of the environment.


Video Image Input
2D feature detection
Camera pose prediction
Build 2D-3D correspondence
Camera pose correction and feature auto-calibration
Render real scene and virtual object

-*Camera pose computing*: estimating camera pose based on the correspondences of tracked features.
-*2D features tracking*: tracking the inter-frame motion of natural features for pose estimation.
-*2D-3D correspondence*: corresponding 2D image measurement with their calibrated features.
-*Natural feature calibration*: dynamically calibrating the 3D positions of a-priori uncalibrated natural features.

Figure 3: The flow chart of tracking

1) 2D feature detection and tracking

Two types of features, artificial landmarks and natural features are used as tracking primitives.

We adopt a multi-ring color fiducial originally designed by Scalable Fiducial Tracking [11].

Figure 4: User with HMD and other AR kit

A robust motion tracking approach is used for natural feature tracking. The part of the approach is its integration of three motion analysis functions, feature selection, tracking, and verification, in a closed-loop cooperative manner to cope with complex imaging conditions. Firstly, in the feature selection module, 0D (point) and 2D (region) features are selected for their tracking and motion estimation suitability. This selection and evaluation process also uses data from a tracking evaluation function that measures the confidence of a feature's prior tracking estimates.

Once selected, features are ranked according to their evaluation scores and fed into the tracking module. The tracking method is a differential-based local optical-flow calculation that utilizes normal-motion information in local neighborhoods to perform a least-squares minimization to find the best fit to motion vectors. Unlike traditional single-stage implementations, the approach adopts a multi-stage robust estimation strategy. For every estimated result, a verification and evaluation metric assesses the confidence of the estimation. If the estimation confidence is low, the result is refined iteratively until the estimation error converges.

To achieve robust tracking, two different verification strategies are used for the point and region tracking and motion models. Basically, in both cases, an estimated motion field generates a predicted frame that is used to measure the estimation residual. The difference between the predicted frame and the true target frame measures the error of the estimate. This error information is fed back to the tracking module for feature re-evaluation. The process acts as a "selection-hypothesis-verification-correction" strategy that make it possible to discriminate between good and poor estimation features, which maximizes the quality of the final motion estimations.

2) Build 2D-3D correspondence

For tracking and computing camera pose relating to a world coordinate frame, we need to correspond a number of calibrated 3D features and their 2D projections on the image plane. As stated before, two kinds of features, artificial landmarks and natural features are used as tracking primitives in our system. Since the artificial landmarks have their 3D coordinates recalibrated, the tracking system has to reliably detect them and build the correspondence between the detected 2D projections and their tabulated 3D positions. For each new video image, the 2D projections of fiducials are detected. At the same time, the 3D fiducial positions are also projected onto the image plane using a predicted camera pose derived from the previous frame. We note the fact that interframe motion is often small relative to the larger distances between fiducials corresponded. Combining this method with simple color matching and cluster identification to provide the initial correspondences produces a very robust system. Even with only one or two observable fiducials, the system nominally determines the correct 2D-3D correspondences.

3) Camera tracking and natural feature auto-calibration

To track camera pose and dynamically estimate the 3D positions of natural features, we developed an auto-calibration approach based on an iterative Extended Kalman Filter (iEKF). Basically, this Kalman Filter consists of two main processes: pose prediction and measurement correction. In the prediction process, the 2D feature motion and "history" information from pervious estimates are combined to predict the current camera pose. This predicted pose is used to establish the 2D-3D correspondences, as described above. With correspondences, the new image feature measurements are used to correct the pose. This measurement correction step is iterative in that we refine the estimate by applying corrections from one feature at a time. By processing the requirement for a minimum number of available features in order to obtain a pose and we accommodate corrections from features with varying position certainty. In order to calibrate the 3D positions of feature we also maintain each feature's state. The feature database holds a position and error covariance matrix that represents the uncertainty of each feature.

A acceleration model chosen to model target motion assumes the target undergoes a constant acceleration during period δt, and the acceleration is uncorrelated from period to period. This model is expressed as

$$\vec{x}(t) = A\vec{x}(t - \delta t) + \Gamma\vec{v}(\delta t)$$

where A is the state transition matrix implementing

$$x(t) = x(t - \delta t) + \dot{x}(t - \delta t)\,\delta t$$

and $\dot{x}(t) = \dot{x}(t - \delta t)$.

Γ is the noise gain, and $\vec{v}$ is the zero-mean vector whose elements are independent of each other.

Besides the dynamic model we also need a measurement model to correct the state prediction. For our tracking application we use the projections of given 3D features on the image as the measurement model. Our tracking and feature calibration algorithm is introduced as follows:

1) Calculate elapsed time from the last frame and use the camera state in the last frame to predict the current camera state and error covariance matrix. The camera state includes position, orientation, and their derivatives $[x, y, z, \dot{x}, \dot{y}, \dot{z}, \Delta\phi, \Delta\theta, \Delta\psi, \dot{\phi}, \dot{\theta}, \dot{\psi}]$. We maintain incremental orientation in the state vector and the externally accumulated absolute orientation is represented as a quaternion.
2) After computing the 3D-2D correspondence, sort the observable features in increasing order of 3D position uncertainty. For each feature f, augment camera state with feature state and augment camera state error covariance matrix with feature error covariance. Then use the current camera state to project the 3D feature on to the image and use the difference from the corresponding 2D real image measurement to update the camera state and feature state of f.

$$\widehat{x} = [\hat{x}^{-}, x_f, y_f, z_f]$$

$$\widehat{\mathrm{P}}^{-} = \begin{bmatrix} P^{-} & 0 \\ 0 & P_f \end{bmatrix}$$

By iterating the features in increasing order of uncertainty the feature error is small while the camera error is large at the start of the measurement correction iterations. This leads to fast convergence of the estimated camera state. After camera state convergence, The features with relative large errors do not affect camera state significantly. Accurate camera state also leads to rapid feature position convergence. By integrating camera tracking and feature calibration together we can use all of the observable features in the image to obtain smooth and accurate camera pose tracking.

3) For new features without corresponding 3D positions in the database, use the 2D-3D correspondence of feature tracking and the tracked camera pose to estimate their 3D positions and add them to the database. The intersection of two rays connecting the estimated camera pose and the 2D image measurement of the features in two frames create an initial estimate of the 3D position for the feature.

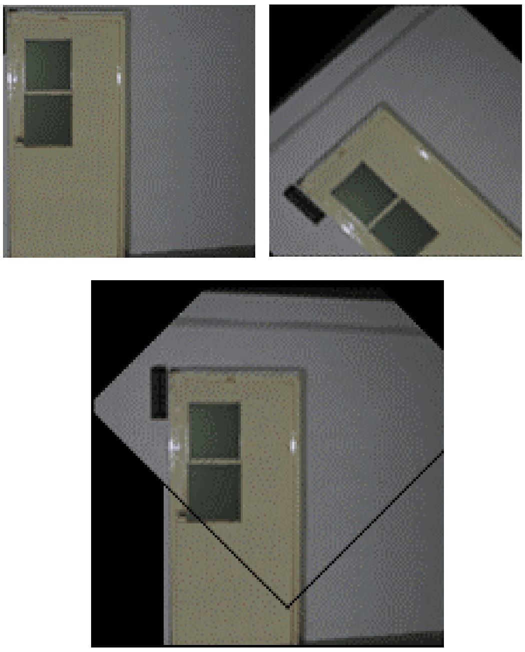

(c)

Figure 5 Registration examples with rotation and translation:

(a) reference image
(b) live video frame
(c) the resembling method adds two images by bilinear interpolation

Camera pose for the first frame is computed by a 3-point method as an initial value for iEKF.

The algorithm's reliability represents the most important aspect of our registration technique choice. Many changes may appear between the live and reference images due to different distance, camera pose and diverse viewpoints. Therefore, we opted for an approach based on the Fast Fourier Transform (FFT) [12] to process the initial 2D image because of its robustness and fast calculation. The FFT enables to recover rotation, scale and translation.

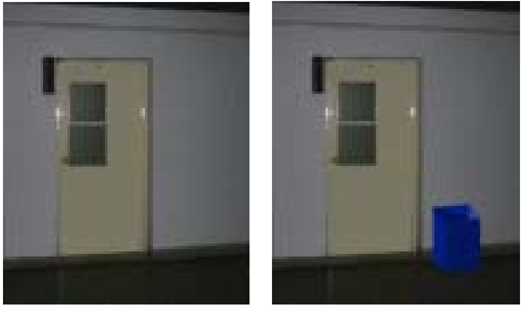

(a) (b)

Figure 6: Rendering example

(a) the gallery in its present state,
(b) augmented gallery with rendered model on the bottom right corner of live video.

Implementation

The computations of 2D image registration for pose estimation are done on basis of the FFT. Thereby, the image must be square and with dimension 2^n. In our implementations, the left and right borders are cut out and the image is scaled down to the next 2^n dimension. Because the Fourier transform assumes a periodic function and the image is truncated, we must apply a window as the Hanning window [13], to the input images.

Another implementation difficultly consists of the numerical instability for coordinates near to the origin, since we have: $\lim_{r\to 0}\rho = \lim_{r\to 0}\ln(r) = -\infty$. Therefore, a high-pass filter is applied on the logarithmic spectra. We use a filter with the following transfer function:

$$H(x,y)=(1.0-\cos(\pi x)\cos(\pi y))(2.0-\cos(\pi x)\cos(\pi y))$$

With: $-0.5 \le x, y \le 0.5$

(We could, instead, directly set to zero the points near the origin and inside a circle of radius e [14].)

The tracking algorithm runs sequentially at 30 frames per second on a laptop and for a camera resolution of 640×480 pixels. Figure 5 shows an example where a translated and rotated video frame (Figure 5b) is registered to a reference image (Figure 5a) from the database.

Rendering

Recalling that the reference images are stored along with (aligned) 3D models of reconstructed scenery. The render can thus render the transformed 3D models on top of each live video frame and present the user with an augmented view.

Figure 6 shows a typical example, where the natural view from the user's viewpoint precedes the same view augmented with the 3D model. This image appears on the augmented reality glasses users wear, and the rendering process follows their movement.

4. CONCLUSIONS

In this paper, we propose a distributed tracking system integrating mobile AR-tools into the communication infrastructure of buildings controlled by computers. The future work includes fulfilling the real-time AR reconstruction by improving the algorithm and enhancing the data speed of the WLAN by IEEE 802.11g. We will also explore the joint analysis of information taken from multiple users.

5. REFERENCES

[1] F. Betting, J. Feldmar, et al. A New Framework for Fusing Stereo Images with Volumetric Medical Images. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pp. 30-39, 1995.

[2] E. Foxlin, M. Harrington, and G. Pfeiffer. Constellation^{TM}: A Wide-range Wireless Motion-tracking System for Augmented Reality and Virtual Set Applications. *Proc. SIGGRAPH*, pp. 371-378, 1998

[3] V. Vlahakis, N. Ioannidis, J. Karigiannis, et al. "Archeoguide: An Augmented Reality Guide for Archaeological Sites," *IEEE Computer Graphics and Applications*, September 2002, pp. 52-59.

[4] R. Azuma, “A Survey of Augmented Reality,” *Proc. Siggraph 95*, course notes no. 9(Developing Advanced Virtual Reality Applications), ACM Press, New York, 1995.

[5] M. Brandner and A. Pinz, “Real-time Tracking of Complex Objects Using Dynamic Interpretation Tree,” *Pattern Recognition, Proc. Of 24th DAGM Symp., Lecture Notes in Computer Science 2449*, Springer-Verlag, Berlin, 2002, pp. 9-16.

[6] W.E.L. *Grimson, Object Recognition by Computer: The Role of Geometric Constraints*, MIT Press, Cambridge, Mass., 1990

[7] M. Ribo, P. Lang, et al. “Hybrid Tracking for Outdoor Augmented Reality Application,” *IEEE Computer Graphics and Applications*, November 2002, pp. 54-63.

[8] A. Gelb (ed.), *Applied Optimal Estimation*. MIT Press,Cambridfe, MA, 1974.

[9] R. T. Azuma, “A Survey of Augmented Reality,” *Teleoperators and Virtual Environment*, Aug. 1997, pp. 355-385.

[10] M. M. Clothier, “Overcoming Difficulties of Tracking in Changing Lighting Conditions,” Department of Computer Science, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093.

[11] Young Kwan Cho, Scalable Fiducial-Tracking Augmente Reality, Ph.D. Dissertation, Computer Science Department, University of Southern California, January 1999.

[12] B.Jiang, S. You, and U. Neumann, “Camera Tracking for Augmented Reality Media,” Integrated Media Systems Center,University of Southern California, Los Angeles, CA 90089-0781.

[13] D. Stricker, “Tracking with Reference Images: A Real-Time and Markless Tracking Solution for Out-Door Augmented Reality Applications,” Fraunhofer Institute for Computer Graphics, Rundeturmstrae 6, 64283 Darmstadt, Germany.

[14] M. Bilinghurst, H. Kato, and I. Poupyev, “The Magic Book: An Interface that Moves Seamlessly between Reality and Virtuality,” IEEE Computer Graphics and Applications, vol.21, no.3, May/June 2001, pp.6-8.

[15] M. Uenohara and T. Kanade, “Vision-Based Object Registration for Real-Time Image Overlay,” *Int’l J. Computers in Biology and Medicine*, vol. 25, no. 2, Mar. 1995, pp. 249-260.

Approach on Visual Federation Member Relationship

Xu Dong-ping Qin Juan
Science and Technology of Computer, Wuhan University of Technology
Wuhan 430063, China
Email: DPXU@public.wh.hb.cn **Tel**.: +86(0) 27-86551167

ABSTRACT

In network that has limited bandwidth, in order to get a fluent vision of motion, few parameters are transferred. Visual Federation is an effective method that can control the parameters. Industry design, a sham battle, and city plan simulation, virtual reality simulation train, and interactive entertainment simulation fields will use such technique. In this article, we define and introduce the approach on Visual Federation Members and it's relationship to solve the problem of interactive scene.

Keywords：Visual Federation, Member of Visual Federation, Relationship

1、 INTRODUCTION

In network that has limited bandwidth, in order to get a fluent vision of motion, few parameters are transferred. Visual Federation is an effective method that can control the parameters. Industry design[1], a sham battle, and city plan simulation, virtual reality simulation[2] train, interactive entertainment simulation fields will use such technique. In Visual Simulation[3], we set up distributing simulation Visual Federation Object Model (VFOM) [4], using data management thought of HLA, referring to OMT of HLA[5]. It resolves the problem of interactive scene.

2、 MEMBER OF VISUAL FEDERATION AND RELATIONSHIP

We can construct VFOM with three tables: motion command table, announced motion object table, and motion object's parts table which has main-parts and sub-parts. Main-parts is publicized and ordered directly by other federation members, while sub-parts is not publicized and it can be driven by main-parts to reduce the transferring information.

2.1. Motion Command Table

Motion command table is an relation table between simulation federation members and visual simulation members. Its definition refer to Table 1.The data of command window will refresh each △tc（as 300ms）, or refresh when simulation member give an order .

ID of announced motion object , made up of ASCⅡ string , can identify every announced motion object ;

ID of announced-main-parts , made up of ASCⅡstring , can identify every announced-main-parts . The data of this ID is the name of announced motion object's main- parts.

Command Speed can make Visual Simulation System to repair speed. It is expressed by a 6-dimension vector （C_x, C_y, C_z, C_h, C_p, C_r）.

Command Fitting Location is expressed by a 6-dimension vector（P_x, P_y, P_z, P_h, P_p, P_r）.

Location Flag is a Location Accordance Flag from Control Simulation Federation Member. The initial data is 1, and the data will be changed to 0 after adjusted by Visual Simulation Federation Location Accordance.

2.2. Announced Motion Object Table

Announced motion object is an object which could be controlled by another federation member. This table includes the motion attribute of object. Control Simulation System maybe public motion command for such object. For example Table 2.

ID of announced motion object , made up of ASCⅡ , can identify every announced motion object ;

Current Speed is expressed by a 6-dimension vector (V_x, V_y, V_z, V_h, V_p, V_r) . It represent the parallel speed at the X ,Y , Z axes and the angle speed at circumrotating Z，X，Y axes by right theorem .

Speed Granularity is a kind of speed increment in the time of frame refreshment. Speed Granularity is expressed by a 6-dimension vector (a_x, a_y, a_z, a_h, a_p, a_r). It represent the parallel speed increment at the X ,Y , Z axes and the angle speed increment at circumrotating Z，X，Y axes by right theorem .

Command Speed can make Visual Simulation System to repair speed. It is expressed by a 6-dimension vector （C_x, C_y, C_z, C_h, C_p, C_r）. Command Fitting Location is expressed by a 6-dimension vector （C_x, C_y, C_z, C_h, C_p, C_r）. Command Fitting Location is expressed by a 6-dimension vector（P_x, P_y, P_z, P_h, P_p, P_r）.

The method to Locate:
①refresh w follow the steps (w contain x , y , z , h , p and r . Suppose the value of x , y , z , h , p and r point the current location of this object),
②if (C_w != V_w and $|C_w – V_w|>k$) then $V_w = V_w + (C_w - V_w) / |C_w — V_w|*a_w$*the time of this frame refreshment
③$w=w + V_w$ * the time of this frame refreshment
④if (Location Flag = = 1) $w=(w ＋ P_w) / 2$，Location Flag =0; Adopting w to locate

Announced motion object's VSOM adopt the global Coordinate to locate. VSOM which can be controlled certainly is in the announced motion object table. These announced motion object queue, for example Figure 1.

Table 1 Motion Command Table

ID of announced motiong object	<ID of announced motiong object>
ID of announced-main-parts	<ID of announced-main-parts >
Command Speed	<Command Speed>
Command Fitting Location	<Command Fitting Location>
Location Flag	<Location Flag>

Table 2 Announced motion object table

ID of announced motion object	<ID of announced motion object>
Current Speed	<Current Speed>
Speed Granularity	<Speed Granularity>
Command Speed	<Command Speed>
Command Fitting Location	<Command Fitting Location>

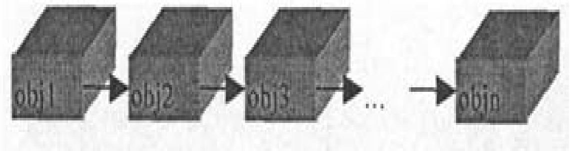


Figure1 Announced motion object

2.3. Motion Object Parts Table

Motion object parts table descript the parts' moving chain and the moving attributes. Motion object parts adopt part coordinate to locate and this is related to one object or another parts. Motion object parts include announced main-parts which can directly get the moving command from other federation member, and unpublicized sub-parts which can move by the pull power from the motion of main-parts. The relation of traction motion is defined by the traction locating equation.

Object parts motion chain refers to Figure 2.

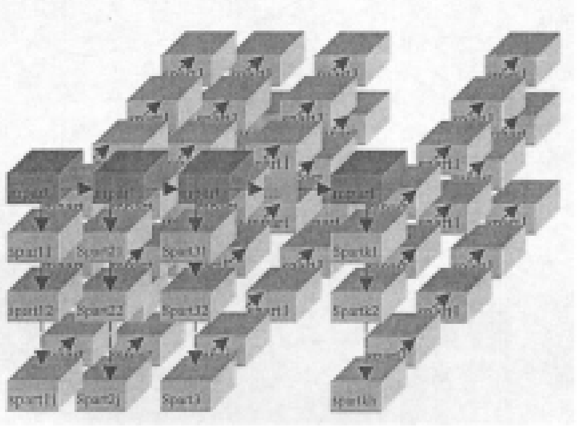


Figure 2 Object parts motion chain

Motion Object Main-parts Table:
For example Table 3.

Motion object main-parts is publicized, in other words , it can get the moving command from another federation member . In the table we can know:

ID of object made up of ASCⅡ represent the name of VSOM;ID of parts made up of ASCⅡ represent the name of Visual Simulation Object Parts;

M_w , a 6-dimention vector (M_x, M_y, M_z, M_h, M_p, M_r), means the partes'upper limit of freedom degree;

L_w , a 6-dimention vector (L_x, L_y, L_z, L_h, L_p, L_r), means the partses' lower limit of freedom degree;

V_w, a 6-dimention vector (V_x, V_y, V_z, V_h, V_p, V_r), means the partses' current speed;C_w, a 6-dimention vector (C_x, C_y, C_z, C_h, C_p, C_r), means this partses' moving speed from another federation member;a_w, a 6-dimention vector (a_x, a_y, a_z, a_h, a_p, a_r) , means the a kind of speed increment in the time （$\triangle t_R$ (30ms ─ 50ms)） of frame refreshment , and reflects inertial character of such parts;

M_p is a point, which direct to the prior main motion parts. In the motion chain, the M_p of first main-parts direct to itself;

S_p is a pointer which direct to next main motion parts. In the motion chain, the value of last main-parts' M_p is zero;

P_w, a 6-dimention vector (P_x, P_y, P_z, P_h, P_p, P_r), means the data of fitting lcation;

B_L is a pointer, which direct to next sub motion parts. Each main-parts can directly lead several sub-partses to move, but only B_L that pointed to one sub-parte need to preserve in the expression of the main-parts .

It is at intervals of the vision frame refreshment period $\triangle t_R$(30ms ─ 50ms) , making to the main parts in all announcements as follows position:

①presses as follows the step make the w(the w takes the x, y, z, h, p, r, establish current position in this object as the x, y, z, h, p, r) get the renewal;
②if (C_w!=V_w and $|C_w - V_w| > k$) then $V_w = V_w + (C_w - V_w) / |C_w - V_w| * a_w * \triangle t_R$
③$w=w+V_w*\triangle t_R$;
④if ($w>M_w$) then $w= M_w$ else if ($w<L_w$) then $w= L_w$;
⑤if (Location flag==1) $w= (w+P_w) / 2$，Location flag=0 ; Adoption w to locate

Motion Object sub-parts Table: Motion object sub-parts is the unannounced motion parts. It can't accept other federation member's motion orders, and it be led by its main parts or other sub-parts (current sub-parts and father-parts) . Such as

Table 4:

Table 3 motion object main-parts table

ID of Object	ID of Parts	M_w	L_w	V_w	C_w	a_w	M_p	S_p	B_L	P_w
< ID of Object>	<ID of Parts>	$<M_w>$	$<L_w>$	$<V_w>$	$<C_w>$	$<a_w>$	$<M_p>$	$<S_p >$	$<B_L>$	$<P_w>$

Table 4 Motion object sub-parts table

ID of Object	ID of part	M_w	L_w	W	M_p	S_p	B_L	B_R
<ID of Object>	<ID of part>	$<M_w>$	$<L_w>$	<W>	$<M_p>$	$<S_p>$	$<B_L>$	$<B_R>$

Object of ID, part of ID, M_w, L_w inside this table are the same meaning as inside motion object main parts table.
W means the equation of leading motion. This equation is expressed by 6 strings of ASCⅡ which mean the current sub-parts of 6 free fixed position data . The definition of equation as follow:

W ∷= <expression>
<expression> ∷= <expression> <⊕> T | T
T ∷= T <⊙> F | F
F ∷= (<expression>) |
<function> (<expression> [，<expression>]) |
< a fixed positiong parameter of father node in chain > | <constant> | <variable>
< a fixed positiong parameter of father node in chain > ∷ = x | y | z | h | p | r
<constant> ∷= <integer> | <real> |
< symbol constant>
<variable> ∷= ASCⅡ string started by char
<function> ∷= sin | cos | log | tan | acos |asin | atan | sqrt | pow | exp | ceil | floor | fabs | fmod | round | sqr
⊕ ∷= + | -
*⊙ ∷= * | /*

W means the 6 free coordinate of sub-parts’ current position . The variables , x, y, z, h, p, r ， mean the fixed positiong parameter of father-node in motion chain .

S_p means the indicator pointing to the sub-parts which is led by the current parts’ father-parts (lead the sub- parts for the first time, also can call to lead head). If it has no the subsequence sub- parts, its value is empty.

M_p means that the pointer of father-parts’ table-node.The father-parts may be a main parts, also may be a sub- parts . If the current parts is not for the first time to lead the sub- parts, its value is empty.

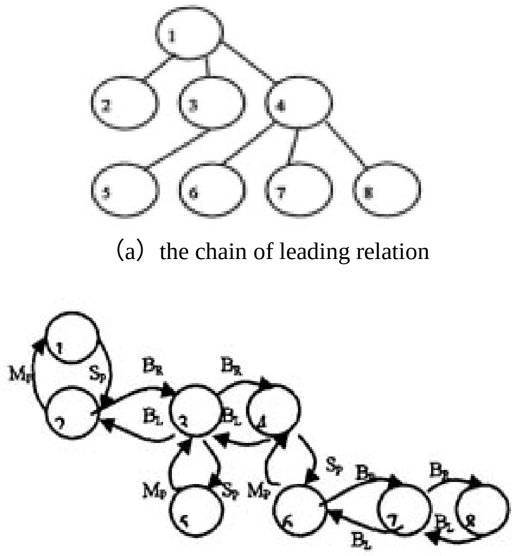
The all sub- partses which in double -chain that Mp , Sp consist constitutes to drawing-chain.

B_L is an indicator . It point to the sub-parts of pre-brothers which in the same drawing-source .The B_L value of the first sub-parts in the same drawing-source brothers is empty.

B_R is similar to the above. It point to together the sub- parts of pre-brothers which in the same drawing-source . And the B_R value of the last one which in this chain is empty.

The all sub- partses which in double -chain that B_L , B_R consist have the same drawing-source, but among these sub-partses have not any sport relation.

The above motion chain is abstract to express with "forest" integrated by the "tree" of foot-node, which regarded as a series motion-objects announcement main parts. These announced main parts can accept other federal member's motion orders directly. They become the final drawing-source of their "descendant" node parts. Adopting the above chain, you can apply unified fixed node construction to express, in order to solute the node out-degree uncertain problem in tree construction. Figure 3(a) shows that the announced main-parts lead to the related token-tree. The quantity of the other parts, draw by one parts, is uncertain. In the diagram,"1" lead "2", "3", "4" ; "3" lead" 5" ; "4" lead " 6", "7", "8" . Figure 3(b) shows the chain of leading relation . Each node simply includes 4 indicator pointers. It can express that each parts draw uncertain number other parts. Each part at most have a drawing-source. That to say, each token-node have a father-node at most.



(a) the chain of leading relation

(b) the token-tree of leading relation

Figure 3 Leading relation

3、 CONCLUSIONS

All the tables we mentioned can be divided into two classes: one is static state data which can be dealt with beforehand. It not only include inherence attribute of object such as speed granularity of Announced Motion Object, motion object mainpart’s motion upper limit of freedom degree, but also include the motion object part, the Initial value of announced motion object motion parameter. Traction location equation also can be give as static data beforehand. The second is dynamic data, viz. the data in motion command table, which is provided by control simulation member when the system running. Above method ensure motion visual object statement parameter error in vision simulation within the permitted range. As the same time It can maximumly debase the data redundancy and the data transmission quantity , so we can get the fluent vision .

4、REFERENCES

[1] Xiao Tianyuan, et al. Next Generation Manufacturing---distributed Interactive Virtual Product Development[A]. System Simulation and Scientific Computing[C]. October 19-21,1999,Beijing China. 23-27.

[2] Li Bohu，Two Focuses in the Development of Contemporary Simulation Technology——ADS，SBA，Journal of System Simulation，2001 vol.13 no.1，101-105.

[3] Loftin R B. Aerospace Applications of Virtual Environment Technology[J]. Computer Graphics, November 1996: 33-35.

[4] Xu Dongping. Real Time Dynamic Interactive Visual Simulation: [Ph. D. thesis]. Wuhan University of Technology，2001

[5] DMSO. HLA Rules 1.0, 21 August 1996. Http://www.dmso.mil

Security Analysis and Improvement of Some Threshold Proxy Signature Schemes

Xue Qingshui **Cao Zhenfu**
Dept. of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200030, China
Email: {xue-qsh, zfcao}@cs.sjtu.edu.cn **Tel.:** 86-021-62932951

ABSTRACT

We review Hsu et al's threshold proxy signature scheme with known signers, describe the Tsai et al's attack to Hsu et al's scheme and point out that their attack can't work. We also review Hsu et al's another scheme with unknown signers, analyze its security and point out that it can not resist the public key substitution attack. Based on Hsu et al's second scheme, an improved version which can resist the weakness is proposed.

Keywords: Cryptography, Digital signatures, Proxy signature, Threshold proxy signature

1. INTRODUCTION

The proxy signature scheme [1], a variation of ordinary digital signature schemes, enables a proxy signer to sign messages on behalf of the original signer. Proxy signature schemes are very useful in many applications such as electronics transaction and mobile agent environment.

Mambo et al. [1] provided three levels of delegation in proxy signature: full delegation, partial delegation and delegation by warrant. In full delegation, the original signer gives its private key to the proxy signer. In partial delegation, the original signer produces a proxy signature key from its private key and gives it to the proxy signer. The proxy signer uses the proxy key to sign. As far as delegation by warrant is concerned, warrant is a certificate composed of a message part and a public signature key. The proxy signer gets the warrant from the original signer and uses the corresponding private key to sign. A lots of proxy signature scheme were proposed [2]-[16].

The threshold proxy signature schemes were proposed [2[[6]-[14]. In the threshold proxy signature scheme, a group of *n* proxy signers share the secret proxy signature key. To produce a valid proxy signature on the message *m*, individual proxy signers produce their partial signatures on that message, and then combine them into a full proxy signature on *m*. In a *(t, n*) threshold proxy signature scheme, the original signer authorizes a proxy group with *n* proxy members. Only the cooperation of *t* or more proxy members is allowed to generate the proxy signature. Threshold signatures are motivated both by the demand which arises in some organizations to have a group of employees agree on a given message or document before signing, and by the need to protect signature keys from the attack of internal and external adversaries.

In 1999, Sun proposed a threshold proxy signature scheme with known signers [9]. Then Hwang et al. [7] pointed out that Sun's scheme was insecure against collusion attack. By the collusion, any *t*-1 proxy signers among the *t* proxy signers can cooperatively obtain the secret key of the remainder one. They also proposed an improved scheme which can guard against the collusion attack. After that, [6] showed that Sun's scheme was also insecure against the conspiracy attack. In their attack, *t* malicious proxy signers can impersonate some other proxy signers to generate valid proxy signature. To resist the attack, they also proposed a scheme. [8] Showed that the scheme in [7] was also insecure against the attack by the cooperation of one malicious proxy signer and the original signer. In 2002, Li et al. [2] proposed a threshold proxy signature scheme with lots of good performances. In [14], we point out there are some errors in Sun's and Hwang et al.'s scheme and also proposed an improved version.

Recently, Tsai et al pointed out that Hsu et al's scheme is insecure against the public key substitution attack from the original signer and the forgery attack from the proxy signer [13]. In the letter, we will point out that their attack can't work, thus, their improved version to resist the attack is of insignificance. In addition, Hsu et al proposed another threshold proxy signature scheme with unknown signers based on Sun et al's scheme [12]. We point out their scheme is insecure against the public key substitution attack from the original signer or proxy signers. Furthermore, based on Hsu et al's second scheme, an improved scheme is proposed by us.

In the paper, we will first review Hsu et al's scheme. Then we describe Tsai et al's attack to Hsu et al's scheme. In session 4, we analyze the attack and point out that it can't work. We review Hsu et al's another scheme with unknown signers in session 5. In session 6 we analyze the security of Hsu et al's second scheme and point out that it can not resist the public key substitution attack either. Based on Hsu et al's second scheme, an improved version which can resist the weakness is proposed in session 7. In session 8, we make some remarks on the improved scheme. Finally, we draw the conclusion.

2. REVIEW OF HSU ET AL'S SCHEME [6]

In the scheme, a system authority (SA) whose tasks are to initialize the system and to manage the public directory, the original signer U_O, certificate authority (CA), the proxy group of *n* proxy signers $G_P = \{U_{P_1}, U_{P_2}, ..., U_{P_n}\}$ and the signature verifier *V* are needed.

Throughout the paper, *p* and *q* are two large primes with $q \mid (p-1)$ and *g* is a generator of $GF(p)$ with order *q*. *h* is a secure one-way hash function. M_w is a warrant which records the identities of the original signer and the proxy signers of the proxy group parameters *t* and *n*, the valid delegation time, etc. $ASID$ (Actual Signers' *ID*) denotes the identities of the actual signers. PGID={EM, Time, Group} is the proxy group identity which records the proxy status, in which EM denotes the event mark of the proxy share generation including the parameters *t* and *n*, Time denotes the expiration time of the delegation of signing power, and Group denotes the identities of the original signer and the proxy signers of G_P.

Each user U_i with the public identifier $v_i \in Z_q$, owns a private key $x_i \in Z_q^*$ and a public key $y_i = g^{x_i} \bmod p$ which is certified by CA.

The scheme includes four phases: secret share generation, proxy share generation, proxy signature generation and proxy signature verification.

Secret Share Generation

SA chooses the group private key X_G and computes the group public key $Y_G = g^{X_G} \bmod p$ which is certified by CA. Then SA selects a $(t-1)$ -degree polynomial $f(v) = X_G + a_1 v + a_2 v^2 + ... + a_{t-1} v^{t-1} \bmod q$ where $a_i \in Z_Q (i = 1,2,...,t-1)$ is choosed at random.

For each proxy signer U_{P_i}, SA computes the secret share $\gamma_i = f(v_i)$ and the corresponding public $\tau_i = g^{\gamma_i} \bmod p.$ Then, SA separately sends γ_i to U_{P_i} via a secure channel and publishes all $\tau_i 's.$

Proxy Share Generation

The original signer U_O performs the following steps to delegate the signing capability to G_P.

1) Choose a random number $k \in Z_q^*$ and compute $K = g^k \bmod p.$

2) Compute $\sigma = k + x_o h(M_w, K) \bmod q$ as the proxy signature key.

3) Choose a $(t-1)$ -degree polynomial $f_o(v) = \sigma + b_1 v + b_2 v^2 + ... + b_{t-1} v^{t-1} \bmod q$, where $b_j \in Z_q (j = 1,2,...,t-1)$ is selected at random.

4) Publish $B_j = g^{b_j} \bmod p$ for $j = 1,2,...,t-1.$

5) Send $\sigma_i = f_o(v_i)$ to $U_{P_i} \in G_P$ via a secure channel.

6) Broadcast $(M_w, K).$

Upon receiving σ_i, each $U_{P_I} \in G_P$ confirms its validity by checking the following congruence

$$g^{\sigma_i} = y_o^{h(M_w,K)} K(\prod_{j=1}^{t-1} B_j^{v_u^{\,j}}) \bmod p. \qquad (1)$$

If it holds, U_{P_i} computes $\sigma_i' = \sigma_i + \gamma_i h(M_w, K) \bmod q$ as his/her proxy share.

Proxy Signature Generation

Given a message *M*, at least *t* proxy signers of G_P sign *M* on the behalf of the original signer U_O. Without loss of generality, assume that $D_P = \{U_{P_1}, U_{P_2}, ..., U_{P_t}\}$ are the actual proxy signers. D_P as a group performs the following steps to generate the proxy signature.

1) Each U_{P_i} selects a random number $k_i \in Z_q^*$ and broadcasts $r_i = g^{k_i} \bmod p.$

2) Receiving all of $r_j 's(j = 1,2,...,t; j \neq i)$, each U_{P_i} computes

$$R = \prod_{j=1}^{t} r_j \bmod p.$$

$$s_i = k_i R + (L_i \sigma_i' + x_{P_i}) h(R, ASID, M) \bmod q. \qquad (2)$$

where $L_i = \prod_{j=1, j\neq i}^{t} (-v_j)(v_i - v_j)^{-1} \bmod q$. Here s_i is sent to the designated clerk as the individual proxy signature.

3) Receiving s_i, the designated clerk validates it by checking

$$g^{s_i} = r_i^R (((y_o \tau_i)^{h(M_w,K)} \cdot (\prod B_j^{v_i^{\,j}}) K)^{L_i} y_{P_i})^{h(R,ASID,M)} \bmod p \qquad (3)$$

If it holds, (r_i, s_i) is the valid individual proxy signature of *M*. If all of the individual proxy signatures of *M* are valid, the clerk computes

$$S = \sum_{j=1}^{t} s_j \bmod q \qquad (4)$$

The threshold proxy signature of *M* is $(R, S, K, M_w, ASID)$.

Proxy Signature Verification

Receiving the threshold proxy signature $(R, S, K, M_w, ASID)$ of *M*, any verifier can confirm the validity of the proxy signature and identify the actual signers. The steps of the phase are stated as follows:

1) By M_w and *ASID*, the verifier can identify the original signer and the proxy signers, and get the public keys from the CA. Besides, he/she can also identify the actual proxy signer.

2) The verifier validates the proxy signature by checking

$$g^S = R^R (K(y_o Y_G)^{h(M_w,K)} \prod y_{P_i})^{h(R,ASID,M)} \bmod p \qquad (5)$$

If it holds, the proxy signature $(R, S, K, M_w, ASID)$ is valid.

3. THE ATTACK PROPOSED BY TSAI ET AL TO THE ABOVE SCHEME

Assume that a malicious original signer (U_o) or a malicious proxy signer (U_{P_k}) can forge the valid signature without the other signers' private keys.

A. Public Key Substitution Attack

Suppose the malicious original signer without any private keys of the other proxy signers, attempts to forge a valid proxy signature for a message. The steps of the attack are described as follows:

1) U_o Chooses at random a private key $x_o \in Z_q^*$.

2) U_o Waits until he/she gets any *t* or more proxy signers' public key y_{P_i} . Then, instead of broadcasting $y_o = g^{x_o} \bmod p$, he/she computes

$$y_o' = g^{x_o} (Y_G)^{-1} \prod_{i=1}^{t} y_{P_i}^{-h(M_w,K)^{-1}} \bmod p \qquad (6)$$

and reveals the value of y_o' as his/her public key.

3) U_o selects two random integers *k* and *r*, and computes *K* and *R* as follows:

$$K = g^k \bmod p\,.$$

$$R = g^r \bmod p\,.$$

4) U_o selects a message *M* at will and computes *S* as

$$S = rR + (k + x_o h(M_w, K) h(R, ASID, M) \bmod q \qquad (7)$$

Then, the proxy signature of *M* is $(R, S, K, M_w, ASID)$.

B. Insider Forgery Attack

Assume a malicious proxy signer U_{P_k} without any private key of the other proxy signers, attempts to forge a valid proxy signature for an arbitrary message. The attack can take the following steps.

1) U_{P_k} At random chooses private key $x_{P_k} \in Z_q^*$.

2) U_{P_k} Waits until he/she obtains any $t-1$ or more proxy signers' public key y_{P_i}. Then instead of broadcasting $y_{P_k} = g^{x_{P_i}} \bmod p$, he/she computes

$$y_{P_k}' = g^{x_{P_k}} (K(y_o Y_G)^{h(M_w, K)} \prod_{i=1}^{t-1} y_{P_i})^{-1} \bmod p \qquad (8)$$

and reveals y_{P_k}' as his/she public key.

3) U_{P_k} selects a random number *r* and computes *R* as follows:

$$R = g^r \bmod p\,.$$

4) U_{P_k} selects an arbitrary message *M* and computes *S* as

$$S = rR + x_{P_i} h(R, ASID, M) \bmod q \qquad (9)$$

Then, the proxy signature for *M* is $(R, S, K, M_w, ASID)$.

4. THE INVALIDITY OF THE ABOVE ATTACK

From the above two kinds of attack, it can be seen that they are all public substitution attack, one of which is from the malicious original signer and another of which is from the malicious proxy signer. In fact, the two kinds of attack can't work to Hsu et al's scheme. The causes are described as follows.

As described in Hsu et al's scheme, a system authority (SA), whose tasks are to initialize the system and to manage the public directory, and a certificate authority (CA), whose tasks are to certified the original signer's or proxy signers' public key, are needed. That is, if the original signer or any proxy signer wants to substitute a new public key for his/her former public key, CA will confirm whether he/she have the knowledge of the corresponding private key. If he/she does, CA will allow he/she to change his/her public key and refresh the public key directory; or else CA will refuse he/she to change his/her public key. In the two kinds of public key substitution attack, the original signer or the proxy signer does not have the knowledge of corresponding private key, thus they can't change their public keys. In Hsu et al's scheme, when proxy signer signature is verified, the verifier will first identify the original signer and the proxy signers by M_w and *ASID*, then get the necessary public keys from the CA. From above analysis, the original signer or any proxy signer can't substitute a new public key for his/her former public key, therefore, the verifier is unable to get the new public key from the CA. In addition, even if the malicious original signer or any proxy signer can change his/her public key, because of he/she have no the knowledge of corresponding private key, the CA or the original signer is not capable of changing M_w, which should be changed if the original signer's public key is substituted as some parts of message of M_w is encrypted by the original signer's private key. When M_w is verified by the verifier, M_w is confirmed invalid, which will result in the invalidity of the proxy signature. All in all, the two kinds of public key substitution attack can't succeed.

As far as the public key substitution attack is concerned, many literatures mentioned it. Generally speaking, there are two main ways to resist public key substitution attack. The first main way, which is applicable to all of proxy signature schemes, is proposed by Li et al [2]. Li et al's way is based on the idea of zero-knowledge. That is, when the proxy signature is verified by the verifier, he/she first randomly selects a number *r,* which is named "challenger" by Li et al, then encrypts *r* by the signers' public key and sends the encrypted message to the signers. Upon receiving the encrypted message, the signers will decrypted the encrypted message by their private keys and have to send the decrypted message, which is *r* if the signers have the corresponding private key, to the verifier. The first way cost more computation than the second way. The second way is that a certificate authority (CA) or a system authority (SA) is needed. The CA/SA will be responsible for initializing the system parameters, managing the public directory and certificating the signers' public key. If the signers want to change their public keys, the CA/SA will confirm whether the signers have the knowledge of the corresponding private key. If the signers have corresponding private keys, the CA/SA will allow the signers to change their public key and update the signers' public keys in the public directory, which is only managed by CA/SA and any other third party can't update; or else the signers are unable to change their public keys. When the proxy signature is verified, the verifier will obtain the signers' public keys from the CA/SA, not from the signers. The advantage of the second way is that in proxy signature verification phase, the verifier will perform least computation than the first way. The weakness of the second way is that CA/SA is needed and makes the proxy signature scheme insecure if the CA/SA is not honest. To date, to resist the public key substitution attack, when we design proxy signature schemes, one of the two ways can be selected. That is to say, the public key substitution attack is relatively easy to be resisted.

In the following session, we will review Hsu et al's another threshold proxy signature scheme with unknown signers [12], and point out it can't resist the public key substitution attack from the original signer or the proxy signers.

5. HSU ET AL'S ANOTHER THRESHOLD PROXY SIGNATURE SCHEME WITH UNKNOWN SIGNERS [12]

Proxy Share Generation

1) The original signer U_o at random selects $\tilde{k} \in Z_q^*$, computes and broadcasts $\tilde{r} = g^{\tilde{k}} \bmod p$.

2) Upon receiving $\tilde{r}$, each proxy signer $U_{P_i} \in G_P$ randomly selects $\alpha_i \in Z_q^*$, computes and broadcasts $r_i = g^{\alpha_i} \cdot \tilde{r} \bmod p$.

3) Once collecting all $r_i's$ from $U_{P_i} \in G_P$, U_o computes

$$\tilde{s} = x_o h(r, PGID) + n\tilde{k} \bmod q \text{ , } r = \prod_{i=1}^{n} r_i \bmod p \qquad (10)$$

And performs a (t,n) verifiable threshold secret sharing scheme [16], denoted as (t,n) -VSS scheme to share $\tilde{s}$ among n proxy signers in G_P . That is, U_o chooses a $(t-1)$ -degree polynomial $f''(x) = \tilde{s} + a_1''x + a_2''x^2 + ... + a_{t-1}''x^{t-1} \bmod q$ and sends $s_i' = f''(i)$ to U_{P_i} for $i = 1,2,...,n$ secretly. Meanwhile, U_o publishes $c_i'' = g^{a_i''} \bmod p$ for $i = 1,2,...,t-1$.

4) Upon receiving the $s_i' = f''(i)$ from U_o , each $U_{P_i} \in G_P$ can confirm it by checking that

$$g^{\tilde{s}_i} = y_o^{h(r,PGID)} \cdot \tilde{r}^{n} \prod_{j=1}^{t-1} (c_j'')^{i^j} (\bmod p) \qquad (11)$$

If it holds, each $U_{P_i} \in G_P$ performs a (t,n) -VSS and acts as a dealer to distribute proxy sub-shares to other $n-1$ proxy signers for generating their valid proxy shares. That is, each $U_{P_i} \in G_P$ chooses a $(t-1)$ -degree polynomial

$$\begin{aligned} f_i(x) = \alpha_i + x_i h(r, PGID) + a_{i,1}x + \\ a_{i,2}x^2 + ... + a_{i,t-1}x^{t-1} (\bmod q) \end{aligned} \qquad (12)$$

Then $U_{P_i} \in G_P$ sends the proxy sub-share $f_i(j)$ to proxy signer U_{P_j} (for $1 \le j \le n$ and $j \ne i$) via a secure channel and broadcasts $c_{i,k} = g^{a_{i,k}} \bmod p$ for $k = 1,2,...,t-1$. To ensure the validity of $f_j(i)$ sent from U_{P_j} , U_{P_i} can check whether the equality

$$g^{f_j(i)} = r_j y_j^{h(r,PGID)} \cdot \tilde{r}^{-1} \cdot \prod_{k=1}^{t-1} (c_{j,k})^{i^k} (\bmod p) \qquad (13)$$

holds. If all $f_j(i)'s$ (for $1 \le j \le n$ and $j \ne i$) are verified, then U_{P_i} computes $x_i' = \sum_{j=1}^{n} f_j(i) \bmod q$ as his/her proxy share. Let $f(x) = \sum_{j=1}^{n} f_j(x) \bmod q$. The proxy share can be rewritten as $x_i' = f(i)$ and will be used for generating proxy signatures. The shared secret is regarded as

$$f(0) = \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} x_i h(r, PGID) \bmod q \qquad (14)$$

Proxy signature generation

Without loss of generality, let $D_P = \{U_{P_1}, U_{P_2}, ..., U_{P_t}\}$ be t proxy signers who want to cooperatively generate a proxy signature.

1) Each participant proxy signer U_{P_i} performs a (t,t) -VSS scheme by randomly selecting a $(t-1)$ -degree polynomial $f_i'(x) = \sum_{j=0}^{t-1} a'_{i,j} x^j$ and broadcasts $c'_{i,j} = g^{a'_{i,j}} \bmod p$ for $j = 0,1,2,...,t-1$. Furthermore, U_{P_i} calculates $f_i'(j)$ and sends it to U_{P_j} via a secure channel for $j = 1,2,...,t$ and $j \ne i$. The validity of $f_i'(j)$ can be verified by checking the equality $g^{f_i'(j)} = \prod_{k=0}^{t-1} (c'_{i,k})^{j^k} \bmod p$. Moreover, each participant proxy signer U_{P_j} can get

$$x_i'' = f'(i) = \sum_{j=1}^{t} f_j'(i) \bmod q \qquad ,$$

where $f'(x) = \sum_{j=1}^{t} f_j'(x) \bmod q$, and $Y = \prod_{k=1}^{t} c'_{k,0} \bmod p$.

2) Each participant proxy signer U_{P_j} computes and broadcasts

$$T_i = (x_i' + \tilde{s}_i) h(M) + x_i'' Y \bmod q \qquad (15)$$

3) T_i can be verified by checking the following equality:

$$g^{T_i} = \left(\prod_{j=1}^{n} r_j \left(y_o \prod_{j=1}^{n} y_j \right)^{h(r,PGID)} \left(\prod_{j=1}^{t-1} \left(c_i'' \prod_{k=1}^{n} c_{k,j} \right) \right)^{i^j} \right)^{h(M)} \cdot \left(Y \prod_{j=1}^{t-1} \prod_{k=1}^{t} c'_{k,j}{}^{i^j} \right)^{Y} (\bmod p) \qquad (16)$$

If all $T_j's$ are verified, then each U_{P_j} computes

$$T = (f(0) + \tilde{s}) h(M) + f'(0) Y \bmod q \qquad (17)$$

by applying Lagrange interpolating polynomial [15]. As a result, the proxy signature of *M* is $(r, PGID, Y, T)$.

Proxy Signature Verification

The verification of the proxy signature of *M* with respect to the proxy group G_P and the original signer is

$$g^T = \left(\left(y_o \prod_{i=1}^{n} y_i \right)^{h(r,PGID)} r \right)^{h(M)} Y^Y (\bmod p) \qquad (18)$$

6. SECURITY ANALYSIS OF HSU ET AL'S SECOND SCHEME

By our observation, Hsu et al's second scheme can't resist the public key substitution from the original signer or any proxy signer. We state it as follows.

1) Public Key Substitution Attack From The Original Signer

A malicious original signer U_o at random selects three numbers $k_1 \in Z_q^*$, $k_2 \in Z_q^*$ and $k_3 \in Z_q^*$, then computes

$$y_o' = \left(\prod_{i=1}^{n} y_i^{-1} \right) \cdot g^{k_1} \bmod p \text{ ,}$$

$r = g^{k_2} \bmod p$, $Y = g^{k_3} \bmod p$ and

$$T = (k_1 h(r, PGID) + k_2) h(M) + k_3 Y (\bmod q) \qquad (19)$$

The original signer U_o publishes y_o' as his/her new public key, and forges a valid proxy signature $(r, PGID, Y, T)$ as

$$\left(\left(y_o'\prod_{i=1}^{n}y_i\right)^{h(r,PGID)} r\right)^{h(M)} Y^{Y}(\mathrm{mod}\ p)$$
$$=\left(\left(y_o'\prod_{i=1}^{n}y_i\right)^{h(r,PGID)} r\right)^{h(M)} Y^{Y}(\mathrm{mod}\ p)$$
$$=\left(\left(\left(\prod_{i=1}^{n}y_i^{-1}\right)\cdot g^{k_1}\prod_{i=1}^{n}y_i\right)^{h(r,PGID)} g^{k_2}\right)^{h(M)} g^{k_3 Y}(\mathrm{mod}\ p)$$
$$= g^{(k_1 h(r,PGID)+k_2)h(M)+k_3 Y)}(\mathrm{mod}\ p)$$
$$= g^{T}(\mathrm{mod}\ p)$$

2) **Public Key Substitution Attack From Any Proxy Signer**

A malicious proxy signer U_{P_i} at random selects three numbers $k_1 \in Z_q^{\ *}$, $k_2 \in Z_q^{\ *}$ and $k_3 \in Z_q^{\ *}$, then computes

$y_i' = y_o^{-1}\left(\prod_{j=1, j\neq i}^{n} y_j^{-1}\right)\cdot g^{k_1} \bmod p$,

$r = g^{k_2} \bmod p$, $Y = g^{k_3} \bmod p$ and

$$T = (k_1 h(r,PGID)+k_2)h(M)+k_3 Y(\mathrm{mod}\, q) \qquad (20)$$

The proxy signer U_{P_i} publishes y_i' as his/her new public key, and forges a valid proxy signature $(r, PGID, Y, T)$ as

$$\left(\left(y_o(\prod_{j=1, j\neq i}^{n} y_j)y_i'\right)^{h(r,PGID)} r\right)^{h(M)} Y^{Y}(\mathrm{mod}\ p)$$
$$=\left(\left(y_o(\prod_{j=1, j\neq i}^{n} y_i)\cdot y_o^{-1}\left(\prod_{j=1, j\neq i}^{n} y_j^{-1}\right)\cdot g^{k_1}\right)^{h(r,PGID)} g^{k_2}\right)^{h(M)} \cdot$$
$$g^{k_3 Y}(\mathrm{mod}\ p)$$
$$= g^{(k_1 h(r,PGID)+k_2)h(M)+k_3 Y)}(\mathrm{mod}\ p)$$
$$= g^{T}(\mathrm{mod}\ p)$$

In the following, based on Hsu et al's second scheme, an improved version is proposed.

7. THE IMPROVED THRESHOLD PROXY SIGNATURE SCHEME WITH UNKNOWN SIGNERS

Proxy Share Generation

1) The original signer U_o at random selects $\tilde{k} \in Z_q^{\ *}$, computes and broadcasts $\tilde{r} = g^{\tilde{k}} \bmod p$.

2) Upon receiving $\tilde{r}$, each proxy signer $U_{P_i} \in G_P$ randomly selects $\alpha_i \in Z_q^{\ *}$, computes and broadcasts $r_i = g^{\alpha_i}\cdot\tilde{r} \bmod p$.

3) Once collecting all $r_i's$ from $U_{P_i} \in G_P$, U_o computes $\tilde{s} = x_o y_o h(r,PGID) + n\tilde{k} \bmod q$, where $r = \prod_{i=1}^{n} r_i \bmod p$

And performs a (t,n) verifiable threshold secret sharing scheme [16], denoted as (t,n) -VSS scheme to share $\tilde{s}$ among n proxy signers in G_P. That is, U_o chooses a $(t-1)$ -degree polynomial $f''(x) = \tilde{s} + a_1''x + a_2''x^2 + ... + a_{t-1}''x^{t-1} \bmod q$ and sends $s_i' = f''(i)$ to U_{P_i} for $i = 1,2,...,n$ secretly. Meanwhile, U_o publishes $c_i'' = g^{a_i''} \bmod p$ for $i = 1,2,...,t-1$.

4) Upon receiving the $s_i' = f''(i)$ from U_o, each $U_{P_i} \in G_P$ can confirm it by checking that

$$g^{\tilde{s}_i} = y_o^{\,y_o h(r,PGID)}\cdot\tilde{r}^{\,n}\prod_{j=1}^{t-1}(c_j'')^{i^j}(\mathrm{mod}\, p) \qquad (21)$$

If it holds, each $U_{P_i} \in G_P$ performs a (t,n) -VSS and acts as a dealer to distribute proxy sub-shares to other $n-1$ proxy signers for generating their valid proxy shares. That is, each $U_{P_i} \in G_P$ chooses a $(t-1)$ -degree polynomial

$$f_i(x) = \alpha_i + x_i y_i h(r,PGID) + a_{i,1}x + a_{i,2}x^2 + ... + a_{i,t-1}x^{t-1}(\mathrm{mod}\, q) \qquad (22)$$

Then $U_{P_i} \in G_P$ sends the proxy sub-share $f_i(j)$ to proxy signer U_{P_j} (for $1 \le j \le n$ and $j \neq i$) via a secure channel and broadcasts $c_{i,k} = g^{a_{i,k}} \bmod p$ for $k = 1,2,...,t-1$. To ensure the validity of $f_j(i)$ sent from U_{P_j}, U_{P_I} can check whether the equality

$$g^{f_j(i)} = r_j y_j^{\,y_j h(r,PGID)}\cdot\prod_{k=1}^{t-1}(c_{j,k})^{i^k}(\mathrm{mod}\, p) \qquad (23)$$

holds. If all $f_j(i)'s$ (for $1 \le j \le n$ and $j \neq i$) are verified, then U_{P_I} computes $x_i' = \sum_{j=1}^{n} f_j(i) \bmod q$ as his/her proxy share. Let $f(x) = \sum_{j=1}^{n} f_j(x) \bmod q$. The proxy share can be rewritten as $x_i' = f(i)$ and will be used for generating proxy signatures. The shared secret is regarded as

$$f(0) = \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n} x_i y_i h(r,PGID) \bmod q \qquad (24)$$

Proxy Signature Generation

Without loss of generality, let $D_P = \{U_{P_1}, U_{P_2}, ..., U_{P_t}\}$ be t proxy signers who want to cooperatively generate a proxy signature.

1) Each participant proxy signer U_{P_i} performs a (t,t) -VSS scheme by randomly selecting a $(t-1)$ -degree polynomial $f_i'(x) = \sum_{j=0}^{t-1} a'_{i,j}\, x^j$ and broadcasts $c'_{i,j} = g^{a'_{i,j}} \bmod p$ for $j = 0,1,2,...,t-1$. Furthermore, U_{P_i} calculates $f_i'(j)$ and sends it to U_{P_j} via a secure channel for $j = 1,2,...,t$ and $j \neq i$. The validity of $f_i'(j)$ can be verified by checking the equality $g^{f_i'(j)} = \prod_{k=0}^{t-1}(c'_{i,k})^{j^k} \bmod p$. Moreover, each participant proxy signer U_{P_j} can get

$$x_i'' = f'(i) = \sum_{j=1}^{t} f_j'(i) \bmod q \qquad ,$$

where $f'(x) = \sum_{j=1}^{t} f_j'(x) \bmod q$, and $Y = \prod_{k=1}^{t} c'_{k,0} \bmod p$.

2) Each participant proxy signer U_{P_j} computes and broadcasts

$$T_i = (x_i' + \tilde{s}_i)h(M) + x_i''Y \bmod q \qquad (25)$$

3) T_i can be verified by checking the following equality:

$$g^{T_i} = \left(\prod_{j=1}^{n} r_j \left(y_o{}^{y_o} \prod_{j=1}^{n} y_j{}^{y_j} \right)^{h(r,PGID)} \left(\prod_{j=1}^{t-1} \left(c_i{}'' \prod_{k=1}^{n} c_{k,j} \right) \right)^{i^j} \right)^{h(M)} \cdot \left(Y \prod_{j=1}^{t-1} \prod_{k=1}^{t} c'_{k,j}{}^{i^j} \right)^{Y} (\text{mod } p) \qquad (26)$$

If all T_j's are verified, then each U_{P_j} computes

$$T = (f(0) + \tilde{s})h(M) + f'(0)Y \bmod q \qquad (27)$$

by applying Lagrange interpolating polynomial [15]. As a result, the proxy signature of M is $(r, PGID, Y, T)$.

Proxy Signature Verification

The verification of the proxy signature of M with respect to the proxy group G_P and the original signer is

$$g^{T} = \left(\left(y_o{}^{y_o} \prod_{i=1}^{n} y_i{}^{y_i} \right)^{h(r,PGID)} r \right)^{h(M)} Y^{Y} (\text{mod } p) \qquad (28)$$

As to the security analysis and discussion of the improved version, for the space limitation, we will describe them in another paper.

8. REMARKS ON THE IMPROVED SCHEME

1) **Remark 1**

The improved scheme don't use the CA/SA, so when the proxy signature is verified, the verifier will obtain proxy signers' public keys from the proxy signers. The new scheme can resist the public key substitution attack from the original signer or proxy signers. And after the verifier gets the original signer or proxy signers' public keys, it is unnecessary for the verifier to check if the public key owners have the corresponding private key, as the verification congruence shows that they are confronted with solving difficult discrete logarithm questions.

2) **Remark 2**

The new scheme is one with unknown signers. That is to say, the verifier can't know who actual proxy signers are. This kind of threshold proxy signature scheme is useful in case that the verifier need only know that the proxy signature is signed by *t* proxy signers on behalf of a group of *n* proxy signers and want not to know who are the actual proxy signers.

3) **Remark 3**

The computation complexity of the new scheme is a little more than that of Hsu et al's second scheme. In the new scheme, because the verifier need not confirm whether public key owners have corresponding private keys, the communication cost is small.

9. CONCLUSIONS

We review Hsu et al's scheme, describe Tsai et al's attack to it and point out that their attack can't work. Also we review Hsu et al's another scheme with unknown signers, analyze its security and point out that it can not resist the public key substitution attack. Based on Hsu et al's second scheme, an improved version which can resist the weakness is proposed.

ACKNOWLEDGMENT

The author would like to thank anonymous referees for their suggestions to improve the paper. This letter is supported by the National Natural Science Foundation of China under Grant No. 60072018, the National Science Fund for Distinguished Young Scholars under Grant No. 60225007 and the National Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20020248024.

10. REFERENCES

[1] M. Mambo, K. Usuda, E. Okamoto, "Proxy signature for delegating signing operation", in Proceedings of the 3.th ACM Conference on Computer and Communications Security, New Dehli, India, ACM Press, New York, 1996, pp. 48-57.

[2] Li Jiguo, Cao Zhenfu, "Improvement of a threshold proxy signature scheme", Journal of Computer Research and Development, Vol. 39, Nov. 2002, pp. 515-518 (in Chinese)

[3] Li Jiguo, Cao Zhenfu, Zhang Yichen, "Improvement of M-U-O and K-P-W proxy signature schemes", Journal of 13, April 2003 pp. 1-5 (in Chinese)

[4] C.L Hsu, T.S. Wu and T.C. Wu, "New nonrepudiable threshold proxy signature scheme with known signers", *The Journal of Systems and Software*, Vol. 58, 2001, pp.119~124.

[5] M.S. Hwang, I.C. Lin and J.L. Lu Eric, "A secure nonrepudiable threshold proxy signature scheme with known signers", International Journal of Informatica, Vol. 11, Feb. 2000, pp.1-8.

[6] S.J Hwang and C.C. Chen, "Cryptanalysis of nonrepudiable threshold proxy signature scheme with known signers", INFORMATICA, Vol.14**,** Feb. 2003, pp.205-212.

[7] H.M. Sun, "An efficient nonrepudiable threshold proxy signature scheme with known signers", Computer Communications, Vol.22, Aug. 1999, pp.717-722.

[8] H.M. Sun, N.Y. Lee and T. Hwang, "Threshold Proxy Signature", IEEE Proceedings-computers & Digital Techniques, Vol.146, May 1999, pp.259-263.

[9] K. Zhang, "Threshold proxy signature schemes" Information Security Workshop, Japan, 1997, pp.191-197.

[10] C.L. Hsu, T.S. Wu and T.C. Wu, "Improvement of threshold proxy signature scheme", Applied Mathematics and Computation, Vol.136, 2003, pp.315-321.

[11] C.S. Tsai, S.F. Tzeng and M.S. Hwang, "Improved Non-Repudiable Threshold proxy signature scheme with known signers", INFORMATICA, Vol.14, March 2003, pp.393-402.

[12] Xue Qingshui, Cao Zhenfu, "On two nonrepudiable threshold proxy signature schemes with known signers", INFORMATICA, to be published.

[13] D.E.R. Denning, "Cryptography and Data Security", Addison-Wesley, Reading, MA, 1983.

[14] T. Pedersen, Distributed provers with applications to undeniable signatures .New York: Springer-Verlag, vol.547, 1991.Harbin Institute of Technology (New Series), Vol. 9, Feb. 2002, pp. 145-148.

[15] Li Jiguo, Cao Zhenfu, Zhang Yichen, "Nonrepudiable proxy multi-signature scheme", Journal of Computer Science and Technology, Vol. 18, May 2003, pp. 399-402.

[16] Li Jiguo, Cao Zhenfu, Zhang Yichen, Li Jianzhong, "Cryptographic analysis and modification of proxy multi-signature scheme", High Technology Letters, Vol.

Bilinear Pairings-based Threshold Proxy Signature Schemes with Known Signers

Xue Qingshui Cao Zhenfu Qian Haifeng
Dept. of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200030, China
Email: {xue-qsh, zfcao}@cs.sjtu.edu.cn, ares@sjtu.edu.cn **Tel.:** 86-021-62932951

ABSTRACT

So far, all threshold proxy signature schemes are based on discrete logarithm problems in the modular multiplicative group of a large prime. The kind of threshold proxy signature scheme becomes more and more complex and cost more and more computation. In the paper, we propose a bilinear pairings-based threshold proxy signature scheme with known signers and its security is analyzed and discussed. The scheme can provide the properties of nonrepudiation, unforgeability, identifiability, distinguishability, verifiability, prevention of misuse of proxy signing right, etc. Furthermore, we show that the proposed scheme is more efficient than Sun's and Hsu et al's scheme in terms of computational complexities and communication costs in some cases.

Keywords: Cryptography, Digital Signatures, Proxy Signature, Threshold Proxy Signature, Bilinear Pairings

1. INTRODUCTION

The proxy signature scheme [1], a variation of ordinary digital signature schemes, enables a proxy signer to sign messages on behalf of the original signer. Proxy signature schemes are very useful in many applications such as electronics transaction and mobile agent environment.

A proxy signature scheme comprises three entities: original signer, proxy signer and verifier. Mambo et al. [1] provided three levels of delegation in proxy signature: full delegation, partial delegation and delegation by warrant. In full delegation, the original signer gives its private key to the proxy signer. In partial delegation, the original signer produces a proxy signature key from its private key and gives it to the proxy signer. The proxy signer uses the proxy key to sign. As far as delegation by warrant is concerned, warrant is a certificate composed of a message part and a public signature key. The proxy signer gets the warrant from the original signer and uses the corresponding private key to sign. So far, a lots of proxy signature scheme have been proposed [2-14].

Afterwards, threshold proxy signature schemes were proposed [2, 6, 7-14]. In the threshold proxy signature scheme, a group of *n* proxy signers share the secret proxy signature key. To produce a valid proxy signature on the message *m*, individual proxy signers produce their partial signatures on that message, and then combine them into a full proxy signature on *m*. In a (t,n) threshold proxy signature scheme, the original signer authorizes a proxy group with *n* proxy members. Only the cooperation of *t* or more proxy members is allowed to generate the proxy signature. Threshold signatures are motivated both by the demand which arises in some organizations to have a group of employees agree on a given message or document before signing, and by the need to protect signature keys from the attack of internal and external adversaries.

In 1999, Sun proposed a threshold proxy signature scheme with known signers [9]. Then Hwang et al. [7] pointed out that Sun's scheme was insecure against collusion attack. By the collusion, any $t-1$ proxy signers among the *t* proxy signers can cooperatively obtain the secret key of the remainder one. They also proposed an improved scheme which can guard against the collusion attack. After that, [6] showed that Sun's scheme was also insecure against the conspiracy attack. In their attack, *t* malicious proxy signers can impersonate some other proxy signers to generate valid proxy signatures. To resist the attack, they also proposed a scheme. [8] shows that the scheme in [7] was also insecure against the attack by the cooperation of one malicious proxy signer and the original signer. In 2002, Li et al. [2] proposed a threshold proxy signature scheme with lots of good performances. In [14], we pointed out there were some errors in Sun's and Hwang et al.'s scheme and also proposed an improved version.

The bilinear pairings, namely the Weil pairing and the Tate pairing of algebraic curves, are important tools for study on algebraic geometry. Their usage in cryptography goes back to Victor Miller's [17] unpublished paper in 1986, and particularly the results of Menezes-Okamoto-Vanstone [18] and Frey-Ruck [19].

Currently, all threshold proxy signature schemes are based on discrete logarithm problems in the multiplicative group Z_p^* where p is a large prime. There are always some weaknesses in all proposed schemes. To resist the found weakness, the threshold proxy signature scheme based on the discrete logarithm problems in the multiplicative group Z_p^* become more and more complex and cost more and more computation. As far as I know, now there are no threshold proxy signature schemes based on bilinear pairings. In the paper, we propose a bilinear pairings-based threshold proxy signature scheme.

In the paper, we will first introduce some related work about the bilinear pairings, and then state the proposed threshold proxy signature scheme based on bilinear pairings. Next we analyze the security of the proposed scheme. After that, we also showed that the proposed scheme is more efficient than Sun's and Hsu et al's scheme in terms of computational complexities and communication costs in some cases. Finally, we draw the conclusion.

2. THE RELATED WORK

Bilinear Pairings

Let G_1 and G_2 be additive and multiplicative groups with the same prime order *q*, respectively. *P* is a generator of G_1 suppose that it is hard to solve the discrete logarithm problems in both G_1 and G_2. $e: G_1 \times G_1 \to G_2$ is a pairing which satisfies the following three properties:

1) Bilinear: $e(aP, bP') = e(P, P')^{ab}$ for all $P, P' \in G_1$ and all $a, b \in Z$.

2) Non-degenerate: If $e(aP, bP') = 1, \forall P' \in G_1$, then $P = O$.

3) Computable: There is an efficient algorithm to compute $e(aP, bP')$ for any $P, P' \in G_1$.

We can use the Weil pairing or revised Tate pairing associated with supersinglar elliptic curves to construct the bilinear pairing. With such a group G_1, we define the hard cryptographic problems in the following:

1) Discrete Logarithm (**DL**) Problem: Given $P, P' \in G_1$, find an integer *n* such that $P = nP'$ whenever such integer exists.

2) Computational Diffie-hellman (**CDH**) Problem: Given a triple $(P, aP, bP) \in G_1^3$, for $a, b \in Z_q^*$, find the element abP.

3) Decision Diffie-hellman (**DDH**) Problem: Given a quaternion $(P, aP, bP, abP) \in G_1^4$, for $a, b, c \in Z_q^*$, decide whether $c = ab (\bmod q)$ or not.

4) Gap Diffie-hellman (**GDH**) Problem: A class of problems where the **CDH** problem is hard but the **DDH** problem is easy.

Groups where the **CDH** problem is hard but the **DDH** problem is easy are called Gap Diffie-hellman (**GDH**) groups [15].

A GDH group-based signature scheme

Now we review the **GDH** group-based signature scheme [16]. Let G_1 be a **GDH** group, $[\{0,1\}^* \to G_1^*]$ be a hash function family, each member of which maps arbitrary long strings to group G_1^* and *H* be a random member of the family. *M* is the massage to be signed. The **GDH** group-based signature scheme is as follows:

Step 1. The signer chooses $x \in Z_q^*$ and computes $Y = xP$. Then the signer publishes *Y* as his/her public key and retains *x* as his/her private key.

Step 2. The signer computes $\sigma = xH(M)$. The signature of *M* is (M, σ).

Step 3. The verifier or receiver validates the (M, σ) by checking $e(P, \sigma) = e(Y, H(M))$. If it holds, the verifier/receiver accepts it, or he/she rejects it.

The authors of [15] stated and proved the following theorem:

Theorem 1. The above **GDH** group-based signature scheme is secure in the random oracle model.

The proof of **Theorem 1** is omitted here.

3. THE PROPOSED BILINEAR PAIRINGS-BASED THRESHOLD PROXY SIGNATURE SCHEME WITH KNOWN SIGNERS

In the scheme, the original signer has a private key $x_o \in Z_q^*$ and a corresponding public key $Y_O = x_o P$ which is certified by **CA** (Certification Authority). $\{P_1, P_2, ..., P_n\}$ are *n* proxy signers. Each proxy signer $P_i (i = 1,2,...,n)$ has a private key $x_i \in Z_q^*$ and a corresponding public key $Y_i = x_i P$ which is certificated by **CA** as well. m_w is a warrant which records the identities of the original signer and the proxy signers of the proxy group, parameters *t* and *n*, the valid delegation time, etc. *ASID* (Actual Signers' **ID**) denotes the identities of the actual signers. Group G_0 and G_1 have prime order *q*. *P* is a generator of **GDH** group G_0. $e: G_0 \times G_0 \to G_1$ is a secure bilinear pairing. In addition, $H_1: \{0,1\}^* \times G_0 \to Z_q^*$ and $H_2: \{0,1\}^* \to G_0 \setminus \{1\}$ are two hash functions.

The proposed scheme includes three phases: proxy share generation phase, proxy signature generation without revealing shares phase and proxy signature verification phase.

Proxy share generation phases

1) The original signer selects a random number $r \in Z_q^*$ and computes $U = rP$, $h = H_1(m_w, U)$, $Q = H_2(m_w)$, $V = (r + hx_o)Q$, $\sigma = (U, V)$ and $s = n^{-1}(hr + x_o)$.Then the original signer sends (m_w, σ, s) to each proxy signer.

2) Each proxy signer P_i confirms the validity of (m_w, σ, s) by checking whether the following congruences hold.

$$e(P, \sigma) = e(U + hY_O, H_2(m_w)) \quad (1)$$

$$nsP = hU + Y_O \quad (2)$$

If they hold, the proxy signer P_i will accept (m_w, σ, s) and calculates $s_i = s + x_i + k_i$, where $k_i \in Z_q^*$ is selected at random by the proxy signer P_i.

3) Each proxy signer P_i chooses at random a $(t-1)-$degree polynomial

$$f_i(x) = s_i + a_{i,1}x + a_{i,2}x^2 + ... + a_{i,t-1}x^{t-1}.$$

Here $a_{i,0} = s_i$. Thus $f_i(0) = s_i$. Then P_i computes and broadcasts $a_{i,j}P$ for $j = 1,2,...,t-1$ and $k_i P$. In addition, P_i computes and sends $f_i(j)$ for $j = 1,2,...,n; j \neq i$ secretly to other $(n-1)$ proxy signers.

4) Upon receiving $f_j(i)$ from P_j, for $j = 1,2,...,n, j \neq i$, P_i verifies $f_j(i)$ by checking

$$f_j(i)P = \sum_{k=0}^{t-1} i^k \cdot a_{j,k}P, \quad (3)$$

where $a_{j,0}P = n^{-1}hU + n^{-1}Y_o + Y_j + k_j P$.

If it holds, P_i computes $x_i' = \sum_{k=1}^{n} f_k(i)$ as the secret proxy share and computes $Y_i' = x_i'P$ as the public proxy share. Let $f(x) = \sum_{k=1}^{n} f_k(x)$, then we can rewrite $x_i' = \sum_{k=1}^{n} f_k(i)$ as $x_i' = f(i)$ and $Y_i' = x_i'P$ as $Y_i' = f(i)P$.

Proxy signature generation phase without revealing shares phase

Let *M* be a message to be signed. Without loss of generality, assume that $P_1, P_2, ..., P_t$ are the *t* proxy signers who want to cooperate to generate the proxy signature on behalf of the original signer.

1) Each proxy signer P_i for $i = 1,2,...,t$ computes $w_i = \prod_{j=1, j \neq i}^{t} \frac{j}{j-i}$ and $\sigma_i = (x_i'w_i + x_i)H_2(M)$. Thus $P_i's$ partial signature on the message *M* is σ_i. P_i sends σ_i to the clerk who is responsible for collecting partial signature and generates the threshold proxy signature.

2) After the clerk receives σ_i from P_i, the clerk will validate σ_i by checking

$$e(P, \sigma_i) = e(w_i Y'_i + Y_i, H_2(M)) . \qquad (4)$$

If all $\sigma_i's$ for $i = 1,2,...,t$ are valid, the clerk computes $\sigma' = \sum_{i=1}^{t} \sigma_i$. Thus the clerk sends the threshold proxy signature $(M, m_w, ASID, U, \sigma', k_1P, ..., k_nP)$ to the verifier or users.

Proxy signature verification phase

Receiving the threshold proxy signature $(M, m_w, ASID, U, \sigma', k_1P, ..., k_nP)$ of *M*, any verifier can confirm the validity of the proxy signature and identify the actual signers. The steps of the phase are stated as follows:

1) By M_w and *ASID*, the verifier can identify the original signer and the proxy signers, and get the public keys from the **CA**. Besides, he/she can also identify the actual proxy signers.

2) The verifier validates the proxy signature

$$(M, m_w, ASID, U, \sigma', k_1P, ..., k_nP)$$

by checking

$$e(P, \sigma') = e(H_1(m_w, U)U + Y_O + \sum_{i=1}^{n} Y_i + \sum_{i=1}^{t} Y_i + \sum_{i=1}^{n} k_i P, H_2(M)) \qquad (5)$$

If it holds, the proxy signature

$$(M, m_w, ASID, U, \sigma', k_1P, ..., k_nP)$$

is valid. The verification of the proxy signature

$$(M, m_w, ASID, U, \sigma', k_1P, ..., k_nP)$$

is justified by the following equation:

$$\begin{aligned}
e(P, \sigma') &= e(P, \sum_{i=1}^{t} \sigma_i) \\
&= e(P, \sum_{i=1}^{t} (x_i'w_i + x_i)H_2(M)) \\
&= e(P, (\sum_{i=1}^{t} (x_i'w_i) + \sum_{i=1}^{t} x_i)H_2(M)) \\
&= e(P, (f(0) + \sum_{i=1}^{t} x_i)H_2(M)) \\
&= e(P, (\sum_{i=1}^{n} f_i(0) + \sum_{i=1}^{t} x_i)H_2(M)) \\
&= e(P, (\sum_{i=1}^{n} s_i + \sum_{i=1}^{t} x_i)H_2(M)) \\
&= e(P, (ns + \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} k_i + \sum_{i=1}^{t} x_i)H_2(M)) \\
&= e(P, (H_1(m_w, U)r + x_o + \sum_{i=1}^{n} x_i + \sum_{i=1}^{t} x_i + \sum_{i=1}^{n} k_i)H_2(M)) \\
&= e((H_1(m_w, U)r + x_o + \sum_{i=1}^{n} x_i + \sum_{i=1}^{t} x_i + \sum_{i=1}^{n} k_i)P, H_2(M)) \\
&= e(H_1(m_w, U)U + Y_O + \sum_{i=1}^{n} Y_i + \sum_{i=1}^{t} Y_i + \sum_{i=1}^{n} k_i P, H_2(M))
\end{aligned} \qquad (6)$$

4. SECURITY ANALYSIS OF THE PROPOSED SCHEME

In the following, we will prove that the proposed scheme can resist all kinds of known attack including the forgery attack, conspiracy attack, public key substitution attack etc.

First, the proposed scheme can resist the forgery attack from the original signer. In Eq (7),

$$\begin{aligned}
\sigma' = \sum_{i=1}^{t} \sigma_i &= \sum_{i=1}^{t} (x_i'w_i + x_i)H_2(M) \\
&= (\sum_{i=1}^{t} (x_i'w_i) + \sum_{i=1}^{t} x_i)H_2(M) \\
&= (f(0) + \sum_{i=1}^{t} x_i)H_2(M) \\
&= (\sum_{i=1}^{n} f_i(0) + \sum_{i=1}^{t} x_i)H_2(M) \\
&= (\sum_{i=1}^{n} s_i + \sum_{i=1}^{t} x_i)H_2(M) \\
&= (ns + \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} k_i + \sum_{i=1}^{t} x_i)H_2(M) \\
&= (H_1(m_w, U)r + x_o + \sum_{i=1}^{n} x_i + \sum_{i=1}^{t} x_i + \sum_{i=1}^{n} k_i)H_2(M)
\end{aligned} \qquad (7)$$

Although the original signer has the knowledge of m_w, U and x_o , he/she has no the knowledge of each proxy signer's private key x_i for $i = 1,2,...,n$ or $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{t} x_i$. From the proxy share generation phase, it can be seen that the original signer is unable to obtain the knowledge of proxy signers' private keys either. Thus the original signer can't forge a valid proxy signature. So the proposed scheme is a proxy-protected scheme.

Second, any proxy signer can't forge valid proxy signatures. From the proxy signature

$$(M, m_w, ASID, U, \sigma', k_1P, ..., k_nP)$$

any proxy signer can't obtain the knowledge of other proxy signers' private keys $k_1, k_2, ..., k_{n-1}$ and k_n . Also, he/she can't get $\sum_{i=1}^{t} (x_i' + x_i)$ because of difficult discrete logarithm problems. Therefore, proxy signatures can’t be forged by any proxy signer.

Third, $(t-1)$ proxy signers can’t cooperate to generate valid proxy signatures. From Eq (7), only *t* proxy signers cooperate to generate proxy signatures. Because $(t-1)$ proxy signers don’t have the knowledge of $t'th$ proxy signer’s proxy share x_i' and private key x_i, and can’t get them from the known proxy signature $(M, m_w, ASID, U, \sigma', k_1P, ..., k_nP)$ because of the same cause as the above case. So proxy signatures can’t be generated by $(t-1)$ proxy signers.

Fourth, *t* proxy signers can’t impersonate other *t* proxy signers to generate valid proxy signature. From the verification Eq (5) of proxy signatures, it shows that the *t* actual proxy signers’ public keys are needed. If *t* malicious proxy signers want to forge other *t* proxy signers’ signature, they need have the knowledge of other *t* proxy signers’ secret proxy share $x_i's$, private keys x_is or $\sum_{i=1}^{t}(x_i'+x_i)$. From the above case, we know that it is impossible to get them.

Fifth, the scheme can resist the public key substitution attack from the original signer or any proxy signer. In the scheme, **CA** is need. If the original signer or any proxy signer wants to substitute a new public key for the original public key, he/she must have the knowledge of the corresponding private key. In the public key substitution attack, generally speaking, the attacker doesn’t have the knowledge of the corresponding private key. Thus, the attacker can’t change its public key in the system public directory which is managed by **CA**. So the public key substitution attack doesn’t work.

Sixth, the original signer and any proxy signer can’t cooperate to get the valid proxy signature. The reason is the same as that of the first and second case.

Seventh, the original signer and $(t-1)$ proxy signers can’t cooperate to generate valid proxy signatures. The cause is the same as that of the first and third case.

Eighth, the third party can’t forge valid proxy signatures. In the scheme, the third party has least knowledge of the proxy signers’ proxy share and private key than the original signer and $(t-1)$ proxy signers. Because the original signer and $(t-1)$ proxy signers can’t cooperate to generate valid proxy signatures, the third party can’t generate valid proxy signatures either.

Ninth, the clerk can’t forge valid proxy signatures. From the partial signature σ_i, the clerk can’t get the knowledge of $(x_i'w_i + x_i)$ because of difficult discrete logarithm problems. Of course, the clerk is unable to obtain the knowledge of x_i' or x_i either. From the equation $\sigma' = \sum_{i=1}^{t}\sigma_i$, the clerk can’t get $\sum_{i=1}^{t}(x_i'w_i + x_i)$ either because of the same cause. Therefore, the proxy signature can’t be forged by the clerk. In common cases, one proxy signer or the third party will be regarded as the clerk. All in all, from the security analysis of the proposed scheme, the scheme provides the properties of unforgeability, nonrepudiation, distinguishability, identifiability, verifiability and prevention of misuse of proxy signing right.

5. PERFORMANCE EVALUATION OF THE PROPOSED SCHEME

In the section, the proposed scheme, Sun’s scheme and Hsu et al’s scheme are compared in terms of computational complexities and communication costs. We denote the following notations to facilitate the performance evaluation: T_h: The time for performing a one-way hash function *h* or H_1 in the paper. T_H: The time for performing a one-way hash functions H_2 in the paper. $T_{\exp}$: The time for performing a modular exponentiation computation. T_{mul}: The time for performing a modular multiplication computation. T_{inv}: The time for performing a modular inverse computation. $|x|$: The bit-length of an integer *x*. T_{pm}: The time for performing a point addition computation. T_{sm}: The time for performing a scalar multiplication computation. T_p: The time for performing a pair computation. NA: Not available.

The comparison of computational complexities and communication costs among Sun’s Scheme [9], Hsu et al’s scheme [6] and the proposed scheme are stated in **Table 1** and **2**, respectively. From **Table 1**, it can be seen that if we select appropriate algorithms of H_2, point addition, scalar multiplication and pair computation, the computation cost in each phase can be smaller so that the computation performance of the proposed scheme can be better than Sun’s and Hsu et al’s schemes. Also from **Table 2**, in secret share generation phase, the communication cost is least than that of Sun’s and Hsu et al’s schemes. In other three phases, if we choose appropriate *p* and *q*, the proposed scheme can be more efficient than Sun’s and Hsu et al’s schemes. Through the comparisons, we conclude that the proposed scheme outperforms Sun’s and Hsu et al’s schemes in terms of computational complexities and communication costs in some cases.

Table 1 Comparison of computational complexities [a]

	Sun’s scheme	Hsu et al’s scheme	The proposed scheme
Secret share generation [b]	$n(t-1)T_{\exp} + (n^2-1)\cdot(t-1)T_{mul}$	$(n+1)T_{\exp} + n(t-1)T_{mul}$	$2T_{sm} + T_h + T_H + 2T_{mul} + T_{inv}$
Proxy share generation	The original signer: $tT_{\exp} + (nt-n+1)T_{mul} + T_h$ Each proxy signer: $(t+1)T_{\exp} + (2t-1)T_{mul} + T_h$	The original signer: $tT_{\exp} + (nt-n+1)T_{mul} + T_h$ Each proxy signer: $(t+1)T_{\exp} + (2t-1)T_{mul} + T_h$	The original signer: NA Each proxy signer: $2T_p + 2T_{pm} + (t+2)T_{sm} + T_H + T_{mul} + (n-1)[(t-2)T_{mul} + (t-3)T_{\exp} - 3T_{pm} - 2T_{sm} + 2T_{inv}]$

Proxy signature generation [c]	Generating the individual proxy signature: $t^2T_{\exp} + (4t^2 - 7t+5)T_{mul} + T_h$ Generating the proxy signature: $(3t^2 - t - 2) \cdot T_{\exp} + (6t^2 - 7t+1)T_{mul} + (t^2 - t)T_{inv} + 2T_h$	Generating the individual proxy signature: $T_{\exp} + (3t-1) \cdot T_{mul} + (t-1)T_{inv} + T_h$ Generating the proxy signature: $(t^2+4t)T_{\exp} + (4t^2 - t - 1)T_{mul} + (t^2 - t)T_{inv} + 2T_h$	Generating the individual proxy signature: $(t-1)T_{mul} + T_H + T_{sm}$ Generating the proxy signature: $2T_p + T_{sm} + T_H + (t-1) \cdot T_{pm}$
Proxy signature verification [c]	$4T_{\exp} + (t+3)T_{mul} + 2T_h$	$4T_{\exp} + (t+3) \cdot T_{mul} + 2T_h$	$2T_p + T_h + T_{sm} + (n+t+1)T_{pm} + T_H$

[a] The comparison excludes the computation costs for validating. [b] The totally required computation costs are considered in the stage. [c] Assume that $Y_G = \prod_{i=1}^{n} y_i \bmod p$ is pre-computed.

Table 2 Comparison of communication costs [a]

	Sun's scheme	Hsu et al's scheme	The proposed scheme
Secret share generation	$n(t-1)\mid p\mid + n(n-1) \cdot \mid q\mid$	$(n+1)\mid p\mid + n\mid q\mid$	$5\mid q\mid + \mid m_w\mid$
Proxy share generation	$t\mid p\mid + n \cdot \mid q\mid + \mid m_w\mid$	$t\mid p\mid + n\mid q\mid + \mid m_w\mid$	$(n+2t-3) \cdot \mid q\mid$
Proxy signature generation	$(t^2+1)\mid p\mid + t(t-1) \cdot \mid q\mid + \mid m\mid + \mid m_w\mid$	$(t+1)\mid p\mid + t\mid q\mid + \mid m\mid + \mid m_w\mid$	$(2n+2t+2)\mid q\mid + \mid m\mid + \mid m_w\mid$
Proxy signature verification	$2\mid p\mid + \mid q\mid + \mid m\mid + \mid m_w\mid + \mid ASID\mid$	$2\mid p\mid + \mid q\mid + \mid m\mid + \mid m_w\mid + \mid ASID\mid$	$4\mid q\mid + \mid m\mid + \mid m_w\mid + \mid ASID\mid$

[a] Here assume that *m* is the message to be signed.

6. CONCLUSIONS

We have proposed a bilinear pairings-based threshold proxy signature scheme with known signers, analyzed and discussed its security. The proposed scheme can provide the properties of nonrepudiation, unforgeability, identifiability, distinguishability, verifiability, prevention of misuse of proxy signing right, etc. Furthermore, we have showed that the proposed scheme is more efficient than Sun's and Hsu et al's schemes in terms of computational complexities and communication costs in some cases either.

7. ACKNOWLEDGMENT

The author would like to thank anonymous referees for their suggestions to improve the paper. This paper is supported by the National Science Fund for Distinguished Young Scholars under Grant No. 60225007 and the National Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20020248024.

8. REFERENCES

[1] M. Mambo, K. Usuda, E. Okamoto, "Proxy Signature for Delegating Signing Operation", Proceedings of the 3.th ACM Conference on Computer and Communications Security, New Dehli, India, ACM Press, New York, 1996, pp. 48-57.

[2] Li Jiguo, Cao Zhenfu, "Improvement of a Threshold Proxy Signature Scheme", Journal of Computer Research and Development, Vol. 39, No.11, Nov. 2002, pp. 515-518 (in Chinese)

[3] Li Jiguo, Cao Zhenfu, Zhang Yichen, "Improvement of M-U-O and K-P-W Proxy Signature Schemes", Journal of Harbin Institute of Technology (New Series), Vol. 9, No.2, Feb. 2002, pp. 145-148.

[4] Li Jiguo, Cao Zhenfu, Zhang Yichen, "Nonrepudiable Proxy Multi-signature Scheme", Journal of Computer Science and Technology, Vol. 18, No.3, May 2003, pp. 399-402.

[5] Li Jiguo, Cao Zhenfu, Zhang Yichen, Li Jianzhong, "Cryptographic Analysis and Modification of Proxy Multi-signature Scheme", High Technology Letters, Vol. 13, No.4, April 2003 pp. 1-5 (in Chinese)

[6] C.L Hsu, T.S. Wu and T.C. Wu, "New Nonrepudiable Threshold Proxy Signature Scheme with Known Signers", The Journal of Systems and Software, 58, 2001, pp.119~124.

[7] M.S. Hwang, I.C. Lin and J.L. Lu Eric, "A Secure Nonrepudiable Threshold Proxy Signature Scheme with Known Signers", International Journal of Informatica, Vol. 11, No.2, Feb. 2000, pp.1-8.

[8] S.J Hwang and C.C. Chen, "Cryptanalysis of Nonrepudiable Threshold Proxy Signature Scheme with Known Signers", INFORMATICA, Vol.14, No2, Feb. 2003, pp.205-212.

[9] H.M. Sun, "An Efficient Nonrepudiable Threshold Proxy Signature Scheme with Known Signers", Computer Communications, Vol.22, No.8, Aug. 1999, pp.717-722.

[10] H.M. Sun, N.Y. Lee and T. Hwang, "Threshold Proxy Signature", IEEE Proceedings-computers & Digital Techniques, Vol.146, No.5, May 1999, pp.259-263.

[11] K. Zhang, "Threshold Proxy Signature Schemes" Information Security Workshop, Japan, 1997, pp.191-197.

[12] C.L. Hsu, T.S. Wu and T.C. Wu, "Improvement of Threshold Proxy Signature Scheme", Applied Mathematics and Computation, 136, 2003, pp.315-321.

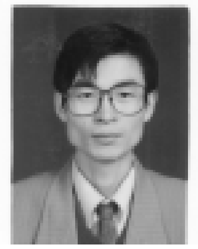
[13] C.S. Tsai, S.F. Tzeng and M.S. Hwang, "Improved Nonrepudiable Threshold Proxy Signature Scheme with Known Signers", INFORMATICA, Vol.14, No.3, March 2003, pp.393-402.

[14] Xue Qingshui, Cao Zhenfu, "On Two Nonrepudiable Threshold Proxy Signature Schemes with Known Signers", INFORMATICA, to appear.

[15] D. Boneh, B. Lynn and H. Shacham, "Short Signature from the Well Pairing", Advances in Cryptology-Asiacrypt'01, Springer Verlag, 2001.

[16] A. Boldyreva, "Threshold Signature, Multisignature and Blind Signature Scheme Based on the Gap-Diffie-Hellman-group Signature Scheme", Public Key Cryptography-PKC2003, Springer Verlag, 2003, pp. 31-46.

[17] V. Miller, “Short Programs for Functions on Curves”, unpublished manuscript, 1986.
[18] A. Menezes, T. Okamoto, and S. Vanstone, “Reducing Elliptic Curve Logarithms to Logarithms in a Finite Field”, IEEE Transaction on Information Theory, Vol.39, 1993, pp.1639-1646.
[19] G. Frey and H. Ruck, “A Remark Concerning M-Divisibility and the Discrete Logarithm in the Divisor Class Group of Curves”, Mathematics on Computation, Vol.62, 1994, pp.865-874.



Xue Qingshui was born in Shandong Province, China, in 1971. He received a BS degree in computer science and technology in Shandong Normal University and a master degree in computer application from Shandong University, China in 1995 and 2000, respectively. From 1995 to 2000, he was an assistant teacher and from 2000 to 2002, a docent, respectively, in Shandong P.E. Institute, China. He is now a doctoral candidate in the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include network security and cryptography.

Cao Zhenfu was born in Jiangsu Province, China, in 1962. He received a BS in computer science and technology from the Harbin Institute of Technology, China in 1983, and a Ph.D. in mathematics from the same university. Since 1987 and 1991, he has been the Assistant Professor and Professor respectively. Main research areas are number theory and modern cryptography, theory and technology of information security, etc. Since 2000, he has been granted to become the doctoral supervisor. In 2002, he came to Shanghai Jiao Tong University. Prof. Cao is the member of many academic organizations such as Expert consultation group of the national information area and Expert group of the national information security technology. And he is a reviewer of Mathematical Reviews (USA) and Zentrallbatt MATH (Germany). More than 20 academic research projects have been completed, and he is the head of the projects. He is the gainer of the first prize of award for science and technology in Chinese University and other 4 term awards for Heilongjiang Province and Chinese Ministry of Aeronautics and Astronautics. Up to now (1981—2002), more than 200 academic papers have been published in Journals, and was invited to give talks more than 20 times in some universities and academic institutes.

Qian Haifeng received a BS degree in mathematics in East China Normal University and a master degree in Algebraic Geometry in East China Normal University, China in 2000 and 2003, respectively, and is now a doctoral candidate in the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include network security and cryptography.

A Load Balancing Algorithm for High Speed Intrusion Detection*

Gong Jian, Lu Sheng, Rui Suying
Department of Computer Science and Engineering, Southeast University
Nanjing 210096, P.R.China
Email: jgong@njnet.edu.cn; **Tel.:** 025-83794341

ABSTRACT

Using the concept of bit entropy and bit flow entropy, a novel load-balancing algorithm named Dimension-based Classification Algorithm (DCA) is introduced in this paper, aiming at implementing NIDS in high-speed network with traffic load up to Gbps. Based on the contents of fields in IP packet header and some simple operations, this algorithm can keep the semantic relativities among packets in a high bandwidth network environment while distributing workload to different processing node. It has a fairly good load-balancing feature in both macroscopical and microscopical senses for high-speed intrusion detection. The selection of operation and operand of DCA is discussed in detailed, and their efficiency is evaluated.

Keywords: Load Balance, Bit Entropy, Packet Classification, Intrusion Detection.

1. INTRODUCTION

Nowadays, attackers can find more and more valuable targets in the Net. Network attack incidents keep happening almost all the time. So NIDS (Network-based Intrusion Detection system) has to analyze more security related audit data in a shorter instant from the packets collected from the network being monitored ever than before. On the other hand, the traffic in high bandwidth network increases from Mbps to Gbps, which causes a number of performance problems in NIDS, and makes many traditional detection methods unfeasible any more. Therefore, the conflict between performance of NIDS and the arrived mass of packets must be dealt with in high bandwidth network.

When processors' performance does not meet the requirement, using clustered architecture for load balance is a very common solution. But HIDS has some specific requirements of load distribution that should be met. That is, the context-sensitive relationship among packets must be kept and the communication among processors should be minimized. For example, some sophisticated attackers will divide their remote exploit packets into fragments; and NIDS should be able to reassemble those packets in order to detect such an action. If one wants to correlate attacks or intrusions, all the related packets must be sent to the same processing node. Those requirements definitely restrict the selection of load balancing algorithms for NIDS. Round-robin and some other traditional methods will not fit for such an application environment, and some classification-based algorithms are required.

Inspired from the concept of bit entropy, this paper proposes a new load balancing algorithm, named **Dimension-based Classification Algorithm (DCA)**, to be used in a high bandwidth network for NIDS and other applications in which the context-sensitive feature of flows need to be kept. With this algorithm, a good load balance can be achieved and the integrity of context-sensitive packets can be maintained as well.

The DCA is defined in section 2, and its parameter selections are discussed in section 3. Section4 and 5 analyze its macroscopic and microscopic load balance features, respectively. Some conclusions are given in section 6.

2. THE DEFINITION OF DCA

2.1 Classification based load balance

The most intuitive method to keep the content integrity of packets is to distribute workload of NIDS according to IP addresses because the content related packets should go to the same destination or come from the same source. Therefore, workload distribution is a typical packet classification problem.

To discuss the solution, following presentations are used in this paper.

Packet Classification (PC): PC(*O*, Л, *M*)->i, i∈{1, 2, …, m}; m is the number of processors}; where three operations are involved.

- ***Field Generate Operation** (O)*：O(F_1, F_2, …, F_n)->F
  F_1，F_2，…，F_n are *n* fields in packet *P*. *O* is an operation on F_1~F_n. F is a set of bits, which is the result of operation *O*.
- ***Classification Operation*** (Л)： Л($\mathfrak{R}$, F)->R
  $\mathfrak{R}$={R_1, R_2, …, R_m}， F is the result of operation *O*. R is a subset of ruleset $\mathfrak{R}$, which is true for F on operation Л.
- ***Mapping Operation** (M)*：M(R)->i
  |R|=1, i∈{1, 2, …, m} is required in this specific context.

Operation O generates the bits required by the detection algorithm. Operation Л finds the rules which are true for F. $\mathfrak{R}$ is a rule set because each processing node may have different detection task so that they may have different detection rule subset. Because |R|=1, Operation M will be very simple. the content integrity will be guaranteed by operation Л with the workload also balanced. One can use state to keep track of the related packet dispatching, but this is unacceptable for high-speed intrusion detection due to its

* This paper is supported by NSFC (90104031) and National 973 program (2003CB314803)

cost.

G.Cheng, et.al.[1] proposed a packet classification model which can be used for sampling network traffic in high bandwidth network. With this model, certain bits in packet are used to classify packets into different group for processing, so that the effect of load balancing is achieved. The chosen bits, e.g. Identification field in IP packet header, have a good randomness, and are suitable for sampling. The specialty of this model is that it can maintain the consistency of the packets sampled at different sampling point. That is, if one packet is sampled at a point, it will be sampled at all the points, with which packet can be distinguished and classified in high speed. This feature can be applied to IDS load balance. Instead of selecting a set of samples by specific bits value, one can just separate the set of packets into subsets by the value of a bit or a set of bits, so that these packets are grouped for load balancing while the (context-sensitive) relativity among them is reserved.

2.2 Dimension-based Classification Algorithm

With the concepts described above, the dimension-based classification algorithm (DCA) can be defined generally as following: For each incoming packet P

1) Get the values of F_1, F_2, …, F_n in P;
2) Perform the operation O on F_1, F_2, …, F_n, and gain the result field F;
3) According to the rule set $\Re$ and field F, do the classification operation ת, and gain a result rule set R;
4) Based on a predefined mapping operation *M*, map the result rule set R to a classification number *i* (i=M(R));
5) Using the classification number *i*, assign P to *ith* processor.

Notice: any well-known packet classification algorithms, e.g. given in [2], could be directly used in step 3.

From above one can see that no communication among processors is required when dealing with the incoming packet. The relationship among packets will be assured by the selection of operation O and field F_i ($1 \le i \le n$) since the context-sensitive relationship is maintained by invariant of field. For example, for the segmented IP packet, most of the packet header content will be the same; and for a TCP session, all packets will at least have same source IP address and destination IP address. Therefore, the selection of Fields F_i, Field Generate Operation O, and Rule Set $\Re$ become the key points of the algorithm.

Table 1 Truth table of two bits operations

0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
A	B	0	&		A		B	$\oplus$	|		$\overline{\oplus}$	$\overline{B}$		$\overline{A}$			1

3. SELECTION OF OPERATION AND OPERATIONS IN DCA

3.1 Bit Entropy and Bit Flow Entropy

According to information theory, bit entropy is defined as:

$$H(b) = -(p\log_2 p + (1-p)\log_2(1-p))$$

Where p is the probability of b=0, and (1-p) is the probability of b=1.H(b) is the average indeterminacy of bit occurrences and can be used as a metrics of the randomness of bit *b*. The higher the randomicity of a bit *b*, the larger its bit entropy, and vice versa.

In reality, a flow of bits is more useful than a single bit, so that the bit flow entropy can be defined as

$$H(s) = -\sum_{i=0}^{2^s-1} p_i \log_2 p_i \text{ ,}$$

where *s* is the length of a bit flow which has $n=2^s$ events all together[1], and p_0, p_1, ..., p_{n-1} are probabilities of each event.

According to the Maximal Bit Flow Entropy Theorem, the maximum of bit flow entropy is $H_{max}(s)=s$. Then the Information Efficiency E of a bit flow is defined by [1] as:

$$E = H(s)/H_{\max}(s) = H(s)/s \text{ .}$$

E is a metric of bit flow randomness, indicating the ratio between H(s) and $H_{max}(s)$.

An instance of DCA is determined by the selection of operators and operands, which could be seen as the arguments of algorithm. To choose best-fit fields in a packet when instance an algorithm, bit entropy and bit flow entropy analysis is a very useful method. Larger bit flow entropy of a field is a strong assurance of good macroscopic load balance.

3.2 Selection of Fields Fi and Field Generate Operation O

The DCA does not use the fields in packet directly, but a result of an operation on them instead. Because the result of a bit flow operation is also a bit flow (field), its Information Efficiency E could also be used to test such a virtual field.
The Maximal Bit Flow Entropy Theorem shows that the Information Efficiency E of a bit flow is satisfied if and only if all the composing bits of this bit flow have large Bit Entropy[1]. That is, good stochastic field is composed of good stochastic bits.

To find good operations and good fields for DCA, one can look at two bits operations first. There are only 16 types of operators between two bit operands shown in table 1.

[1] [1] defines the length of s as $n+1=2^s$, the bit flow from p_0 to p_n. We defines the length of s is n, the bit flow from p_0 to p_{n-1}.

If the result of the operation is all 0 or 1, the entropy of result bit is 0. It is obvious that such two operations aren't fit for DCA. There are 8 types of operations that the ratio of 0 and 1 in result is 3:1 or 1:3 ($C_4^1 + C_4^3$). For those operations, if the entropies of A and B are both large and the correlation between A and B is nearly zero, the difference of the ratio of 0 and 1 in the result could be significant, so that the entropy of result bit should not be large enough. Of course, one could try to gain a large result entropy by using two bits whose entropies are both less. For example, if both A and B have high probability of being 1, the bit entropy of the result of operation (A & B) could be much larger than that of A or B. However, it is very difficult to find the bit A and bit B that has such a strange characteristics to be used to those operations.

So, using the operations which have two 0s and two 1s in result is a more reasonable choice. There are only six (C_4^2) types of operations have such feature. For operations A and $\overline{A}$, the bit entropies are both as same as A's bit entropy, because $H(\overline{A}) = -((1-p)\log_2(1-p) + p\log_2 p) = H(A)$. For the same reason, the bit entropies of operations B and $\overline{B}$ are both as same as the bit entropy of operation *B*. Thus, there are only two types of operations with same entropy will need to be analyzed. They are exclusive OR (XOR) and NOT exclusive OR. Let the probability that result equals to 1 is *p*, p_1 is the probability of *A*=1 and p_2 is the probability of *B*=1. Suppose that there is no correlation between *A* and *B*, then $p = p_1 p_2 + (1-p_1)(1-p_2)$, and the entropy of p is:

$$H = -(1-p_1)(1-p_2)\log_2(1-p_1)$$
$$- p_1 p_2 \log_2 p_1$$
$$- (1-p_1)(1-p_2)\log_2(1-p_2)$$
$$- p_1 p_2 \log_2 p_2$$

For $p_1 \in (0,1)$ and $p_2 \in (0,1)$, the entropy of p is shown as figure 1.

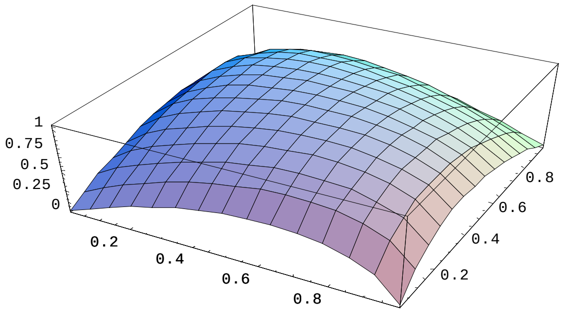


Figure 1 Bit Entropy of Two Bits Exclusive OR

The maximum point is mounted at $p_1=p_2=0.5$, where the entropy H equals to 1. Because $\frac{\partial^2 H}{\partial p_1} = \frac{p_1 + p_2 - 2p_1 p_2}{-p_1 \ln 2 + p_1^2 \ln 2}$ will less than 0 in area $p_1 \in$ (0,1) and $p_2 \in$(0,1), so that H shows as an arch structure. With the discussion above, one can draw the conclusion that if fields with large Information Efficiency are chosen and XOR or NOT XOR operations are used over them, the Information Efficiency of the result will be good. Fortunately, those fields we need to keep the relativities among packets all have good bit flow entropy and Information Efficiency, as shown in section 4.

3.3 Choose Rule Set ℜ, Classification Operation л and Mapping Operation *M*

There are less constraints about the selection of rule set ℜ, classification operation л and mapping operation *M*. If these operations do not influence the efficiency of load balancing, the simpler, the better. A simple method described below is effective. Choose $log_2 m$ (m must be an exponential of 2) bits in field F based on the number m of processors, and select the processors by those bits. For example, it would select three bits to denote 0~7 when m equals to 8. If there are m processors, ℜ={R1, R2, …, Rm}, each Ri is a sequence of bits with length $log_2 m$, as ℜ={0..0, 0..1, …, 1..1} (each rule of ℜ has $log_2 m$ bits). The mapping operation is just the EQUAL, using the value of Ri directly.

Classification operation л could be any one as long as it meets the performance and precision requirement of the problem, so it will not be discussed in this paper.
In the following discussion, only fields Fi and field generate operation *O* will be concerned.

4. MACROSCOPIC LOAD BALANCE OF DCA

To compare the load balancing performance of instances with different arguments, some metrics are needed. The macroscopic metric is just the randomness of packets, which means that if the packets dealt to the processors with similar probability, it is axiomatic that each processor will handle the packets with almost same amount in a long period. In this section, we will show that DCA can achieve good load balance in microscopy (in a long period) by choosing the arguments properly. Because the calculation of the macroscopic load balance is very simple, it would also be a good way to select the fields and field generate operation for an instance of DCA. 67,870,553 packets were analyzed that gained from CERNET backbone in one week with the static interval of dumping all packets in 2 seconds for every 75 seconds.

Only those fields that keep the relative information of the packets will be analyzed, which yields to different algorithms.

Algorithm 1: Let F1 be the identification field of IP head, field generate operation O be EQUAL. O: (=, F1)[2]

[1] had gained a conclusion that identification field of IP header has very good Information Efficiency. Figure 2 proved such declaration. This field is analyzed just because it involved in some attacks toward TCP/IP protocol implementations. Of course, the highest bit is not a fair bit that would be chosen.

Algorithm 2: Let F1 be the source IP address, F2 be destination IP address, and field generate operation O be exclusive OR (XOR). O=(^, F1, F2)
Both Source IP address and destination IP address are good

[2] O is represented by s-expression leaded by operator, same as in other algorithms' description

stochastic variables. Figure 3 shows bit entropies of source IP address, destination IP address and the result of XOR operation between them.

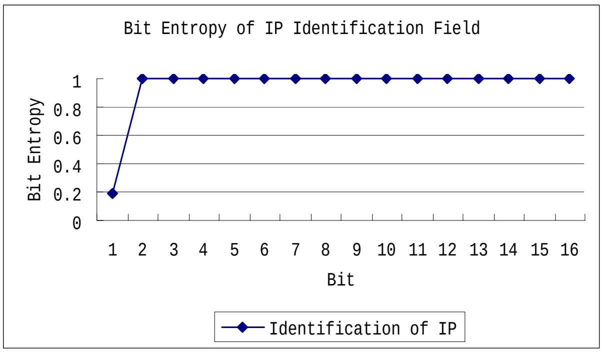


Figure 2 Bit Entropy of IP Identification Field

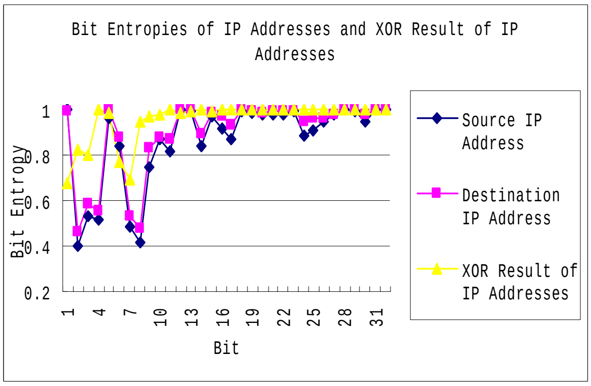


Figure 3 Bit Entropies of IP Addresses and XOR Result of IP Addresses

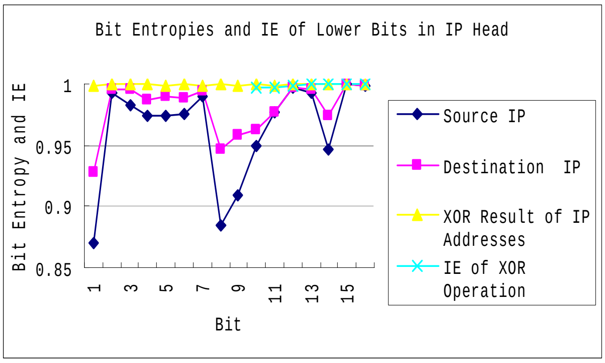


Figure 4 Bit Entropies and IE of Lower Bits in IP Head

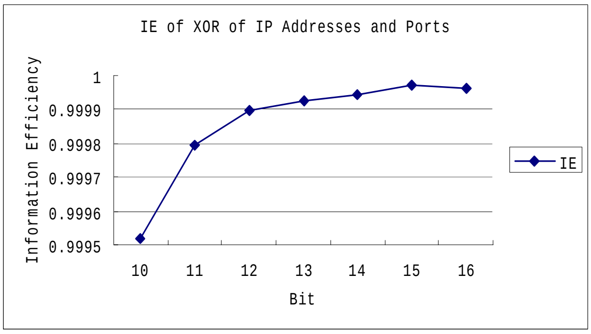


Figure 5 Information Efficiency of Algorithm 4

The entropies of lower 16 bits and the Information Efficiencies (IE) of lower 7 bits are shown in figure 4. (The Information Efficiency point at the 16th bit is the Information Efficiency of the 16th bit, The Information Efficiency point at the 15th bit is the Information Efficiency of the bit flow consisted of the 15th bit and the 16th bit, and so on)

Figure 4 illustrates that the entropy of the result is better than any other operands. It means that the XOR operation will ameliorate the randomness of bit flows.

Algorithm 3: Let F_1 be source IP address, F_2 be destination IP address, F_3 be source port, F_4 be destination port, and field generate operation O be XOR. O=(^, F_1, F_2, F_3, F_4)
The Information Efficiency of last seven bits is shown in figure 5. Its presentation is similar to figure 4. Figure 5 tells if more bits are used, the Information Efficiency will decrease, which means that increasing the number of processors will decrease the macroscopic load balance.

For all of the algorithms mentioned above, algorithm 2 will guarantee that all packets with same source IP address and destination IP address will be assigned to same processor. Algorithm 3 will promise that all packets in one TCP session will be assigned to same processor, and both of them would achieve good load balance.

5. MICROSCOPIC LOAD BALANCE OF DCA

Good bit entropy and Information Efficiency promise a good macroscopic load balance, but they cannot assure the load balance in microscopic (short time period). Many commonly used metrics in parallel computing are not suitable in this situation, for example the machine balance metric and other benchmarks [3]. Two measures are used to evaluate the algorithms defined in section 4 in microscopic. The basic definitions of these two measures are originated from [4], but the final metrics has been changed accordingly to be applicable for packet and flow.

- **$load_{i,j}$** - Load of processor i (of n processors) at the jth sampling point (of m such points)
- $peak\_load_j$ - highest load on any processors at the jth sampling point
- **pead_to_mean ratio-** $\frac{peak\_load_j}{(\sum_{i=1}^{n} load_{i,j})/n}$
- **LBM**(Load Balance Metric) -

$$\frac{\sum_{j=1}^{m} (\frac{peak\_load_j}{(\sum_{i=1}^{n} load_{i,j})/n}) \times \frac{\sum_{i=1}^{n} load_{i,j}}{n})}{\sum_{j=1}^{m}\sum_{i=1}^{n} load_{i,j}/n}$$

$$= \frac{\sum_{j=1}^{m} peak\_load_j}{(\sum_{j=1}^{m}\sum_{i=1}^{n} load_{i,j})/n}$$

The major difference between **[4]** and this paper is the definitions of load$_{i,j}$. pps$_{i,j}$(packet per second) and bps$_{i,j}$(bits per second) are defined here for different usages. Two types of LBMs are discussed, PLM (pps Load Balance Metric) and BLM (bps Load Balance Metric).

The PLM and BLM of algorithm 2 are shown in figure 6, and the PLM and BLM of algorithm 4 are shown in figure 7.

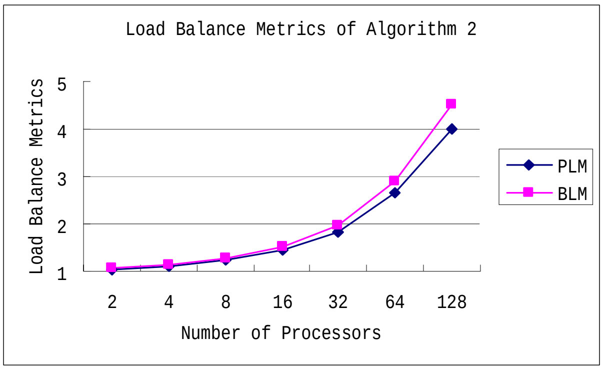


Figure 6 LBM of Algorithm 2

These two figures show that the increasing of the number of processors decrease the microscopic load balance. This conclusion is similar to the one obtained from macroscopic analysis.

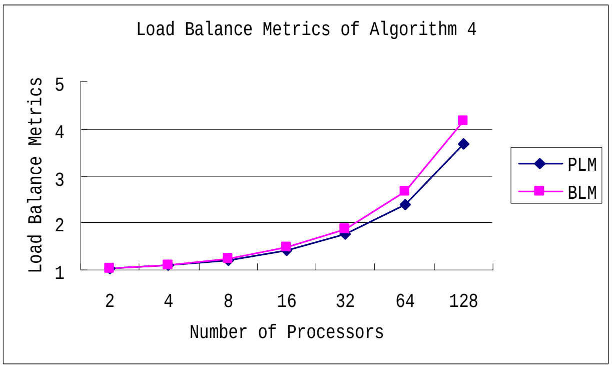


Figure 7 LBM of Algorithm 4

Comparing figure 6 with figure 7, it is clear that the microscopic load balance of algorithm 4 is better than that of algorithm 2. It is another proof that using more stochastic variable will improve the load balance.

6. CONCLUSIONS

DCA deals the packets to different processors. It is very similar with the algorithms in server cluster, but quite different from the algorithms in parallel computing. Because the arrival rate of the packets is much higher than that of any other requests in server cluster, communication among processors becomes unacceptable due to the restriction of processing time.

Based on fields in IP packet header and the concepts of bit entropy and bit flow entropy, this novel load-balancing algorithm can keep the relativities among packets, and no communication among processing nodes is required.

It is shown that the higher the randomness of the bits chosen, the better the balance achieved. And the XOR operation can also give help to improve the macroscopical balance. Both the macroscopical and microscopical analysis shows that the size of the cluster used to balance the work is limited, and the best number of node is between 4 to 8. However, this conclusion also suggests that if some context is introduced, e.g. the current workload of nodes or the prediction of workload in certain period of time, the size of the cluster could be expanded.

The algorithm described in the paper not only can be used in high speed Intrusion Detection System, but also can used in any other situation need to keep the relativities among packets, for example, flow recognition, and application layer behavior analysis etc. in network measurement.

7. REFERENCES

[1] CHENG Guang, GONG Jian, DING Wei, “Network Traffic Sampling Measurement Model on Packet Identification”, Acta Electronica Sinica (Tien Tzu Hsueh Pao), Vol.30, No.12A, Dec. 2002: 89-93(in Chinese)

[2] Pankaj Gupta, NIck McKeown, “Algorithms for Packet Classification”, IEEE Network, March/April 2001, pp.24.

[3] Kai Hwang, Zhiwei Xu, “Scalable Parallel Computing: Technology, Architecture, Programming”, China Machine Press, ISBN 7-111-07176-X, pp.91, May. 1999

[4] Richard B. Bunt, Derek L. Eager, Gregory M. Oster, and Carey L. Williamson, “Achieving Load Balance and Effective Caching in Clustered Web Servers”, Proceedings of the forth International Web Caching Workshop, San Diego, California, April 1999, 159-169.

A Measure and Design Method of Security Protocol *

Wang Tao, Guo Heqing, Yao Songtao
Computer Science and Engineering Department,
South China University of Technology
Guangzhou, Guangdong Province 510642, P.R. China
Email: filion@163.net **Tel.:** 0086-020-85285432

ABSTRACT

In this paper we introduce a method for security protocol measurement and redundancy measurement. We formally give the definition of protocol security property (goal) satisfaction measurement and discuss the relating factors of it. By this we give a method to measure the redundancy of the protocol and propose the method of reduction. We then give two application of the method both using reverse inference: analysis method of implicit assumptions and improper assumptions involved in modal logic based protocol analysis, protocol design and generation.

Keywords: security protocol; measure; redundancy; protocol generation; reverse inference;

1. INTRODUCTION

In a decade, the research on security protocol has made a great improvement. Many analysis methods of security protocol have been proposed. The first one was the BAN logic [1], which gave an innovative work to this area. Then, other method based on modal logic emerged, such as AT logic, GNY logic, SVO logic [2], etc. Besides that, the methods based on model checking (CSP + FDR method [3], Murφ [4]), the methods based on theorem proving (strand space [5]), the methods based on formal specification (LOTOS [6], CASPL [7]), the methods based on calculus (Spi calculus [8]) are proposed.

These existing protocol analysis methods have some disadvantages. First, using these methods can prove whether the protocol is correct, i.e. whether the protocol as whole can reach the expectation on functional and secure properties. But they do not give a measurement to the contribution of protocol components, i.e. a step, a field in a step or an operation (encryption, hashing, etc.), to the whole function of the protocol. Second, as a result of the former disadvantage, they do not give a method to measure the redundancy of the protocol and its components, the redundancy are still be found out with experience and direct inspecting to each component of the protocol, which is informal and the improper adjustment may bring new error. Third, these analysis methods do not directly give a related protocol design method. The protocol design without computer aid is an informal, manual work as ever which was error-prone, while the recent automatic protocol generation is based on searching of a large protocol space [9].

In this paper, we intend to discuss a modal logic based method that can measure the satisfaction of the protocol to the security properties and the redundancy of the protocol, and a new way of protocol generation.

In section 1, we give a set of definitions of the method, including protocol goals, satisfaction, redundancy and contribution of protocol components. In section 2 we will give a simple instance of the method's application that show the finding and reduction of redundancy. Section 3 will be the discussion on the method application on improper assumption checking and protocol generation. Section 4 will be the comparison to some related works. In section 5 we give a conclusion of the paper, then give a prospect of this method in section 6.

2. DEFINITIONS

We give a set of definitions using in the method.

Let's firstly recall the "logically entail" from propositional calculus [10]: if for a set of propositions *S*, $S = \{p_1 ... p_n\}$, a proposition *p* to all p_i=TRUE, we have *p*=TRUE, i.e. $p_1 \wedge ... \wedge p_n \supset p$, then *S logically entail p*, written as $S \models p$

p_i is one of the sufficient preconditions of *p*, written as $p_i \geqq p$ (not us $\Rightarrow$ here for differentiation); if $a_1 \geqq a_2$, $a_2 \geqq a_3, \ldots, a_n \geqq p$, then a_1 is a indirect sufficient preconditions of *p*, written as $a_1 > p$.

2.1 Reachable Proposition Set

The formulae in SVO [2] is call "proposition" here, so the set of proposition is defined as $F_\tau$which is the language of formulae in SVO logic

A protocol description gives us a set of initial propositions.

If their is a step in the protocol likes "$A \rightarrow B : n_B, n_A, \{X\}_K$ " then proposition "*TTP received* $(n_B, n_A, \{X\}_K)$" is added to the protocol proposition set.

The protocol's initial assumptions that represent the internal feathers of the protocol should be added into the assumption set.

Definition 1.1: *PDEF* is a set of protocol steps of 4-tuple (*n, src, trg, msg*) where *n* is the step number, *src* is the source participant identifier, *trg* is the target participant identifier, *msg* is the protocolmessage of this step. Each element of *PDEF* can be rewrote as "*#n src→trg: msg*" and *#n* can omitted.

Definition 1.2: The initial proposition set of protocol is *IPS* = $P \cup A$, protocol description proposition set P = $\{p \mid p = Q$ *received X*, "$P \rightarrow Q$: *X*" $\in PDEF$ }, assumption set A={ a_1,..

* Supported by the Science Foundation of Guangzhou City under Grant No.B2-109-550: the Research on Enterprise Information System Integration and Security.

$a_n\}$, $a_i \in F_\tau$.

Definition 1.3: The reachable proposition set of proposition set P is $RPS_P = \mathrm{Y}\ RPS_{Pi}$, $RPS_{Pi} = \{p \mid S \models p, p \in F_\tau, S \in P(RPS_{Pi-1})\}$. $P(RPS_{Pi})$ is the power set of RPS_{Pi} where $RPS_{P0} = P$.

This is a recursive definition of *RPS*. Then, given a initial proposition set *IPS* of the protocol, RPS_{IPS} is the reachable proposition set of the protocol, and we write *RPS* for simple in the remaining part of this paper.

A reachable proposition set of a protocol under a logic is determined by the steps of the protocol, the fields and their operation of every step, the assumptions that used in protocol analysis.

Here logic includes the inference rules of the logic, the axioms, the terms, the semantics and notations [10]. For example, the SVO logic use *believes*, *received*, etc. as its notation, the inference rules of SVO logic are based on these notations. While Kailar logic [11] use *CanProve*, *IsTrustedOn* as its notations and surely the inference rules are deferent. Generally, the discussion of protocol reduction and comparison are in certain logic.

Definition 1.4: a protocol is a 3-tuple (*P, A, L*), where *P* is the protocol proposition set, *A* is the assumption set, and *L* is the logic.

We call the steps, fields, operations and assumptions of a protocol are the factors of the protocol. The operations here refer to the functions applied on data, such as encryption and hashing, etc. In these factors, the former three factors are determined by the protocol definition itself, while the last one is made specified in the protocol analysis by the prover. There is no rule for introducing assumptions, which will brings an uncertain factor in the protocol analysis and make the analysis accomplish some what at will. We will give a method to check the assumptions later in this paper.

Given these four factors and using logic for inference, the RPS of a protocol is decided, for the intruder model is not introduced here and the RPS is attained in an ideal situation. The RPS may be infinite if there is a axiom of "*A believes* $\phi \supset$ *A believes (A believes* ϕ*)*" in logic, such as SVO [2].

The *RPS* of a protocol is showed in Fig.1.

IPS = *P* ∪*A*, *P* is the protocol proposition set, *A* is the assumption set.

2.2 Target Proposition Set

The goals of the protocol include authentication, key agreement, secrecy, non-repudiation, fairness, etc. Usually, a protocol's goal can be described as a set of propositions. In our method, the goal of protocol must be described as a set of proposition.

Definition 1.5: The target proposition set (*TPS*) of protocol $T = \{t_1, ..., t_n\}$, $t_i \in F_\tau$ is a set of propositions which define the goals of a protocol.

Key Agreement

The key agreement between participants is the basic goal of key agreement protocol that has been addressed in the early article. The goals of key shared protocol are describe in [12], but here we adjusted then in SVO:

(1) *A believes B says Y*
(2) *A believes B says (Y, R (G (*R_A*), Y)*
(3) *A believes* $A \xleftrightarrow{K} B$
(4) *A believes fresh* $(A \xleftrightarrow{K} B)$
(5) *A believes B believes* $A \xleftrightarrow{K} B$
The more common goals are described in [13]

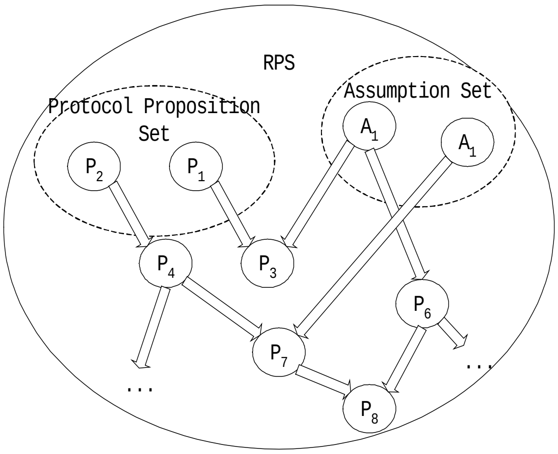


Fig.1 RPS and IPS

Non-Repudiation

One of goals of e-commerce protocol is the non-repudiation goal as in [14]

(1) *A believes B received M*, and
(2) *B believes A said M;*

The proving of the goals in protocol analysis can lead to a conclusion that the protocol is functioned as expected. In the method we proposed here the goals are also the core of the measurement to a protocol.

The protocol may have several goals, or have a complicated goal. In these cases, the TPS may be reduced for there may be some entailment relation between the goals.

2.3 The Satisfaction of the Protocol

Definition 1.6: for a protocol *PT (P, A, L)*, if $RPS \sqsupseteq TPS$, then we say the protocol *satisfies* the protocol goals under an assumption set. Other wise, the protocol is unsatisfied to the goals. The "protocol *satisfies* the protocol goals" is written as *PT (P, A, L)* α *TPS*.

In brief, if all proposition in TPS can be deduced by the proposition from the protocol and assumption using a logic, then the protocol satisfy the goals. This can be represented formally as $\forall T \in TPS$, $P \cup A \models_L T$, *L* is logic using in *P,A* and *TPS*.

The assumption set A is important to the definition to the "satisfy", so it is specified in the definition. In some case of analysis, the protocol it self dose not meet the goals, but there are some unreasonable assumptions or implied assumptions that make the proving of the protocol reasonable and the protocol is claimed right and reach the goal. But there is still flaw in the protocol that will result in an attack. So, the "*satisfy*" is defined on a certain assumption set.

The definition of "satisfy" by RPS and TPS educes the concept of "redundancy" of protocol. We will discuss it later.

Fig.2 and Fig.3 give a graphical representation of "satisfied" and "unsatisfied".

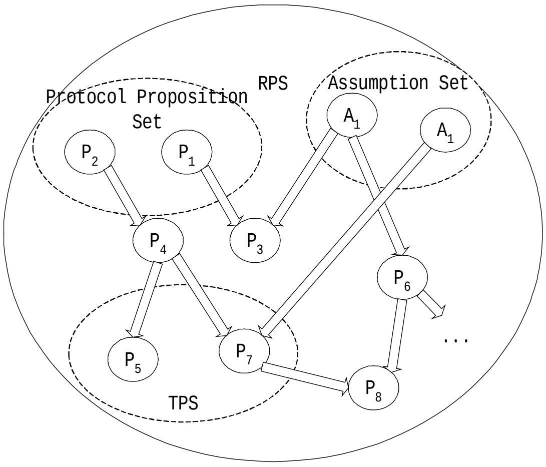


Fig.2 the Protocol satisfies TPS

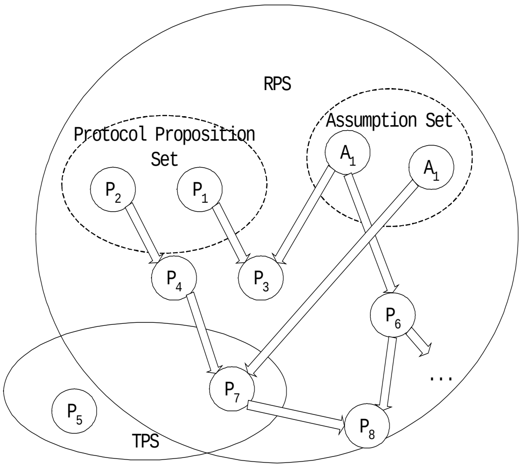


Fig.3 the Protocol does not satisfy TPS

2.4 Contribution

All factors (fields, operations, steps) of the protocol conduce to the *RPS*, but some of them may not be conduce to the achievement of protocol goals. The contribution of the factor should be measured as follow.

Definition 1.7: if a factor corresponding proposition p_a is an indirect sufficient condition of a proposition in *TPS*, $p_a > p_t$, $p_t \in TPS$, then we says that p_a contributes to the protocol goal p_t, this factor is called "core factor" of the protocol (especially the "core factor" to p_t), the factors that don't contribute to a goal are call "redundant factor" to this goal
Obviously, the core factors set of a protocol are reflected by the goals. Given a protocol, to deferent set of goals, the core factors to the protocol maybe (but not must be) deferent.

For instance, a nonce in protocol can do help to authentication goal rather than secrecy goal (at least not directly). An encrypting operation with signature key and public key will serve deferent goals. This is a semantic problem of protocol component.

2.5 Redundancy

Definition 1.8: if a factor is a "*redundant factor*" of each protocol goal, it is the redundant factor of the protocol under the *TPS*.

But sometimes the factor is partially redundant when the sub factor of this factor can make the protocol satisfy the goal but lacking of the whole factor will lead to the unsatisfactory situation of the protocol goals.

Fields
If a field can be take out of the protocol and the protocol still satisfies the goals, then we call the field is completely redundant.

If a field can be replaced by the sub term of that field and the protocol still satisfies the goals, then we call the field is partially redundant. If all M in *IPS* can replaced by M's sub term M', and the RPS is still provable to be correct, and we still have $TPS \subseteq RPS$, then M is partially redundant. The concept of sub term is from strand space [15].
The example of completely redundant field can be found in [6], which we will discuss later.

Steps
If just by some of steps of a protocol, the protocol can achieve the protocol goals, then other steps are redundant. The instance of this is the NSSK's refined protocol, Denning-Sacco protocol [16]].

Operations
If an operation is not executed and the goals are still satisfied, this implies that the operation is not necessary. If in the IPS, X can replace all $\{X\}_K$, or $H(X)$ etc., we say that operation (encryption, hashing) is redundant.

Assumptions
Like field redundancy, there are completely redundancy and partially redundancy for assumptions redundancy.

If the goals can be achieved without an assumption, then this assumption is completely redundant.

If the goals can be achieved by a (or many) weaker assumption that is a sufficient condition of the origin assumption, then this assumption is partially redundant. E.g., if $p \geqq a$ (it means p is one of the sufficient precondition), and using p as the assumption will do for the proving of the protocol correctness, then assumption a is partially redundant.

Reduction of a Protocol
To clarify the problem, we introduce "deduction bundle".

Definition 1.9: a deduction bundle for a proposition x ($x\in$ RPS) is a set of propositions $D_x = \{p| x> p, p\in RPS\} \cup \{p| p> x, p\in RPS\}$, i.e., the deduction bundle of x is the set of all logic predecessors and successors of x (direct or indirect, predecessors or successors).

Then, supposed that a core proposition set

$CPS = \bigcup D_p$ where $p \subseteq TPS$, then RPS - CPS ("-" for set minus) is the redundant proposition set, which implied the proposition in redundant proposition set can be reduced.The redundancy of RPS is showed as Fig.4

All nodes in dark are in the contributing bundle to the TPS. p_6 and p_8 is in the redundant proposition set.

But it should be noticed that the assumption redundancy should be reduced firstly. It is to reduce the effect of improper assumption that was introduced in fault.

Now we give some definitions about protocol comprise, protocol equivalent, simplest protocol.

Definition 1.10: for two protocols, they have deferent protocol proposition set PT_1 (P_1, A, L) and PT_2 (P_2, A, L), for all possible assumption set A, $IPS_1= P_1 \cup A$, $IPS_2= P_2 \cup A$, we have $RPS_{IPS1} \subseteq RPS_{IPS2}$,then we call protocol PT_2 is *logically comprise* protocol PT_1 (written as $PT_1 \le PT_2$). If the result is $RPS_{IPS1} = RPS_{IPS2}$, then we call protocol PT_2 is *logically equivalent to* protocol PT_1 (written as $PT_1 = PT_2$).

Definition 1.11: for two protocols PT_1 and PT_2, $PT_1 \le PT_2$, and for the simplest assumption set, $PT_1 \alpha$ TPS, and there is no PT', $PT' \le PT1$ and $PT' \alpha$ TPS, then we call PT_1 is the *simplest description* of PT_2 on TPS, written as $PT_1=SD_{TPS}(PT_2)$

Note that if $PT1=SD_{TPS}(PT2)$, $PT3=SD_{TPS}(PT4)$, there may not be necessary that $PT_1 = PT_3$ or $RPS_{PT1} =RPS_{PT3}$. Because RPS is a result of inference, this relation "≤" is a partial order relationship, not full order.

3. INSTANCE ANALYSIS

Now we give a simple example to clarify the method. The sample protocol for analysis is a registration protocol proposed in [6]. Choosing it as a subject is because that it is simple but representative and has a redundant field.

The protocol is used in a multimedia service authentication and is an enhanced version of the original one in [17], which using the classical Guillou-Quisuater zero-knowledge identification scheme [18].

The sample protocol from [6] is as follow, which had been refined for a flaw:

1 : *User* → *TTP* : *Register Request* <*UserID*, K_U^P, {*n*} K_{TTP}^P >

2 : *TTP* → *User* : *Register Challenge* <*d*, {*n*} K_{TTP}^S >

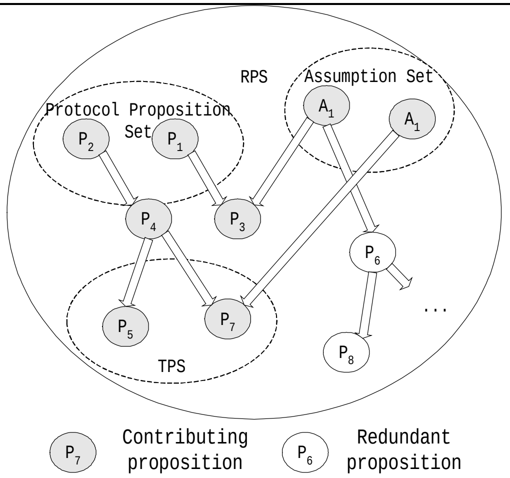


Figure 4 : The deduction bundle and redundancy

3 : *User* → *TTP* : *Register* Response < *F* (*B*, *d*)>

4+: *TTP* → *User* : *Register Ack* <{*Yes*, *UserID*, *n*, *d*} K_{TTP}^S >

4−: *TTP* → *User* : *Register Ack* < {*No*, *UserID*, *n*, *d*} K_{TTP}^S >

F (*B*, *d*) in message 3 represents the special encryption of the GQ model where B is the credentials of User. As GQ model is usually used as a signature model [19] [20], it can be rewrite as $\{d\}_B$. *d* is the nonce of TTP and *n* is the nonce of User.

The first step to analysis this protocol using the method in this paper is to initialize the protocol's factors.

(1) Steps, fields and operations.

TTP received (UserID, K_U^P, *{n}* K_{TTP}^P *)*

TTP received (d, {n} K_{TTP}^S *)*

TTP received $(\{d\}_B)$

TTP received ({Yes, UserID, n, d} K_{TTP}^S *)*

TTP received ({No, UserID, n, d} K_{TTP}^S *)*

(2) Assumed propositions:

TTP believes fresh (d)

U believes fresh (d)

TTP received $\{d\}_B$ →*TTP sees d*

U believe PK$_\sigma$*(TTP,* K_{TTP}^P *)*

(3) Goals:

TTP believes U says d, U believes TTP says Yes

The logic is base on SVO logic summarized in [21] and we used a simple inference engine of it implemented in Prolog to get the propositions set and found the TPS (goals) lied on the RPS, which implied that the simple goals are satisfied. We also found that *n* in the protocol didn't do any effective contribution to the goals listed above, which was already pointed out in [6]. *n* did contribute to the authentication between U and TTP if the goals included U should identified the TTP, as we did in another test, but to this protocol and the listed goals, *n* is redundant.

The proving is showed in appendix in this paper although the goals are simple and the trace of proving is short.

Then redundant factor *n* should be reduced. For a nonce is atomic, n could not be truncated, so the only way to reduce it is to take it out of the protocol. Then we also get a new protocol as in [6]:

1 : User $\rightarrow$ *TTP : Register Request <UserID,* K_U^P *>*

2 : TTP $\rightarrow$ *User : Register Challenge <d>*

3 : User $\rightarrow$ *TTP : Register Response <F (B, d)>*

4+: TTP $\rightarrow$ *User : Register Ack <{Yes, UserID, d}* K_{TTP}^S *>*

4 —: TTP $\rightarrow$ *User : Register Ack < {No, UserID, d}* K_{TTP}^S *>*

4. APPLICATIONS

We can use this measure method in protocol design and other intentions. These using are all based on reverse deduction.

4.1 Reverse Deduction

Reverse deduction is a way to get precondition from conclusion and rules: given a conclusion, a set of deduction rules and some of the preconditions, one or few unknown precondition can be listed manually or generated by inference engine.

4.2 Implicit Assumption

When using modal logic method to prove protocol correctness and idealization a protocol is done, we have introduced some assumption consciously or unconsciously. Not all of these assumptions are proper. An example is the misuse of assumption "*B believes fresh(*$P \xleftrightarrow{K} Q$*)*" in [1] (rewriting in SVO logic [2]).

Using the method of this paper, we can check the implicit assumptions and whether they are proper for the protocol. The steps should be as follow:

(1) List the proposition that can just be inferred from the fields, operations and steps of the protocol, not from the assumptions than we get an IPS without assumptions.
(2) Get a RPS from this initial proposition set.
(3) Focus on the unsatisfied goals and infer reversely the preconditions that are needed in making the goals satisfied.
(4) Check these preconditions with known assumptions and find the implicit assumption.

If the explicit assumptions are trusted to be proper, we can also us them in initial proposition set which will bring us a more efficient way of finding any implicit assumptions. In this case, finding implicit assumptions become an accessional validation of protocol.

Note that this implicit assumption is a proposition that can't be a conclusion of IPS, for the implicit assumption is one of the preconditions of an unsatisfied goal.

4.3 Protocol Design

Fields, operations and steps are factors in protocol analysis, while in protocol generation they become the protocol components.

As components of protocol, a field, a step or an operation in the protocol definition will bring certain proposition to the initial proposition set and determined partially the eventual *RPS*. Using the reverse inference, given a certain goals, a protocol and assumptions can be composed by these protocol components. This gives another way of protocol generation.

5. RELATED WORKS

In [14] Zhou et al pointed out that Kailar logic using in E-commerce protocol analysis have some disadvantages and gave some adjustment to Kailar logic in [22]. They proposed a new logic in analyzing non-repudiation and fairness. They introduced "initial possess set" and "final possess set" into the logic which represent the knowledge of participants and like the set of "seen message" in SVO logic [2]. Here they use "initial possess set" to substitute the assumptions in BAN-like logics.

This adjustment of Kailar logic is deference to the method in this paper in two aspects. First, the set using in our measure method is a set of proposition, not a set of "possess knowledge" that are composed by known messages and keys as in [22]. Because the protocol goals are directly described as a set of propositions, we can gave a direct measure to the satisfaction of the protocol goals and introduce the measurement of redundancy naturally which are not involved in the adjusted Kailar logic. Second, the initial possess set in [22] is not essentially deferent with the initial assumptions of protocol as used in BAN-like logic [1], and the assumptions in the form of proposition is better for analyzed on their redundancy as discuss above, so we still use initial assumptions in our method.

The automatic protocol generation had been hot topic in recent years. Perrig and Song gave a way in automatic protocol generation in their paper [9]. Their method is based on a concept of "protocol space" which contains all possible protocols generated by programs. Users give their specification and requirements of protocol and under the help of "protocol screener" which represent some specific properties of protocol they get protocol(s) as result. According to their research, the "protocol space" is quite large and as much as 10^{12} protocols are in it [9].

Compare to our method, they didn't bring forward the concept of protocol measure and redundancy measure specifically, although the cost is one of the considering factors in their protocol choosing. Another significant deference is that when the method using in protocol design reverse inference is involved which limits the options into a smaller range.

6. CONCLUSION

In this measure and design method we had introduce protocol measurement and redundancy measurement. This gives a precise way to examine the component in a protocol. To do this, we formally give the definition of protocol security property (goal) satisfaction measurement and discuss the relating factors of this satisfaction. From this we give a method to measure the redundancy of the protocol and propose the method of reduction, which had been a new aspect of protocol analysis. We then give two application of the method based on reverse inference: analysis method of implied assumptions and improper assumptions using in protocol proving, protocol design. This may give a new way of protocol generation. But before the further application, the research should be taken on some other problems listed below.

7. FUTURE WORK

More Precise Measure of Contribution and Redundancy
The discussion above has given an approximate measure of factors' contribution and redundancy to the protocol goals. But which factor has more contribution to the goal than other factors, which factor should be reduced firstly. All these questions are not precisely defined.

Intruder
The BAN-like logics (BAN, AT, GNY, SVO), as well as other modal logic (Kailar), have a common disadvantage: they cannot find the flaw and attack scheme directly [23]; the analysis of implicit and improper assumptions may imply some hints of flaw and attack. But introduction of intruder factor into this method may give us another way.
Intruder will give some new propositions that describe their abilities of replaying, concatenating, etc. into the assumption set.

The TPS of intruder is related to the protocol's goal. Usually the intruder's intent is to attain illegal benefit or interrupt the normal protocol functioning. So the intruder may have constructive goals and destructive goals correspondingly. As a sample, in SVO logic, the constructive goals and destructive goals to the goals of the instance protocol mentioned above are that:

Goals of the protocol:	*TTP believes U says d,* *U believes TTP says Yes*
Intruder's constructive goals	*(TTP believes I says d) or* *(U believes I says Yes)*
Intruder's destructive goals	*TTP believes ¬ (U says d),* *U believes ¬ (TTP says Yes)*

If the RPS of the system including intruder can cover the TPS of intruder, it may imply that there is an attack.

Extending to Other Non-Modal-Logic Methods

The method here is based on modal logic. The RPS and TPS are represented in some propositions and the applications are based on reverse inference. As the research subject is same, the modal logic may have some inherent relation with other protocol analysis method. The BAN-like logic can be used in give the strand space method a semantics expression [24].

With this inherent relation, the idea of measurement and redundancy in this paper may be also extended to other method such as strand space, Spi calculus and multi-set rewriting [25], which will be the next research step of us.

8. REFERENCES

[1] Burrows M, Abadi M, Needham R. A logic of authentication. In: Proceedings of the Royal Society of London A, Vol 426. 1989. 233~271.

[2] Syverson PF, van Oorschot PC. On unifying some cryptographic protocol logics. In: Proceedings of the 1994 IEEE Computer Society Symposium on Research in Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1994. 14~28.

[3] Roscoe A, Goldsmith M. The perfect 'spy' for model-checking cryptoprotocols. In: DIMACS Workshop on Design and Formal Verification of Security Protocols. 1997.

[4] Mitchell J, Mitchell M, Stern U. Automated analysis of cryptographic protocols using murphi. In: Proceedings of the 1997 IEEE Computer Society Symposium on Research in Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1997. 141~151.

[5] Thayer FJ, Herzog JC, Guttman JD. Strand spaces: Why is a security protocol correct? In: Proceedings of the 1998 IEEE Symposium on Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1998. 160~171.

[6] G. Leduc, F. Germeau. Verification of Security Protocols using LOTOS - Method and Application. Computer Communications, 23(12):1089-1103, July 2000.

[7] Millen JK. CAPSL: Common authentication protocol specification language. Technical Report, MP 97B48, The MITRE Corporation, 1997.

[8] Abadi M, Blanchet B. Secrecy types for asymmetric communication. In: Proceedings of Foundations of the Software Science and Computation Structures. 2001. 35~49.

[9] Adrian Perrig and Dawn Song. Looking for diamonds in the desert extending automatic protocol generation to three-party authentication and key agreement. In lSth IEEE Computer Security Foundations Workshop, pages 64 76. IEEE CS Press, June 2000.

[10] Artificial Intelligence: A New Synthesis, Nils Nilsson, Morgan Kaufmann, 1998

[11] Kailar R. Accountability in electronic commerce protocols. IEEE Transactions on Software Engineering, 1996, 22 (5) : 313〜328

[12] van Oorschot PC. Extending cryptographic logics of belief to key agreement protocols. In: Proceedings of the 1st ACM Conference on Computer and Communications Security. ACM Press, 1993. 233~243.

[13] Lowe G. A hierarchy of authentication specifications. In: Proceedings of the 10th IEEE Computer Security Foundations Workshop. Los Alamitos: IEEE Computer Society Press, 1997. 31~43.

[14] ZHOU Diancui,QING Sihan,ZHOU Zhanfei. Limitations of Kailar Logic. Journal of Software, 1999,10(12):1238~1245. (In Chinese)

[15] Thayer FJ, Herzog JC, Guttman JD. Strand spaces: Proving security protocols correct. Journal of Computer Security, 1999,7(2-3): 191~230.

[16] D, Sacco G. Timestamps in key distribution protocols. Communications of the ACM, 1981,24(8):533~536.

[17] S. Lacroix, J.-M. Boucqueau, J.-J. Quisquater, B. Macq, Providing equitable conditional access by use of trusted third parties, in:Proceedings of ECMAST 96, Louvain-la-Neuve, Belgium, May 1996, pp. 763-782.

[18] L. Guillou, J.-J. Quisquater, A practical zero-knowledge protocol fitted to security microprocessor minimizing both transmission and memory, Proceedings of Eurocrypt 88, Springer, Berlin, 1988 (pp. 123-128).

[19] G. Itkis and L. Reyzin. SiBIR: Signer-base i ntrusionresilient signatures. In Advances in Cryptology---Crypto'02, 499--514, Berlin, 2002. Springer-Verlag.

[20] Li-Shan Liu, Cheng-Kang Chu, and Wen-Guey Tzeng. A Threshold GQ Signature Scheme. In: Applied Cryptography and Network Security (ACNS'03), LNCS

2846, pp. 137-150. Springer-Verlag, 2003.

[21] Paul Syverson and Iliano Cervesato. The logic of authentication protocols. In R. Focardi and R. Gorrieri, editors, Foundations of Security Analysis and Design, volume LNCS 2171. Springer-Verlag, 2001.

[22] ZHOU Diancui,QING Sihan,ZHOU Zhanfei. A New Approach for the Analysis of Electronic Commerce Protocols. Journal of Software,2000,12(9):1318~1328. (In Chinese)

[23] Qing SH. Twenty years development of security protocols research. Journal of Software, 2003,14(10): 1740~1752. (In Chinese)

[24] Paul Syverson. Towards a strand semantics for authentication logics. Electronic Notes in Theoretical Computer Science, 20, 1999.

[25] Iliano Cervesato, Nancy A. Durgin, Patrick D. Lincoln, John C. Mitchell, and Andre Scedrov. A meta-notation for protocol analysis. In P. Syverson, editor, Proceedings of the 12th IEEE Computer Security Foundations Workshop — CSFW'99, pages 55–69, Mordano, Italy, June 1999. IEEE Computer Society Press.

Appendix:

The proving of the first goal:

TTP received $\{d\}_B$
TTP believes TTP sees d
Then using SVO-12: TTP believes TTP received $\{d\}_B$
TTP believes $PK_\sigma(U,B)$
TTP believes (TTP believes $PK_\sigma(U,B)$ *)* *(using SVO-2)*
Then using SVO-4, we get: TTP believes (U said d)
And TTP believes fresh(d), then :
TTP believes (U says d)

The proving of second goal :

U received $\{$ *Yes, UserID, n, d* $\}_{K^S_{TTP}}$
U sees $\{$ *Yes, UserID, n, d* $\}_{K^S_{TTP}}$
U believes U sees (Yes, UserID, n, d)
U believes U received (Yes, UserID, n, d)
For U believes $PK_\sigma(TTP, K^P_{TTP})$*, then :*
U believes TTP said (Yes, UserID, n, d)
U believes fresh(d)
U believes fresh(Yes, UserID, n, d)
U believes TTP says (Yes, UserID, n, d)
U believes TTP says d

Implement Distributed Parallel Computing Based on EJB*

Huang Zhehuang [1] Jiang Xiufeng [2] Wang Meiqing [3]
College of Mathematics and Computer Science, Fuzhou University
Fuzhou，Fujian 350002，China
Email: [1] quwang1@163.com, [2] jxf1963@hotmail.com,[3] mqwang@fzu.edu.cn **Tel:** 0591-3734383

ABSTRACT

The EJB standard of Sun Company is a typical designing criterion of component. In this paper we build up a distributed parallel computing system based on EJB and implement the parallel algorithms of matrix multiplication with the system. Experimental results show that the system can obtain a rapid speedup with the increase of computing nodes and the sizes of matrixes.

Keywords：EJB, distributed parallel computing, Component

1 INTRODUCTION

Many large-scale scientific computations need to use high-performance computers. However, the costs of developing and using high-performance computers are high. These computers also require specialized operating systems and programming languages, which leads to poor portability. As a result, computation-intensive research using high-performance computers suffers from many limitations. Today, the most of such research takes advantage of existing network system by performing distributed parallel computing, which achieves similar results as high-performance computing. Because of the widespread availability of LAN/WAN and Internet to researchers and end users, there are many applications of distributed parallel computing in a variety of end-use situations. Distributed object technology that first arrived on the scene in the 90's provides a powerful solution to software development on network computing platforms. Currently, distributed object technology has become the core technology for building service-oriented architecture and software components, and it has demonstrated tremendous power in the development of large-scale distributed application systems. Through the years, three flavors of distributed computing technology have emerged and matured. They are COM/DCOM from Microsoft, EJB/RMI from Sun, and CORBA from OMG.

In the last two to three years, internet-based Grid computing[1] environment has attracted a great deal of attention from scholars at both home and abroad. Grid-based scientific computing will take the place of supercomputers and become the new computing model because of the following advantages: platform heterogeneity, location transparency, and resource sharing. Among the afore-mentioned distributed computing technologies, EJB/RMI has the unique features of portability, platform independence, object-oriented, simplicity, and security, and thus it is well suited for programming in a Grid computing environment.

In this paper, we analyze the architecture of EJB, build a distributed parallel computing system based on EJB standard. Then, we implement a distributed parallel computing algorithm of matrix multiplication on the system. Finally, we give the experimental results and compare the speedups in the systems with different computing nodes.

2 EJB/RMI TECHNOLIGY

In Java version 1.1 and later, RMI (Remote Method Invocation) technology is used to develop distributed computing. Distributed computing mechanism based on JAVA RMI is shown as the Figure 1. It mainly consists of the works of server sides and client sides [2].

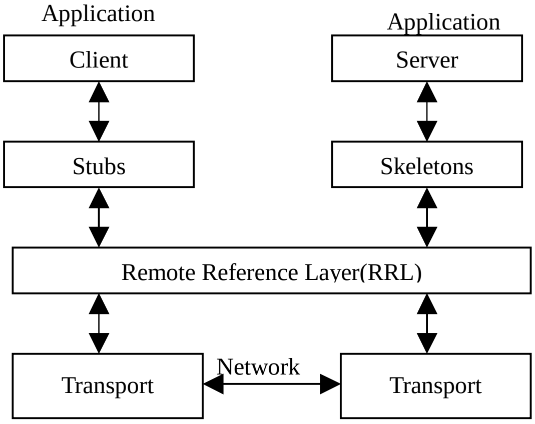


Figure 1 Distributed mechanism based on JAVA RMI

In the server sides following works must be completed:

- Define all Interface for remote object classes,
- Write the Interface Implementation Files,
- Compile the files and use the rmic compiler to generate the stub and skeleton,
- Write the server program to manage remote objects,
- Start the rmiregistry on the server sides to register remote objects,
- Run the server programs in server sides.

In the client sides following works must be completed:

- Create and install a RMI security manager,
- Call the Naming.lookup method that sends a request to a server which returns a remote object reference,
- Invoke remote methods of the remote objects.

As mentioned above, we can use the rmic compiler to generate the stub and skeleton, and start the rmiregistry to register instance references. But the programmers have much work to do in order to build a distributed computing system. So the development of RMI is limited

EJB [3, 4] (Enterprise JavaBeans) is one of the key components of Sun's Java 2 Enterprise Edition (J2EE). The Enterprise

This work was supported by the high school's computer class funds of Ministry of Education, China under Grant No. 0030103002

JavaBeans 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 the Enterprise JavaBeans architecture are scalable, transactional, and multi-user secure. These applications may be written once, and then deployed on any server platform that supports the Enterprise JavaBeans specification.

The architecture of EJB is shown as figure 2, it mainly contains the following parts [3, 4]:

(1) Enterprise JavaBeans

An EJB is a component model in Java, which is similar to DCOM component developed by Microsoft. An EJB is made up of four essential parts: a home interface, a remote interface, a bean class, and an XML deployment descriptor.

Every enterprise bean must have a home interface, which always provides methods to locate or to create instances of the remote interface. An enterprise bean’s home interface controls the life cycle of a bean, and it is responsible for three main operations: create, locate, and destroy instances of a bean.

The remote interface specifies methods according to the application logic encapsulated by the enterprise bean. The remote interface declares business methods at run-time. Clients of the enterprise bean access the methods through its remote interface.

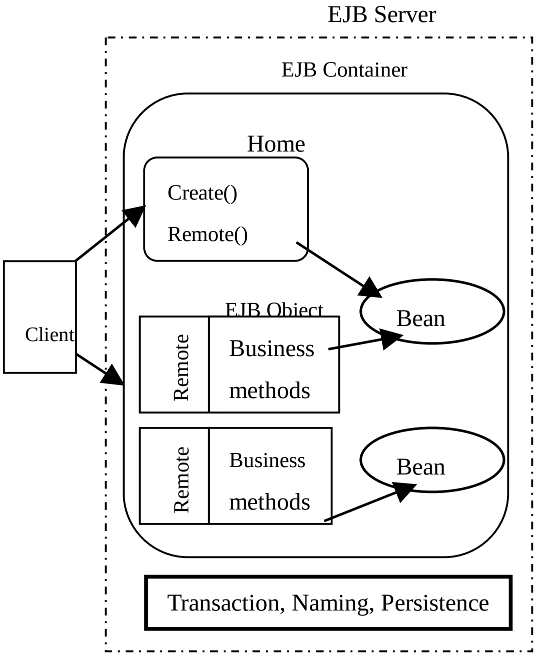


Figure2 The architecture of EJB

A bean class is an implementation of functions declared in the enterprise bean’s home and remote interface. That is, the bean class implements the business methods.

Then the XML deployment descriptor communicates run-time attributes to the server. It specifies the declarative semantics that describe an EJB and the services it requires. The deployment descriptor allows programmers to specify an access control entry. It can be defined for individual methods or for all methods of the enterprise bean.

Deployment also can be implemented by a special deployment descriptor file, which supports parameters that govern enterprise bean behavior, such as if a bean requires transactions.

The developer must include all of these four parts and package them in a JAR file for the EJB to work correctly, then copy the compiled EJB interfaces to the client, and define this location in the client class path.

The EJB specification defines four types of enterprise beans:

● **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.

● **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.

● **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.

● **Message-Driven Beans:**

A message-driven bean (MDB) acts as a JMS message listener. MDBs are different from session beans and entity beans because they have no remote, remote home, local, or local home interfaces. They are similar to other bean types in that they have a bean implementation, they are defined in the ejb-jar.xml file, and they can take advantage of EJB features such as transactions, security, and lifecycle management.

(2) EJB Container

An EJB container is probably the single most important concept in the enterprise beans approach as it provides the most benefit to the developer. Object-based middleware platforms like CORBA or RMI free a distributed application developer from the networking aspects of the application by providing mechanisms for object location, data marshaling, and so forth. An EJB container manages the enterprise beans contained within it. For each enterprise bean, the container is responsible for registering the object, providing a remote interface for the object, creating and destroying object instances, checking security for the object, managing the active state for the object, and coordinating distributed transactions. Optionally, the container can also manage all persistent data within the object.

The container and service providers implement the EJB infrastructure. This infrastructure deals with distribution aspects, transaction management, and security aspects of an application.

(3) EJB server

An EJB server is a process or an application, which provides an environment to support the execution of applications developed using Enterprise JavaBeans (EJB) components. The server manages the resource allocations of the applications and provides the access to system service. Enterprise JavaBeans typically contain the business logic for a J2EE application.

There are many application servers: BEA WebLogic, IBM WebSphere, Oracle 9iAS, JBoss, etc…. We choose BEA WebLogic as application server due to its comprehensive support for EJB, Java, and J2EE.

An EJB server must provide one or more EJB containers.

(4) EJB client

An EJB client may use standard JNDI and EJB calls to obtain a reference to the remote interfaces of the beans.

Clients do not work directly with the beans. Instead, clients communicate with beans using the Java Naming and Directory Interface (JNDI) to locate the bean's home interface, which returns to the client a reference to an object that implements the home interface. Clients invoke a method on the bean's home interface to get a reference to the bean's remote interface. Then the bean instance can be used to make calls to the bean's business methods.

(5) Assistant System(RMI, JNDI, JDBC, JMS, JTS, ETC.)

CORBA clients can access EJB components deployed in CORBA-based EJB servers. A non-bean CORBA client, such as a C++ client, can access the bean and invoke its methods using standard CORBA calls.

3 A DISTRIBUTED PARALLEL COMPUTING SYSTEM BASED ON EJB

Similar to the distributed computing System based on java RMI[5], A distributed computing System based on EJB consists of client computers and a server pool, as shown in Figure 3. The server pool is composed of computing nodes $\{N_1, N_2, \Lambda, N_n\}$.

In node N_i, where $1 \le i \le n$, the following steps need to be completed:

(1) Install WebLogic Server,

(2)Develop EJB components:

- Create a Stateless Session Bean named *Bean_compute_i*, and add a business method in *Bean_compute_i* named $Compute_i(task)$. Where, the parameter 'task' refers to the computing task submitted by clients.
- Compile Bean_compute, and create EJB Deployment Descriptors, then package Bean_compute and related classes into a standard Java archive (*JAR)* file,

(3) Start the WebLogic Application Server, and deploy the JAR file.

After these works done, the node N_i has completed the configuration and is ready to handle clients' invocation.

A client is responsible for submitting computing tasks. A programmer must divide a computing task into sequential and parallel parts. The sequential parts are executed on the client itself. The parallel parts are distributed to *N* computing nodes through multiple threads. The Nodes complete the tasks and return the results to the client. The client combine the return results and outputs the final result.

The main steps to be completed in the client:

(1) Divide the computing task into sequential and parallel parts $\{Par_1, Par_2, \Lambda, Par_n\}$,

(2) Execute the sequential parts in the main thread,

(3) The main thread spawns *N* threads $\{T_1, T_2, \Lambda, T_n\}$.

(4) In thread T_i, $1 \le i \le n$, do the following steps:

- Get a JNDI initial context. Use the initial context to look up a reference to the EJB home of $Bean_compute_$i in Node N_i, named as Ref_EJB_home;
- Obtain a reference to the bean instance by calling create method or find method in Ref_EJB_home, named as Ref_bean_instance;
- Call the business method $Compute_i(Par_i)$ of $Bean\_compute\_i$ by using Ref_bean_instance;
- Wait for the results returning from the N_i

(5)Return to the main thread.

4 THE IMPLEMENTATION OF THE DISTRIBUTEPARALLEL ALGORITHM OF MATRIX MULTIPLICATION BASED ON EJB

In the experiments we implement the distributed parallel algorithm of matrix multiplication based on EJB. That is, we want to calculate

$$C_{2N,2N} = A_{2N,2N} * B_{2N,2N}$$

where $N > 0$ is a interger. In order to implement the calculation in the EJB environment, we partition matrixes $A_{2N,2N}$, $B_{2N,2N}$ and $C_{2N,2N}$ into four $N \times N$ sub-matrixes[6]:

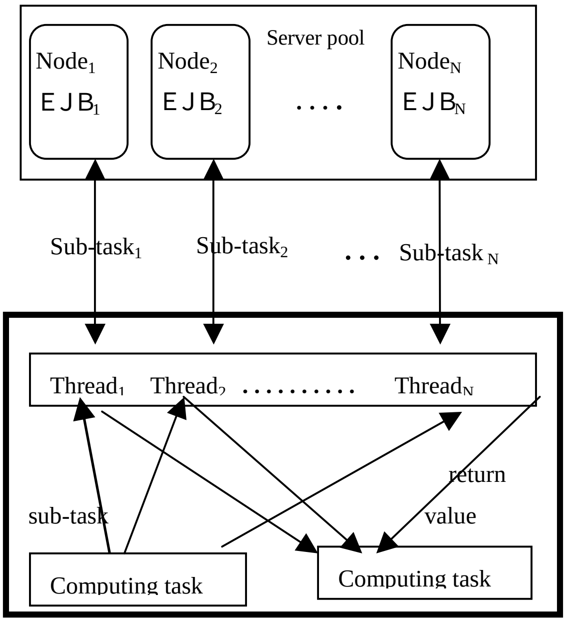


Figure 3 Distributed and Parallel Mechanism based on EJB

$$A_{2N,2N} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B_{2N,2N} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \text{ and}$$

$$C_{2N,2N} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

Where

$$C_{11} = A_{11}B_{11} + A_{12}B_{21}, \quad C_{12} = A_{11}B_{12} + A_{12}B_{22},$$
$$C_{21} = A_{21}B_{11} + A_{22}B_{21}, \quad C_{22} = A_{21}B_{12} + A_{22}B_{22}.$$

In every computing node we create a Stateless Session Bean, there are two methods in the bean class to implement matrix multiplication: AddMatrix() to compute the sum matrix of two matrixes; MutMatrix() to compute the product matrix of two matrixes:

```
public long[ ][ ] MutMatrix (long A[ ][ ],long B[ ][ ],long C[ ][
], int n) {
            for (int i=0;i<n;i++)
                          for (int j=0;j<n;j++)
                          {    for (int k1=0;k1<n;k1++)
                             C[i][j]=C[i][j]+A[i][k1]*B[k1][j];
                          }
        return C;
}

public long[ ][ ] AddMatrix (long A[ ][ ],long B[ ][ ],long C[ ][
],int n)
{       int i,j;
            for (i=0;i<n;i++)
            for (j=0;j<n;j++)
            { C[i][j]=A[i][j]+B[i][j];
            }
            return C;
  }
```

Because the parallel tasks of matrix multiplication are identical in all nodes, we only need to develop an EJB on a computing node, and then deploy it to all other nodes.

The client program use multi-thread technology of Java to connect with computing nodes. In each thread，we look up a reference to the EJB home of the Stateless Session Bean in a server, call the business method of AddMatrix() and MutMatrix() . Then we submit the Sub- matrixes to every computing node respectively, which return the results to the Client. Finally, the Client program merges the result obtained from each thread, and outputs the end result.

(1) If client invoke the local program to complete the matrixes multiplication computing, we use the sequential algorithm to calculate C in the local machine;
(2) If client invoke the methods of one node to complete the matrixes multiplication computing, we use the sequential algorithm to calculate C in the node;
(3) If client invoke the methods of two nodes to complete the matrixes multiplication computing, we use node 1 to calculate C_{11}, C_{12} and node 2 to calculate C_{21}, C_{22};
(4) If client invoke the methods of four nodes to complete matrixes multiplication computing, we can use node 1 to calculate C_{11}, node 2 to calculate C_{12}, node 3 to calculate C_{21} and node 4 to calculate C_{22}.

5 EXPERIMENTAL RESULTS

The experiments are implemented in a local network which is composed of 5 PCs, every PC includes a 1.5GHz CPU with 128M Byte RAM. We use Windows 2000 professional version as the operating system and JBuilder9 and WebLogic8.1 Server [7] as the developing tools.

A summary of the experimental results is shown as table 1.The times (Unit: millisecond) are obtained in client machine when the client program uses different number of nodes to compute matrix multiplications and each data is the average time of ten calculations

Table 1 different times when using different

rank of matrixes Running time method	128	256	512	640
(1)	109	1513	23610	65218
(2)	3141	4620	25563	68172
(3)	3585	4326	19796	39166
(4)	3612	4518	10875	23735

(1) Sequential program in local machine; (2) sequential program in one node; (3) parallel program using two nodes; (4) parallel program using four nodes.

Let T_1 = The running time of sequential program in local machine and

T_n = The running time of parallel program in n nodes;

then we can compute the running time differences between sequential program and the parallel program:

$$\Delta Y = T_1 - T_n$$

and the Speedup

$$S = \frac{T_1}{T_n}$$

Tabel 2 The Speedup of parallel programs

size of matrix ΔY / S method	128*128	256*256	512*512	640*640
Parallel programusing two nodes	−3476 0.03	−2813 0.35	3814 1.19	26052 1.67
Parallel programusing four nodes	–3603 0.03	–3005 0.33	12735 2.17	41483 2.75

In table 2 we compare the values of ΔY and S obtained by different methods.

The figure 4 shows the Speedups of parallel programs using two nodes and four nodes to compute different ranks of matrixes.

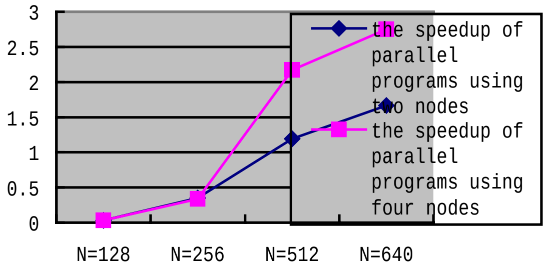


Figure 4 the speedups of parallel programs using two nodes and four nodes

with the increasing of nodes and the ranks of the matrixes. When the size of matrixes is small, we can not obtain a speedup, contrarily, the running times of parallel programs are larger than that of the sequential pro-gram because of the communication time in a distributed computing system. The overhead of invoking a bean is at least several seconds. Using local programs of a single computer to complete the tasks is the simplest method. There is no remote method invocation in this method. The time it cost is the smallest when the ranks of matrixes are small, because EJB is based on the remote method calls of RMI. The overhead of invoking a bean is at least several seconds. When the calculation is small, for example 128*128,the time spent on calculation is only 109ms,far less than other methods. But when the calculation is big enough, such as 512*512, using 4 computers to perform the calculation can reduce the amount of time by half compared to using a single machine. So the distributed computing system based on EJB is suitable for high complexity computing. In future research, we will use the system to run the fractal video compression program and the image restoration program based on PDE [8].

It is worth mentioning that the authors find out the execution time of programs based on EJB is still larger than that based on DCOM. But EJB is more suitable for grid computing environments as Java is a platform-independent language and can be used for heterogeneous systems. With the improvement of the speed of java and the emergence of EJB local interface, the efficiency of EJB will be getting better.

6 CONCLUSIONS

EJB is a new technique based on J2EE platform, it represents the future of server-side component. In this paper, we implement a distributed parallel system based on EJB technique and implement a parallel algorithm of matrix multiplication. The experimental results show that the system can obtain a relative high speedup with the increasing of computing nodes and the sizes of the matrixes.

7 REFERENCES

[1] I. Foster and C Kesselman (ed.). The Grid: Blueprint for a New Computing Infrastructure. Morgan-Kaufman, 1998.

[2] Fundamentals of RMI Short course [DB/OL], http://developer.java.sun.com/develop/online training/rmi/RMI.html, 2002-11.

[3] Demichiel L G, Yalcinalp L, Krishnan S. Enterprise JavaBeans Specification, Version 2.0. Sun Microsystems, Inc., 2001-01-14

[4] John W. Mastering. Enterprise JavaBeans and the Java 2 Platform. Enterprise Edition, Sons Inc. Wiley Computer Publishing, 1999.

[5], M. Wang, S. Zheng, W. Zheng. JDCS: A Distributed Computing System Implementing High Performance Computing. *Computer Engineering and application* Vol.21, 2002. (in Chinese).

[6] LI Xiaozhou ,LI Qinghua. Implementation of Matrix Multiple Cannon Parallel. *ComputerEngineering, June 2002* (in Chinese).

[7] BEA Weblogic Enterprise Platform and Application Infrastructure [EB/OL]. http://www.bea.com/products/weblogic/server/index.shtml

[8] Barcelos C. A. Z., Chen Y. Heat flows and related minimization problem in image restoration. *Journal of Computers and Mathematics with Application*, Vol.39: 81-97, 2000.

Stand Space Theory and its Application on SET Protocol

Xu Feng[1, 2], Li Dake[2], Huang Hao[1]
[1]Department of Computer Science, Nanjing University, Nanjing, Jiang Su 210093, China
[2]College of Computer & Information Engineering, Hohai University, Nanjing, Jiang Su 210098, China
E-mail: njxufeng@163.com **Tel:** 13951835506

ABSTRACT

This paper detailed introduces Strand space theory- an advanced method of protocol analysis. Then, uses it to analyze Secure Electronic Transaction (SET) protocol and prove its secrecy property.

Keywords: Strand space; Secure Electronic Transaction protocol; protocol proof; formal method; secrecy property (For the completeness of this paper, we introduce the theory of strand space firstly. So from section 1 to section 4 is based on [1])

1 INTRODUCTION

A security protocol is an exchange of messages between two or more parties in which encryption is used to provide authentication or to distribute cryptographic keys for new conversations [2]. Even when security protocols have been developed carefully by experts and reviewed carefully by other experts, they are often found later to have flaws that make them unusable (see, for example [3, 4]). In many cases, the attacks do not presuppose any weakness in the cryptosystem being used, and would be just as harmful with an ideal cryptosystem. In other cases, characteristics of the cryptosystem and characteristics of the protocol combine to cause protocol failure [5, 6,7].

Analyzing security protocols consists mainly of two complementary activities. The first is to find flaws in those protocols that are not correct, and the second is to establish convincingly the correctness of those that are. These activities are interrelated, because the discovery of a flaw may suggest an altered protocol that we may wish to prove correct, and because a failure to prove the correctness of a protocol may suggest a particular flaw.

In this paper, however, we focus on the second activity, proving the correctness of protocols when they are in fact correct. Moreover, at this stage, we will study protocol correctness assuming ideal cryptography.

Strand space theory is an advanced approach which has such characteristics. A strand space is a set of strands, and a strand is a sequence of events that a single principal may engage in. Each individual strand is a sequence of message transmissions and receptions, with specific values of all data such as keys and nonces. It is thus a sequential process that exhibits neither internal nor external choice [8].

A strand space is a set of strands, consisting of strands for the various legitimate protocol parties, together with penetrator strands. One may think of a strand space as containing all the legitimate executions of the protocol expected within its useful lifetime, together with all the actions that a penetrator might apply to the messages contained in those executions.

A bundle is a portion of a strand space. It consists of a number of strands legitimate or otherwise hooked together where one strand sends a message and another strand receives that same message. Typically, for a protocol to be correct, each such bundle must contain one strand for each of the legitimate principals apparently participating in this session, all agreeing on the principals, nonces, and session keys [9,10,11]. Penetrator strands or stray legitimate strands may also be entangled in a bundle, even in a correct protocol, but they should not prevent the legitimate parties from agreeing on the data values, or from maintaining the secrecy of the values chosen. One may think of a bundle as collecting all of the activities that were relevant to one run of a protocol, although the definition allows a bundle to contain additional events that need not have been strictly relevant. A strand is a linear structure, a sequence of one principal's message transmissions and receptions. A bundle is a graph-structured entity, representing the communication among a number of strands.

Protocol correctness typically depends essentially on the freshness of data items such as nonces and session keys. For this reason, the strand spaces that concern us are not full, in the sense that they do not contain all the strands that would arise if all possible data items were used. Presumably, the useful lifetime of a protocol is much shorter than the length of time that would be needed for the principals to use every possible session key or random value, and indeed we may reasonably assume that values of these kinds will be invented only once during the lifetime of the protocol.

A strand space models the assumption that some values occur only freshly by including only one strand originating that data item by initially sending a message containing it. Many strands, by contrast, may stand ready to combine with the originating strand by receiving the message and processing its contents further. A strand space will also model the assumption that some values are impossible for a penetrator to guess; in essence, the space simply lacks any penetrator strand in which this value is sent without having first been received.

Strand space theory presented here has several advantages. It gives a clear semantics to the assumption that certain data items, such as nonces and session keys, are fresh, and never arise in more than one protocol run.

Generally speaking, strand space theory has several advantages.

It gives a clear semantics to the assumption that certain data items, such as nonces and session keys, are fresh, and never arise in more than one protocol run.

We work with an explicit model of the possible behaviors of a system penetrator; this allows us to develop general theorems that bound the abilities of the penetrator, independent of the protocol under study.

It allows various notions of correctness, involving both secrecy and authentication, to be stated and proved.

It leads to detailed insight into the reasons why the protocol is correct, and the assumptions required. Proofs are simple and informative: they are easily developed by hand, and they help to identify more exact conditions under which we can rely on the protocol.

2 STRAND SPACES

2.1 Basic Notions

Consider a set A, the elements of which are the possible messages that can be exchanged between principals in a protocol. We will refer to the elements of A as terms. We will later (Section 2.3) impose more algebraic structure on the set A, but in this section we assume only that a subterm relation is defined on A. t0 ⊂ t1 means t0 is a subterm of t1.

In a protocol, principals can either send or receive terms. We will represent transmission of a term as the occurrence of that term with positive sign, and reception of a term as its occurrence with a negative sign.

Definition 2.1 A signed term is a pair 〈σ,a〉 ,with a ∈ A, and σ one of the symbols +,-. We will write a signed term as +t or -t. $(\pm A)^*$ is the set of finite sequences of signed terms. We will denote a typical element of $(\pm A)^*$ by $\langle\langle\sigma_1,a_1\rangle ,\ldots, \langle\sigma_n,a_n\rangle\rangle$. By abuse of language, we will still treat signed terms as ordinary terms. For instance, we shall refer to subterms of signed terms.

Definition 2.2 A strand space over A is a set Σ together with a trace mapping $tr:\Sigma\rightarrow(\pm A)^*$.

We will usually represent a strand space by its underlying set of strands Σ. In particular applications of the theory, the trace mapping need not be injective. We may want to distinguish between various instances of the same trace; for example, we may need to distinguish identical traces occurring at different times to model replay attacks.

Definition 2.3 Fix a strand space Σ

1. A node is a pair 〈s,i, with s ∈Σ and i an integer satisfying 1≤i≤length(tr(s)). The set of nodes is denoted by N. We will say the node 〈s,i〉 belongs to the strand s. Clearly, every node belongs to a unique strand.
2. If n = 〈s,i〉 ∈N then index(n) = i and strand(n) = s. Define term(n) to be (tr(s))*i*, i.e. the *i*th signed term in the trace of s. Similarly, uns_term(n) is ((tr(s))i)2, i.e. the unsigned part of the *i*th signed term in the trace of s.
3. There is an edge n1→n2 if and only if term(n1) = +a and term(n2) = -a for some a ∈ A. Intuitively, the edge means that node n1 sends the message a, which is received by n2, recording a potential causal link between those strands.
4. When n1 = 〈s, i〉 and n2 = 〈s,i+1〉 are members of N, there is an edge n1 ⇒ n2. Intuitively, the edge expresses that n1 is an immediate causal predecessor of n2 on the strand s. We write $n0 \Rightarrow^+ n$ to mean that n0 precedes n (not necessarily immediately) on the same strand.
5. An unsigned term t occurs in n ∈ N iff t ⊂ term(n).
6. Suppose I is a set of unsigned terms. The node n ∈ N is an entry point for I iff term(n) = +t for some t ∈ I, and whenever $n0 \Rightarrow^+ n$, term(n0) ∉ I.
7. An unsigned term t originates on n ∈ N iff n is an entry point for the set I = {t0 : t ⊂ t0}.
8. An unsigned term t is uniquely originating iff t originates on a unique n ∈ N.

If a term t originates uniquely in a particular strand space, then it can play the role of a nonce or session key in that structure.

N together with both sets of edges n1 → n2 and n1 ⇒ n2 is a directed graph 〈N, (→∪⇒).

2.2 Bundles and Causal Precedence

A bundle is a finite subgraph of this graph, for which we can regard the edges as expressing the causal dependencies of the nodes.

Definition 2.4 Suppose $\rightarrow_C \subset \rightarrow$; suppose $\Rightarrow_C \subset \Rightarrow$; and suppose $C = \langle N_C, (\rightarrow_C \cup \Rightarrow_C)\rangle$ is a subgraph of 〈N, (→∪⇒)〉. C is a bundle if:

1. C is finite.
2. If $n2 \in N_C$ and term (n2) is negative, then there is a unique n1 such that $n1 \rightarrow_C n2$.
3. If $n2 \in N_C$ and n1 ⇒ n2, then $n1 \Rightarrow_C n2$.
4. C is acyclic.

In conditions 2 and 3, it follows that $n1 \in N_C$, because C is a graph.

For our purposes, it does not matter whether communication is regarded as a synchronizing event or as an asynchronous activity. This definition formalizes a process communication model with three properties:

- A strand (process) may send or receive a message, but not both at the same time;
- When a strand receives a message m, there is a unique node transmitting m from which the message was immediately received;
- When a strand transmits a message m, many strands may immediately receive m.

Notational Convention 2.5 A node n is in a bundle $C = \langle N_C, \rightarrow_C, \Rightarrow_C\rangle$, written n ∈ C, if n ∈ NC; a strand s is in C if all of its nodes are in NC. If C is a bundle, then the C-height of a strand s is the largest i such that 〈s,i〉 ∈C. C-trace(s) = 〈tr(s)(1),…, tr(s)(m)〉 , where m = C-height(s).

Definition 2.6 If S is a set of edges, i.e. S $\subset \rightarrow \cup \Rightarrow$, then π_S is the transitive closure of S, and $\leq$S is the reflexive, transitive closure of S.

The relations π_S and $\leq$S are each subsets of NS × NS, where NS is the set of nodes incident with any edge in S.

Lemma 2.7 Suppose C is a bundle. Then $\leq_C$ is a partial order, i.e. a reflexive, antisymmetric, transitive relation. Every non-empty subset of the nodes in C has $\leq_C$ -minimal members.

We regard $\leq_C$ as expressing causal precedence, because n π_S n0 holds only when n's occurrence causally contributes to the occurrence of n0. When a bundle C is understood, we will simply write $\leq$. Similarly, “minimal” will mean $\leq_C$ -minimal.

The existence of minimal members in non-empty sets serves as an induction principle, an observation that clarifies the relation of our approach to Paulson's and Schneider's [12,13].

Most of our arguments turn on the $\leq_C$ -minimal elements in some set of nodes. These arguments are motivated by the question, "What did he know, and when did he know it?"

Lemma 2.8 Suppose C is a bundle, and S $\subseteq$ C is a set of nodes such that $\forall$ m,m0. uns_term(m) = uns_term(m0) implies (m $\in$ S iff m0 $\in$ S) If n is a $\leq_C$ -minimal member of S, then the sign of n is positive.

Proof. If term(n) were negative, then by the bundle property, n0 →n for some n0 $\in$ C and uns_term(n) = uns_term(n0). Hence, n0 $\in$ S, violating the minimality property of n. ■

Lemma 2.9 Suppose C is a bundle, t 2 A and n 2 C is a $\leq_C$ -minimal element of {m $\in$ C : t $\sqsubset$ term(m)}. The node n is an originating occurrence for t.

Proof. Because n is a member, t $\sqsubset$ term(n). By Lemma 2.8, the sign of n is positive. If n0 $\Rightarrow^+$ n, then applying Definition 2.4, Clause 3 as many times as necessary, n0$\in$ C. Hence by the minimality property of n, t $\not\sqsubset$ term(n0). Thus n is originating for t. ■

2.3 Terms and Encryption

We will now specialize the set of terms A. In particular we will assume given:

A set T $\subseteq$ A of texts (representing the atomic messages).

A set K $\subseteq$ A of cryptographic keys disjoint from T, equipped with a unary operator inv : K→ K.

We assume that inv is injective; that it maps each member of a key pair for an asymmetric cryptosystem to the other; and that it maps a symmetric key to itself.

Two binary operators

encr : K × A → A

join : A × A → A

We will follow custom and write inv(K) as K^{-1}, encr(K,m) as $\{m\}_K$, and join(a, b) as ab. If k is a set of keys, K^{-1} denotes the set of inverses of elements of k.

2.4 Freeness Assumptions

The proofs in this paper use an assumption we will call the assumption of free encryption; many other authors (e.g. [14, 15,16]) have made similar assumptions, dating back to Dolev and Yao [17], although not all have [18]. It stipulates that a ciphertext can be regarded as a ciphertext in just one way:

Axiom 1 For m,m′$\in$A and k,k′$\in$K

$$\{m\}_k=\{m'\}_{k'} \Longrightarrow \ m=m' \text{且} \ k=k'$$

For clarity of exposition we make a stronger assumption in this paper, namely that A is the algebra freely generated from T and K by the two operators encr and join, as embodied in Axiom 2.

Axiom 2 For m_0,m_0', m_1,$m_1'$$\in$A and k,k′$\in$K,

(1) $m_0m_1= m_0'm_1' \Longrightarrow \ m_0=m_0' \text{且} \ m_1=m_1'$

(2) $m_0m_1 \neq \{m_0'\}_k$

(3) $m_0m_1 \notin$ K Y T

(4) $\{m_0\}_k \notin$ K Y T

This is more than is needed for our method but it leads to the simplest exposition of the main points. In [19] we showed how to weaken this assumption considerably, accounting for the possibility that the join operator is associative, for instance.

Given Axiom 2, we may define the width of terms:

Definition 2.10 If m$\in$K Y T or if m=$\{m_0\}_k$, then width(m) = 1. If m= m_0m_1, then width(m)= width(m_0) + width(m_1).

For any encryption algebra A there is a free encryption algebra A′ and a surjective algebra morphism π: A′→A. Moreover, π and A′ are unique to within isomorphism, this being effectively the definition of free algebra in the theory of universal algebras. In this paper we have shown protocol correctness results for strand spaces over the free algebra A′. Now it is easy to see that if a protocol property fails for strands over A0, then the same protocol property fails for A. However, the converse is not true, since protocol failures may exploit relations in the algebra A that cannot be lifted to A′. Nevertheless, much useful information can be obtained by considering the free message algebra, since we are thereby excluding vulnerabilities based on the structure of the protocol itself, rather than on particular properties of the message algebra. The problem remains to determine which relations among the elements of the free algebra A′ will preserve a protocol correctness result. This is a hard problem, which will doubtless require much future work exploring different approaches. Since we have assumed that our message algebra A is freely generated, we can use a simple inductive definition of the subterm relation.

Definition 2.11 The subterm relation $\sqsubset$ is defined inductively, as the smallest relation such that:

a $\sqsubset$ a;

a $\sqsubset$ $\{g\}_K$ if a$\sqsubset$g;

a $\sqsubset$ gh if a$\sqsubset$g or a$\sqsubset$ h.

We should emphasize that, for $K$$\in$K, $K$$\sqsubset$ $\{g\}_K$ only if $K$$\sqsubset$g already. Restricting subterms in this way reflects an assumption about the penetrator's capabilities: that key can be obtained from ciphertext only if they are embedded in the text that was encrypted. This might not always be the case- for instance, if a dictionary attack is possible-but it is the assumption we will make here.

This notion of subterm does not always mesh perfectly with the definition of origination and unique origination, which refers to the subterm relation (Definition 2.3, Clauses 7 and 8).

An immediate consequence of the freeness assumption and the inductive definition of subterm are:
Proposition 2.12 Suppose $K \neq K'$ and $\{h'\}_{K'} \sqsubset \{h\}_K$. Then $\{h'\}_{K'} \sqsubset h$.

3 THE PENETRATOR

The penetrator's powers are characterized by two ingredients, namely a set of keys known initially to the penetrator and a set of penetrator strands that allow the penetrator to generate new messages from messages he intercepts.

A penetrator set consists of a set of keys KP. It contains the keys initially known to the penetrator. Typically it would contain: all public keys; all private keys held by the penetrator or his accomplices; and all symmetric keys Kpx; Kxp initially shared between the penetrator and principals playing by the protocol rules. It may also contain "lost keys" that became known to the penetrator previously, perhaps because he succeeded in some cryptanalysis.

3.1 Penetrator Strands

The atomic actions available to the penetrator are encoded in a set of penetrator traces. They summarize his ability to discard messages, generate well-known messages, piece messages together, and apply cryptographic operations using keys that become available to him. A protocol attack typically requires hooking together several of these atomic actions.

Definition 3.1 A penetrator trace is one of the following:

M. Text message: $\langle +t \rangle$, where $t \in T$.
F. Flushing: $\langle -g \rangle$.
T. Tee: $\langle -g, +g, +g \rangle$.
C. Concatenation: $\langle -g, -h, +gh \rangle$.
S. Separation into components: $\langle -gh, +g, +h \rangle$.
K. Key: $\langle +k \rangle$, where $k \in K_P$
E. Encryption: $\langle -K, -h, +\{h\}_K \rangle$.
D. Decryption: $\langle -K^{-1}, -\{h\}_k, +h \rangle$.

This set of penetrator traces gives the penetrator powers similar to those in other approaches, e.g. [14, 16]. They ensure that the values that may be emitted by the penetrator are closed under joining, encryption, and the relevant "inverses".

It is also possible to extend the set of penetrator traces given here if it is desired to model some special ability of the penetrator. That requires no essential change to our overall framework, although the proofs in this paper would then need to be modified to take account of the additional penetrator traces. Our theorems characterize a penetrator with just the powers we have described; a penetrator with additional computational or cryptanalytic abilities may not be subject to the same limitations.

One example of an extended penetrator would be a penetrator who can cryptanalyze old session keys, and thus benefit from some kinds of replay attacks [3]; the penetrator we formalize here does not have this ability.

Definition 3.2 An infiltrated strand space is a pair (Σ;P) with Σ a strand space and $P \subseteq \Sigma$ such that tr(p) is a penetrator trace for all $p \in P$. A strand $s \in \Sigma$ is a penetrator strand if it belongs to P, and a node is a penetrator node if the strand it lies on is a penetrator strand. Otherwise we will call it a non-penetrator or regular strand or node.

A node n is an M, F, etc. node if n lies on a penetrator strand with a trace of kind M, F, etc.

We would not expect an infiltrated strand space to realize all of the penetrator traces of type M. In that case, the space could not model unguessable nonces. It is usually assumed that the space Σ lacks M-strands for many text values, which regular participants can use for fresh nonces. In the remainder of this paper, we will examine infiltrated strand spaces in which the regular strands all belong to a single protocol.

4 NOTIONS OF CORRECTNESS

Gavin Lowe studies a range of authentication properties in [9]; strand spaces are a natural model for stating and proving his agreement properties, which are akin to the correspondence properties of Woo and Lam [20].

A protocol guarantees agreement to a participant B (say, as the responder) for certain data items ~x if: each time a principal B completes a run of the protocol as responder using ~x, which to B appears to be a run with A, then there is a unique run of the protocol with the principal A as initiator using ~x, which to A appears to be a run with B.

For a regular strand in an authentication protocol, the principal engaging in that strand, as well as the apparent interlocutor, can be inferred from the contents of the terms occurring in the strand.

A weaker non-injective agreement does not ensure uniqueness, but requires only: each time a principal B completes a run of the protocol as responder using ~x, apparently with A, then there is a run of the protocol with the principal A as initiator using ~x, apparently with B. Non-injective agreement is weaker because it does not prevent the other party A from being duped into executing multiple runs matching a single run by B.

We can prove non-injective agreement by establishing that, whenever a bundle C contains a strand representing a responder run using ~x, then C also contains a strand representing an initiator run that corresponds in the sense that it also uses ~x. We can establish agreement by showing that C contains a unique initiator strand using ~x.

We will also state a simple notion of secrecy for a data value x, which will be sufficient for our purposes here. A value x is secret in a bundle C if for every n ∈ C, term (n) ≠ x.

This notion of secrecy concerns only what is "said on the wire." In this sense, a value is secret if the regular strands never emit it, and the penetrator can never emit it. Regular protocol participants may "know" a secret value in the sense of carrying out computations that depend on it, so long as their behavior in the protocol does not include disclosing it in public.

Moreover, if we prove that the penetrator never emits a value, it follows that he can never derive it from values he receives: for if he derived it, then he would be capable of emitting it. The penetrator strands defined in Definition 3.1 correspond to the ways that a penetrator would derive new values from those he already possesses. For instance, if the penetrator received a value g x, then an S-strand would lead to the penetrator emitting the supposed secret x.

More stringent notions of secrecy are also possible, as for instance information flow security properties, and may be fruitfully applied to security protocols [21].

5 ANALYSIS OF SECURE ELECTRONIC TRANSACTION (SET) PROTOCOL USING STRAND SPACE

In this section, we will formalize the transaction procedure, then we use strand space theory to analyze it. Because we only consider secrecy, the delivery of certificate will be omitted for concision.

The interpretation of symbols:
C: card holder; M: merchant; P: payment gateway;
k_i,i∈{1,2,3,4,5,6}:new session key which is generated stochasticly;
C.acc: accounts of cardholder
K_C^{-1} ：the private key of card holder used in signature；
$K_{M.sig}^{-1}$ ：the private key of merchant used in signature；
$K_{M.sig}$ ：the public key of merchant used in signature；
$K_{M.data}^{-1}$ ：the private key of merchant used in data exchange；
$K_{M.data}$ ：the public key of merchant used in data exchange;
$K_{P.sig}^{-1}$ ：the private key of payment gateway used in signature；
$K_{P.sig}$ ：the public key of payment gateway used in signature；
$K_{P.data}^{-1}$ ：the private key of payment gateway used in data exchange；
$K_{P.data}$ ：the public key of payment gateway used in data exchange；

OI: order information; PI: payment information; deal.res: deal respond;
init.req: initial request; init.res: initial respond;
auth.req: authentication request; auth.res: authentication respond;
cap.req: capture request; cap.res: capture respond.

5.1 The Phase of Transaction Request

$$\mathrm{C} \to \mathrm{M}: \mathrm{init.req}$$
$$\mathrm{M} \to \mathrm{C}: \mathrm{init.res}, \{\mathrm{h(init.res)}\}_{\mathrm{K}_{\mathrm{M.sig}}^{-1}}$$
$$\mathrm{C} \to M: \mathrm{h(PI), OI, \{h(h(PI),\ h(OI))\}}_{\mathrm{K}_\mathrm{C}^{-1}}, \{k_1, c.acc\}_{K_{P.data}},$$
$$\{PI, \{h(h(PI), h(OI))\}_{K_C^{-1}}, h(OI)\}_{k_1}$$
$$\mathrm{M} \to \mathrm{C}: \mathrm{deal.res}, \{\mathrm{h(deal.res)}\}_{\mathrm{K}_{\mathrm{M.sig}}^{-1}}$$

The protocol defines two roles: the card holder and the merchant. Their Strands are:
(1) The strand of card holder: it's message sequence Init[C,M] is:

$$\langle +\mathrm{init.req}\quad -\{\mathrm{init.res}, \{\mathrm{h(init.res)}\}_{\mathrm{K}_{\mathrm{M.sig}}^{-1}}\},$$
$$+\{\mathrm{h(PI), OI, \{h(h(PI),\ h(OI))\}}_{\mathrm{K}_\mathrm{C}^{-1}}, \{k_1, c.acc\}_{K_{P.data}},$$
$$\{PI, \{h(h(PI), h(OI))\}_{K_C^{-1}}, h(OI)\}_{k_1}\}$$
$$-\{\mathrm{deal.res}, \{\mathrm{h(deal.res)}\}_{\mathrm{K}_{\mathrm{M.sig}}^{-1}}\}\rangle$$

(2) The strand of merchant: it's message sequence Resp[C,M] is

$$\langle -\mathrm{init.req}\quad +\{\mathrm{init.res}, \{\mathrm{h(init.res)}\}_{\mathrm{K}_{\mathrm{M.sig}}^{-1}}\},$$
$$-\{\mathrm{h(PI), OI, \{h(h(PI),\ h(OI))\}}_{\mathrm{K}_\mathrm{C}^{-1}}, \{k_1, c.acc\}_{K_{P.data}},$$
$$\{PI, \{h(h(PI), h(OI))\}_{K_C^{-1}}, h(OI)\}_{k_1}\}$$
$$+\{\mathrm{deal.res}, \{\mathrm{h(deal.res)}\}_{\mathrm{K}_{\mathrm{M.sig}}^{-1}}\}\rangle .$$

Figure 1 The Strand of Cardholder in Transaction Request.

Now we prove that the session key k_1 is a secret in SET protocol.
Proposition 1 Suppose:
1. ∑ is the strand space of transaction request;C is a bundle including card holder; s is defined as Init[C,M].
2. $\mathrm{K}_{\mathrm{P.data}}^{-1}, \mathrm{k}_1 \notin \mathrm{K}_\mathrm{P}$
3. k_1 originates uniquely on n_2

then for all the nodes in C, if $k_1 \subset$ term(m), then $\{\mathrm{k}_1, \mathrm{c.acc}\}_{\mathrm{K}_{\mathrm{P.data}}} \subset \mathrm{term(m)}$

Proof: Consider the set
$$\mathrm{F} = \{\mathrm{n} \in \mathrm{C} : \mathrm{k}_1 \subset \mathrm{term(n)} \wedge \{\mathrm{k}_1, \mathrm{c.acc}\}_{\mathrm{K}_{\mathrm{P.data}}} \not\subset \mathrm{term(n)}\} \cdot$$
Let's suppose F isn't empty, then F has at least one ≤-minimal element. Next, we will prove these ≤-minimal elements are neither regular nodes nor penetrator nodes. Therefore F is empty, and the proposition holds.

1. Assume m ∈ F is minimal and a regular node. The sign of m is positive by Lemma 2.8.

$$\Theta\, \mathrm{k}_1 \neq \mathrm{init.req} \wedge \{\mathrm{k}_1, \mathrm{c.acc}\}_{\mathrm{K}_{\mathrm{P.data}}} \subset \mathrm{term(n_2)}$$
$$\therefore \mathrm{m} \notin \mathrm{s}$$

But k_1 originates uniquely on n_2, so m isn't on other strand

$s' \neq s$.
So, m couldn't be a regular node.

2. The $\leq$-minimal elements of F couldn't be penetrator nodes. We will examine every possible penetrator strand:

M. The strand has the form $\langle +t \rangle$ where $t \in T$, if $k_1 \subset t$, then $k_1=t$. So, k_1 should originate from this strand. But that is impossible, as k_1 originates uniquely on the regular node n_2.

F. The strand has the form $\langle -g \rangle$, and thus lacks any positive nodes.

T. The trace tr(p) has the form $\langle -g, +g, +g \rangle$, so the positive nodes are not minimal occurrences.

C. The trace tr(p) has the form $\langle -g,-h, +gh \rangle$, so the positive node is not a minimal occurrence.

S. The trace tr(p) has the form $\langle -gh, +g, +h \rangle$. Assume m=node(+g); there is a symmetrical case if m=node(+h). Since $k_1 \subset g$, $k_1 \subset gh$,and $\{k_1, c.acc\}_{K_{P.data}} \not\subset h$, $\{k_1, c.acc\}_{K_{P.data}} \not\subset gh$, $gh \in F$, contradicting the minimality of m.

K. The trace tr(p) has the form $\langle -k \rangle$ where $K \in K_P$. But $k_1 \notin k_p$, so this case does not apply.

E. The trace tr(p) has the form $\langle -K,-h, +\{h\}_K \rangle$. Suppose m=node($+\{h\}_K$) . Since $k_1 \neq \{h\}_K$, $k_1 \subset h$. Thus node($-h$)$\in$F. So the positive node is not minimal in S.

D. The trace tr(p) has the form $\langle -K^{-1},-\{h\}_k, +h \rangle$. Since $K_{P.data}^{-1} \notin K_P$, $k_1 \notin h$. But +h is the only positive term, so m couldn't be on this kind of penetrator strand.■

Proposition 1 illustrates the occurrences of k_1 in the strand space of transaction request can only take the form of encryption, so k_1 is a secret.

Similarly, we can prove C.acc is a secret.

6 CONCLUSIONS

Strand space theory is a new approach of security protocol. It can provide more details of protocol. According to the procedure of SET protocol analysis, we can find SET is very complicated and its analysis isn't easy.

7 REFERENCES

[1] F.Javier Thayer Fabrega, Jonathan C.Herzog and Joshua D. Guttman. Strand Spaces: Proving Security Protocols correct. Journal of Computer Security, Vol. 7, pp. 191-230,1999.

[2] Roger Needham and Michael Schroeder. Using encryption for authentication in large networks of computers. Communications of the ACM, 21(12), December 1978.

[3] Dorothy Denning and G. Sacco. Timestamps in key distribution protocols. Communications of the ACM, 24(8), August 1981.

[4] Gavin Lowe. An attack on the Needham-Schroeder public key authentication protocol. Information Processing Letters, 56(3):131~136, November 1995.

[5] Judy H. Moore. Protocol failures in cryptosystems. Proceedings of the IEEE, 76(5), May 1988.

[6] John Clark and Jeremy Jacob. On the security of recent protocols. Information Processing Letters, 56(3):151~155, November 1995. 40

[7] Sarvar Patel. Number theoretic attacks on secure password schemes. In Proceedings of the 1997 IEEE Symposium on Security and Privacy, pages 236~247. IEEE Computer Society Press, May 1997.

[8] C. A. R. Hoare. Communicating Sequential Processes. Prentice-Hall International, Englewood Cliffs, New Jersey, 1985.

[9] Gavin Lowe. A heirarchy of authentication speci_cations. In 10th Computer Security Foundations Workshop Proceedings, pages 31~43. IEEE Computer Society Press, 1997.

[10] A. W. Roscoe. Intensional specifications of security protocols. In Proceedings of the 9th IEEE Computer Security Foundations Workshop, pages 28~38, 1996.

[11] Thomas Y. C. Woo and Simon S. Lam. Verifying authentication protocols: Methodology and example. In Proc. Int. Conference on Network Protocols,October 1993. 43

[12] Lawrence C. Paulson. The inductive approach to verifying cryptographic protocols. Journal of Computer Security, 1998. Also Report 443, Cambridge University Computer Lab.

[13] Steve Schneider. Verifying authentication protocols with CSP. In Proceedings of the 10th IEEE Computer Security Foundations Workshop, pages 3~17. IEEE Computer Society Press, 1997.42

[14] Gavin Lowe. Casper: A compiler for the analysis of security protocols. In 10th Computer Security Foundations Workshop Proceedings, pages 18~30. IEEE Computer Society Press, 1997.

[15] Will Marrero, Edmund Clarke, and Somesh Jha. A model checker for authentication protocols. In Cathy Meadows and Hilary Orman, editors, Proceedings of the DIMACS Workshop on Design and Verification of Security Protocols. DIMACS, Rutgers University, September 1997.41

[16] Lawrence C. Paulson. Proving properties of security protocols by induction. In 10th IEEE Computer Security Foundations Workshop, pages 70~83. IEEE Computer Society Press, 1997.

[17] D. Dolev and A. Yao. On the security of public-key protocols. IEEE Transactions on Information Theory, 29:198~208, 1983.

[18] Shimon Even, Oded Goldreich, and Adi Shamir. On the security of ping-pong protocols when implemented using the RSA. In Advances in Cryptology crypto '85, LNCS, pages 58~72. Springer Verlag, 1985.

[19] F. Javier Thayer F_abrega, Jonathan C. Herzog, and Joshua D. Guttman. Honest ideals on strand spaces. In Proceedings of the 11th IEEE Computer Security Foundations Workshop. IEEE Computer Society Press, June 1998.

[20] Thomas Y. C. Woo and Simon S. Lam. Verifying authentication protocols: Methodology and example. In Proc. Int. Conference on Network Protocols,October 1993. 43

[21] Riccardo Focardi and Roberto Gorrieri. The compositional security checker: A tool for the verification of information ow security properties. IEEE Transactions on Software Engineering, 23(9), September 1997.

Xu Feng was born in 1975. Now he is a Ph.D. candidate in the department of computer science, Nanjing University. His research interests concentrate on security protocols, wireless network and mobile security.

Robust Hash Used In the Application of Digital Image Signature

Wu Jin[1, 2], Qiu Ya[2], Huang Honglin[2], Liu Jian[1], Tian Jinwen[1]
[1] **Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology,**
[2]**College of Information Science and Engineering, Wuhan University of Science and Technology,**
Wuhan, Hubei, 430081, China
Email: hust_wu@163.com, www.qiuya@126.com **TEL**: +86 (0) 27 62000502

ABSTRACT

The destroyers can use various methods to attack the digital watermarking; it's hard to protect our intellectual property rights. But by using hash function, we can detect deformations (such as rotation and scaling) applied to original image or against some image processing basic attacks (like compression, filtering, blurring, etc.). In this paper, we present a soft hash algorithm to detect digital image signature. The theory and the experiments has approved that it is a robust algorithm.

Keywords: hash function, copyright, pattern recognition, Radon transform, digital signature, attacks.

1. INTRODUCTION

With the development of digital technology and Internet, various forms of the multimedia-digital works (image, video, audio-file, etc.) are issued in the form of network one after another, and its copyright protection becomes a problem needing to be solved urgently. Because the digital watermark is the effective method of realizing copyright protection, and now it already becomes a focus of the research field about multimedia information safety, at the same time, it is the important branch of the information hiding technology research field.

The digital watermarking is used to protect the intellectual property rights in the multimedia field and E-commerce applications, and it's also used to seek where the digital works is used and to judge whether the use of this intellectual property is legal or illegal. There are many methods used in identifying images securely. Usually hash functions are widely used in this section.

There are two types of hash functions: one is Keyed Hash function; the other is No-Keyed Hash function. The digital signature is based on a hash function – No-Keyed Hash function that is a one-way function, and an encryption algorithm. These types of functions are collision resistant. For example, MD5 [1], SHA1 [2] are customized compression function in cryptographic process. A typical hash function requires the following properties [3], [4]:

[1] If the message M is given, it is easy to compute H(M);
[2] It is hard to find the message M if H(M) is given;
[3] If a message M is given, it is hard to find another message M', such that H(M) =H(M');
[4] It is hard to find two random messages M and M', such that H(M) = H(M');

It is easy to implement the hash function in hardware and in software.

For image application [5], [6], the second property is relevant but the first point need to be corrected in Two different images must have two different message digests. For each document, the digital signature algorithm provides a unique output bits stream. Two images are different if and only if image contents are different. There is impossible existed the same signature which come from two different messages.

In a digital watermark system, an attack is defined as any attempt to remove or isolate the watermark. There are several intentional and unintentional attacks. The goal of the intentional attacks is to remove the owner's watermark and/or to insert a fake one. The unintentional attacks may appear due to the usual signal processing such as: JPEG compression, MPEG-2 compression, filtering, pixel quantisation, etc.

In [6], [7] and [8] the attacks which are specific to a digital watermarking system are classified in:

- ♦ simple attack;
- ♦ detection disabling;
- ♦ removal attack;
- ♦ ambiguity attack

The simple attack (waveform attack, noise attack) produces some distortions of the watermarked data but not in order to remove the watermark. The second type of attacks tries to avoid the watermark detection [9]. The removal attack (collusion attack) is focused directly to the watermark, trying to remove it [10]. The most dangerous kind of attack is the ambiguity attack also known as deadlock, inversion or IBM attack.

In this paper, we want to deal with the first type attack – the simple attack. It contains geometrical transformation (rotation and scaling) and image processing attacks (compression, filtering, blurring).

The Radon transformation largely used in medical image processing [11] provides a good basis for the algorithm. In fact, this transformation is robust against image processing such as sharpening, blurring, adding noise, compression, and has some invariant properties with regards to geometrical transformations such as rotation and scaling. From Radon transformation, some robust and almost invariant elements can be extracted. In the next section, we describe our robust and invariant hash function for images.

2. RADON TRANSFORM

In tomography, when a bundle of X-Rays goes through an organ, its attenuation depends on content of organ, distance, and direction or angle of this projection. This set of projections is called Radon transform.

As we show in figure 1, in two dimensions, the process can be described like this:

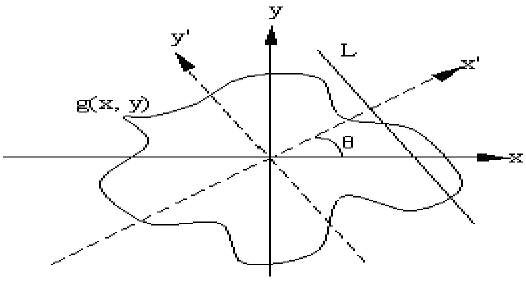


Fig 1 Radon Transform projections

The set of projections can be transmitted from any angle θ, generally f(x, y)'s Radon transform is the integral along with the direction which parallel y' axis (As it is showed in Figure 1). Its mathematic expression is:

$$Rg(x',\theta)=\int_{-\infty}^{+\infty} g(x'\cdot\cos\theta-y'\cdot\sin\theta, x'\cdot\sin\theta+y'\cdot\cos\theta)dy' \quad (1)$$

As we descript just now, the mathematical expression of Radon transforms leads to some very useful properties.

♦ If an image g is rotated by Φ, its Radon transform is

$$g(x\cdot\cos\phi-y\cdot\sin\phi, x\cdot\sin\phi+y\cdot\cos\phi)\leftrightarrow Rg(x',\theta+\phi) \quad (2)$$

♦ If an image g is shifted by (x0, y0), the Radon transform is

$$g(x-x_0, y-y_0)\leftrightarrow Rg(x'-x_0\cdot\cos\theta-y_0\cdot\sin\theta) \quad (3)$$

♦ If an image g is scaled by a factor α, its Radon transform is

$$g(\alpha\cdot x,\alpha\cdot y)\leftrightarrow \frac{1}{|\alpha|}\cdot Rg(\alpha\cdot x',\theta) \quad (4)$$

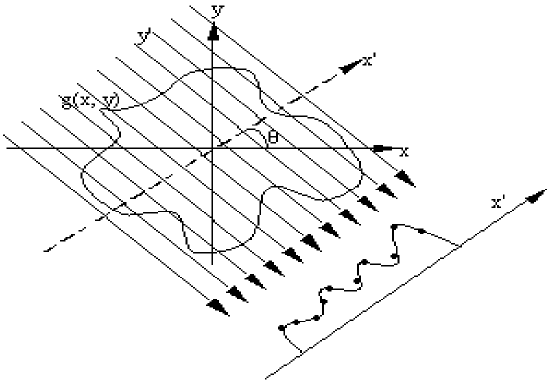


Fig 2 Line Integral of Radon

If we get the integral of all pixels, which are contained in the image, and each pixel must have 180 groups' data because we should take its integral from 1 to 180 degree. So there must have lots of data to deal with. But the data is too big. (You can see [12], they deal with all of the data.) In the next section, we'll do it in another method.

3. HASH FUNCTION

The image content is better described by the variation of the projections rather than the projection themselves. When the projection-angle is fixed, we can get a set of data $v_\theta(i)$– a vector – that depends on the size of the image. Now, in order to give us a convenience of describing the image, we call this vector v_θ Radon vectors.

When we transform the image in the geometrical filed (rotation and scaling) and use the methods of compression, filtering, blurring to attack the image, there is a point in the image should be loss a little data if we get the Radon transform for it. That point is the center of projection of the image. And it is the right point - invariant and robust point - that we want to find. We describe the center of the Radon vector v_θ for each θ as this point.

The figure that is showed below is the invariant-and-robust point's Radon transform.

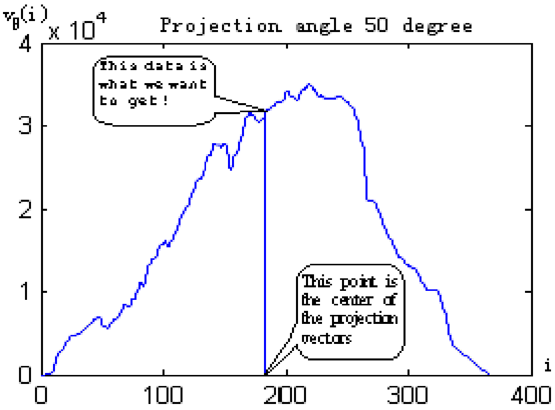
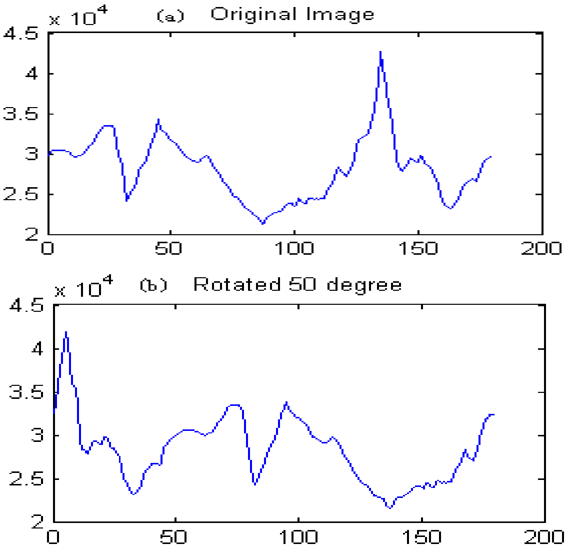


Fig 3 Projection with θ=50° for original image of Lena

4. EXPERIMENTS

In order to validating what we said above, we desire some experiments. All color images are transformed into 8 bits/pixel gray level images. The results like this:



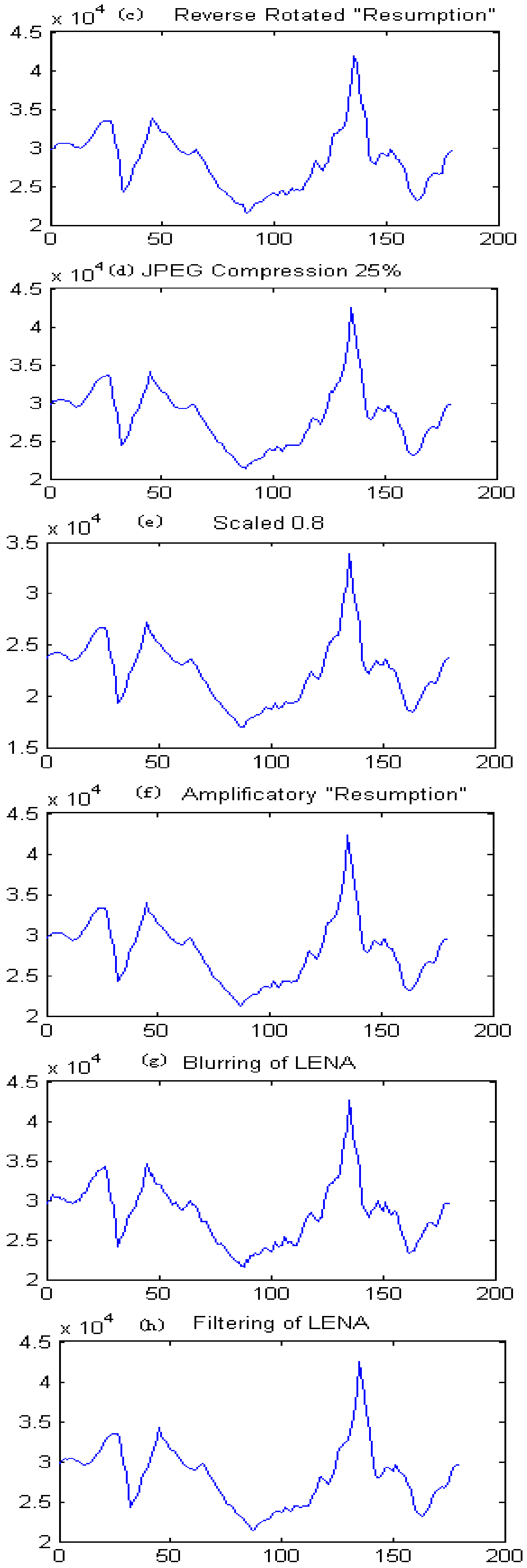


Fig 4 Radon projection for several transformations of Lena and its "Resumption"

There are some equations, we can use them to detect the scaling and the degree of rotation comparing to original image.

♦ If the image is scaled by a factor α, α can be calculated by this:

$$\alpha = \sqrt{\frac{Energy(hash_{original})}{Energy(hash_{scaled})}} \tag{5}$$

♦ If the image is rotated by Φ, now Φ can be calculated by this:

$$\phi = 180 - \arg\max_m (R_{xy}(m)) \tag{6}$$

$$R_{xy}(m) = \sum_{n=0}^{d-m-1} (x_n - \frac{1}{d} \cdot \sum_{i=0}^{d-1} x_i) \cdot (y_{n+m} - \frac{1}{d} \cdot \sum_{i=0}^{d-1} y_i) \tag{7}$$

where d is the length of the signature. When the two signatures come from images with the same content, $R_{xy}(m)$ will be close to 1 for a certain m*. So, at now, we can calculate the angle Φ.

The Mean Square Error (MSE) determines if the signatures come from images with the same content:

$$MSE = \frac{\sum_{i=0}^{d-1} (x_i - y_i)}{d}. \tag{8}$$

After normalization, the MSE of some experiments is showed in the table:

Comparing Lena	MSE
Lena-scale (0.8)	3.3616×10^{-6}
Lena-sharpen*2	8.4752×10^{-5}
Lena-blur*2	4.9016×10^{-5}
Couple	0.0292

From the blank above, we can see that our method is robust, invariant and efficient.

5. CONCLUSION

In this paper, the proposed soft Hash algorithm can easy show us whether the two images come from the same document. And this method not only is effective for geometrical transformation (including scale, rotation), but also can give a accurate decision for the intended or unintended destroys (such as compression, filter, etc.), therefore, it will protect our knowledge property rights effectually.

By far, we have confirmed the robust Radon transform and its robustness against image processing attacks. Because of its stability and uniqueness, it could be used in pattern recognition to retrieve an image in Internet, E-education applications, and in watermarking process as a synchronization block to detect and rectify geometrical deformations such as rotation and scaling.

However, its security and transparency have not been involved yet. We will make an effort in this field afterwards.

6. REFERENCES

[1] Ronald RIVEST "RFC 1321: The MD5 Message-Digest Algorithm." RSA Data Security Inc., April 1992.

[2] National Institute of Standards and Technology (NIST), "Announcement of Weakness in the Secure Hash Standard", 1994.

[3] B. Schneier, "Applied Cryptography", John Wiley and Sons, 1996.

[4] W. Stallings, "Network and Internetwork Security. Principles and Practice", Prentice Hall, 1995.

[5] Jiri FRIDRICH and Miroslav GOLJAN, "Robust Hash Functions for Digital Watermarking", ITCC 2000, Las Vegas, March 27-29, 2000, Nevada, USA.

[6] C.L. SABHARWAL and S.K. BHATIA. "Perfect Hash Table Algorithm for Image Databases Using Negative Associated Values." Pattern Recognition. 28:7. pp. 1091- 1101. July 1995.

[7] I. J. Cox, M. L. Miller, J. A. Bloom, "Digital Watermarking", Morgan Kaufmann Publishers, 2002.

[8] G. Wade, "Watermark Attacks", Research Report, Plymouth, Oct., 1999.

[9] I. Cox, J. Linnartz, "Public watermark and resistance to tampering", IEEE Int. Conf. On Image Processing, CDROM, 1997.

[10] D. Boneh, J. Shaw, "Collusion-Secure Fingerprinting for Digital Data, Advances in Cryptology", CRYPTO'95, Springer Verlag, pp. 452-465, 1995.

[11] F. Hartung, J.K. Su, B. Girod, "Spread Spectrum Watermarking: Malicious Attacks and Counterattacks", Proc. SPIE, Jan., Vol. 3957, 1999.

[12] Fr´ed´eric Lef`ebvre, Benoit Macq and Jean-Didier Legat, "RASH: Radon Soft Hash algorithm", 11th European Signal Processing Conference, september 3-6 2002, Toulouse, France.

An Adaptive Digital Watermarking Algorithm Based on Wavelet Transform*

Zhiqiang Guo, Xuemei Jiang, Quan Liu
School of information engineering, Wuhan University of technology
Wuhan, Hubei, China
Email: guozhiqiang@mail.whut.edu.cn **Tel**: 027-87658209

ABSTRACT

Propose an adaptive color image watermarking algorithm, there are several adaptive factor designed in it which can adept the embedding intensity of watermark according to the color character of image. On condition of the watermarking's invisible to realize the signal of watermarking embedding in the most degree. This watermarking algorithm, especially when watermarking image and original image are all colorful, helps the watermarking signal's robustness to be improved greatly.

Keywords: Watermarking, Wavelet transform, embedding intensity

1. INTROTUCTION

Today, as the digitization develops day by day, the protection of digital information becomes an urgent problem. In order to resist different kinds of infringement, a new information security technology that called watermarking had been put forward to in the international scope. This is a new technology which embeds certain digital information into digital file for the sake of protect copyright owner's benefit. Because of it's extensive application prospect in the fields of copyright protection, identification, private communication, hidden tag and so on, it had already been a very important research domain. In light of the difference among the carrier, watermarking could be mainly classified as image watermark, video watermark, sonic frequency watermark and text watermark. Particularly, digital image watermark is the research hotspot .Now the majority of popular watermarking algorithm are based on null area and transform area, view as the comparison of their general capability, transform area has its own advantage and take the predominant place also will be the mainstream in future trend[1]~ [3]. Author explains a transform area color image watermarking embedding algorithm based on wavelet transform (WT), because WT pay great respect to the human visual character, after the transform of image and watermark, we can get the high frequency and low frequency part individually, then embed the high frequency and low frequency part of watermark into the high frequency and low frequency part of image respectively, this may realize the watermarking signal's robustness[4]. And there are several adaptive factor designed in the algorithm which can adapt the embedding intensity of watermark according to the color character of image. On condition of the watermarking's invisible. Compare with the fact that many image watermarking algorithms on hand are only gray, this algorithm can provide he embedding of color image, so it has more profound realistic significance.

2. TWO-DIMENSIONAL WAVELET ANALYSIS OF IMAGE [5] ~ [8]

2.1 conception of wavelet transform

Wavelet is a function bunch $\psi_{a,b}$ that is generated from a faction ψ which fulfill the condition $\int_R \psi(x)dx = 0$ through translating and stretch. i.e.:

$$\psi_{a,b} = \frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a}) \quad a,b \in R \quad a \neq 0 \tag{1}$$

In the formula: a is stretch factor, b is translating factor concerning arbitrary function $f(t) \in L^2(-\infty,+\infty)$ similar as Fourier transform, define the wavelet transform of $f(t)$ as:

$$W_f(a,b) = \frac{1}{\sqrt{|a|}}\int_{-\infty}^{+\infty} f(t)\overline{\psi}(\frac{t-b}{a})dt \tag{2}$$

2.2 Mallat algorithm

Mallat algorithm offers the analysis and restructuring pyramid algorithm of multiple resolution assay, it's effect in wavelet transform is as the FFT in Fourier transform.

If $f(x,y)$ indicate a image, then its wavelet transform turn up as random form. Provide the filter coefficient matrix which dimension function $\phi(x)$ and wavelet function $\psi(x)$ correspond are H and G. Original image $f(x,y)$ is record as C_0 , then two-dimensional wavelet analysis algorithm could be described as:

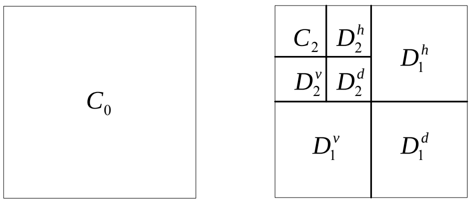
$$\begin{cases} C_{j+1} = HC_jH' \\ D_{j+1}^h = GC_jH' \\ D_{j+1}^v = HC_jG' \\ D_{j+1}^d = GC_jG' \end{cases} \quad j = 0,1,2,\cdots, J \ -1 \tag{3}$$

In the formula: J represents the number of analysis layers, h, v, d represent the horizontal part, vertical part and diagonal part individually., H' and G' are the conjugate convert matrix of H and G. corresponding wavelet restructuring algorithm is as follows:

$$C_{j+1} = H'C_jH + G'D_j^hH + G'D_j^vG + G'D_j^dG \qquad j = J, J-1,\cdots 1 \tag{4}$$

According to Mallat algorithm, after the wavelet analysis, image turns into the low frequency part, horizontal high frequency part, vertical high frequency part, and diagonal high frequency part. Picture 1 illustrates such kind of two layers wavelet image analysis.

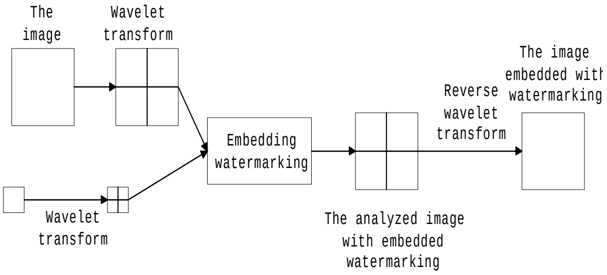
* supported by scientific research foundation of doctor subject of university(20020497006)



Picture 1 illustration of two layers image analysis

3. THE EMBEDDINF PRINCIPLE OF WATERMAKING IMAGE

The basic principle of wavelet transform watermarking algorithm is : proceed the image and watermark as multiple layers wavelet analysis to get the low frequency part (approximate) and high frequency part (detailed) of image, then choose a suitable embedding intensity, embed the low frequency part and high frequency part of watermark into image's, so the low frequency part(approximate) and high frequency part (detailed) of embedding watermark are came into been, afterwards carry out reverse wavelet operation to obtain the image embedded with watermark. The principle of watermark embedding described as follow pictures 2:


The image
Wavelet transform
Embedding watermarking
Reverse wavelet transform
The image embedded with watermarking
Wavelet transform
The analyzed image with embedded watermarking

Picture 2 watermarking algorithm based on Wavelet Transform

The process of wavelet transform watermarking embedding divided as follow procedure:

1) Analysis of image and watermark. Provide image is 24 bit color image of M×N(choose M=N=256 in this paper),watermark is 24 bit color image of m×n(choose m=n=64 in this paper), put them into 3 layers of R,G,B according to the there original color, so every layer of image is M×N two-dimensional matrix, every layer of watermark is m×n two-dimensional matrix.

2) Proceed the image R,G,B 3 layers matrix as multiple layers wavelet analysis, as the result, we get $C_{i,j+1}$ $D^{k}{}_{i,j+1}$、$D^{v}{}_{i,j+1}$、$D^{d}{}_{i,j+1}$ is the high frequency of image after its every layer analysis (approximate), $D^{k}{}_{i,j+1}$、$D^{v}{}_{i,j+1}$、$D^{d}{}_{i,j+1}$ is the high frequency (detailed) of image after its every layer analysis. proceed the watermark R,G,B 3 layers matrix as multiple layers wavelet analysis as the same , as the result, we get $c_{i,j+1}$、$d^{h}_{i,j+1}$、$d^{v}_{i,j+1}$、$d^{d}_{i,j+1}$ (choose $i = R$、G、B)，$c_{i,j+1}$ is the low frequency of image after its every layer analysis (approximate)

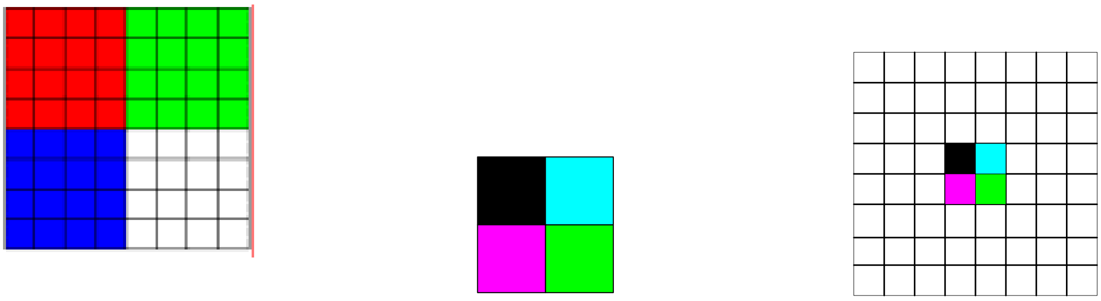
3) Scrambling of watermarking image. In order to resist the attack of clipping, scramble the low frequency $\frac{m}{2}\times\frac{m}{2}$ wavelet coefficient matrix $c_{i,j+1}$ of watermark randomly, and we will get $c_{i,j+1}$. The scrambling algorithm as follow**:**

   Firstly set a $\frac{m}{2}\times\frac{m}{2}$ empty matrix as KEY, in the matrix the elements are orderly $K_i\,(i = 1,2,.....\frac{m}{2},\frac{n}{2})$, encoding the every element $a_i\left(i = 1,2,\Lambda\,,\frac{m}{2}\times\frac{m}{2}\right)$ of the matrix $c_{i,j+1}$, for the every element $a_i(i = 1,2,\Lambda\,,m\times n)$, the process that generates random numbers P_i by the order of $a_1,a_2,a_3,\Lambda\ a_{m\times\cdot(n-1)},a_{m\times n}$ is:

$$P_i = m\times n\times random(1) \quad (i = 1,2,\Lambda\ m\times n) \tag{5}$$

   In the formula the function $random(1)$ generates random numbers between [0,1], so P_i is a random number between [1, m×n], exchange a_i and a_{pi} ,and next record the information of this exchange in the KEY matrix, and like this until the $a_{m\times n}$ also execute this operation, at last we will get wavelet coefficient matrix with scrambled A' The matrix KEY is the encoded privacy key, we can't extract the watermark without this key.

4) Watermarking embedding. Choose the position and intensity of watermarking embedding is the key of watermarking embedding algorithm. Here, we choose m×n area that the low frequency and high frequency intersected after wavelet transform of image. Show as picture 3:



(a) wavelet transform of the image （b）wavelet analyzing of watermark （c）watermarking embedding

Picture 3 position of watermarking embedding

Intensity formula of embedding the every piece of the every layer is:

$$C_{F,i,,j+1}(x,y)=C_{i,j+1}(x,y)+c_{i,j+1}(x,y)\cdot T_i$$
$$D^k_{F,i,j+1}(x,y)=D^k_{i,j+1}(x,y)+d^k_{i,j+1}(x,y)\cdot T_i \qquad (6)$$

In the formula $i=R$、G、B；$k=h$、v、d；$C_{F,i,,j+1}$ is the low frequency part(detailed)，$D^k_{F,i,j+1_i}$ is the high frequency part(approximate).

Embedding Intensity is decided by formula (7):

$$T_R=\frac{aR(x,y)}{R(x,y)+G(x,y)+B(x,y)} \qquad (7)$$
$$T_G=\frac{aG(x,y)}{R(x,y)+G(x,y)+B(x,y)}$$
$$T_B=\frac{aB(x,y)}{R(x,y)+G(x,y)+B(x,y)}$$

in the formula : $R(x,y)$、$G(x,y)$、$B(x,y)$ are respectively the R,G, B value of the image at the coordinate (x, y),

$R(x,y)/[R(x,y)+G(x,y)+B(x,y)]$，
$G(x,y)/[R(x,y)+G(x,y)+B(x,y)]$，
$B(x,y)/[R(x,y)+G(x,y)+B(x,y)]$

are respectively the proportion of the red weight, green weight , blue weight . The α is an adjusting parameter, witch can be adjusted in the experiment.

5) Carry out the wavelet transform rest on formula (4), get the 3 layers—R, G, B of image embedded with watermark on our hands, each one is M×N matrix.
6) Restructure image through 3 layers matrix—R, G, B, which is the image embedded with watermark.

Abstraction of watermarking image is described as below:

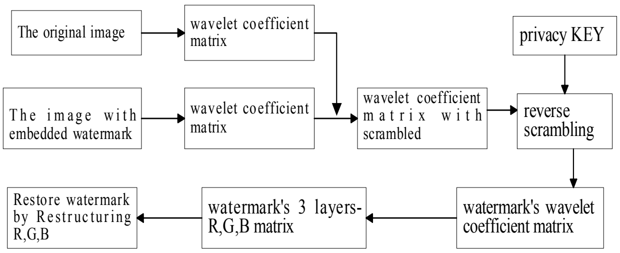
1) Subtraction abstraction. Transform the watermarking image and original image in wavelet form, and then subtract the relevant position of them that to abstract the m×n area containing with the watermarking information. The formula of abstraction is:

$$\begin{cases} c_{i,j+1}(x,y)=[C_{F,i,j+1}(x,y)-C_{i,j+1}(x,y)]/T_i \\ d^k_{i,j+1}(x,y)=[D^k_{F,i,j+1}(x,y)-D^k_{i,j+1}(x,y)]/T_i \end{cases} \qquad (8)$$

2) The reverse scrambling of watermarking image. Reverse scramble the watermarking low frequency wavelet coefficient matrix aided by privacy KEY.

3) Precede reverse wavelet transform in accordance with formula (3), the watermark's 3 layers-R, G, B, each is m×n matrix.

4) Restructure watermark by R, G, B 3 layers.


The original image
wavelet coefficient matrix
privacy KEY
The image with embedded watermark
wavelet coefficient matrix
wavelet coefficient matrix with scrambled
reverse scrambling
Restore watermark by Restructuring R,G,B
watermark's 3 layers-R,G,B matrix
watermark's wavelet coefficient matrix

Picture 4 the process of abstraction of watermarking

4. CALCULATION OF RESEMBLANCE DEGREE NC

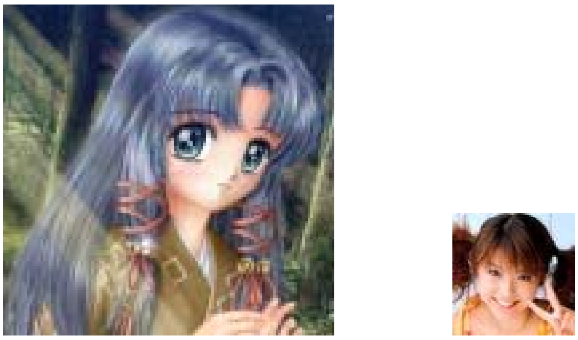
The formula of NC, which is the resemblance between abstracted watermark and original watermark, is as follow:

$$NC=\frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}w_R(i,j)w_R^{'}(i,j)+\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}w_G(i,j)w_G^{'}(i,j)+\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}w_B(i,j)w_B^{'}(i,j)}{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}w_R(i,j)^2+\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}w_G(i,j)^2+\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}w_B(i,j)^2} \qquad (9)$$

Here w is original watermarking matrix, w_R, w_G, w_B each is their R,G,B 3 layers gray value, w' is abstracted matrix $w_R^{'}, w_G^{'}, w_B^{'}$ are their R,G,B 3 layers gray value respectively.

5. RESULT OF EXPERIMENT

We choose original image as 24 bit color image of 256×256 and watermarking image as 24 bit color image of 64×64 individually in the experiment, show in picture 5 (a) and (b), according to the algorithm of embedding and abstraction, adapt them when a=0.1,t=0.7,embedding and abstraction achieve their best condition. The image embedded with watermarking and the watermarking abstracted is shown as Picture 6 (a) and (b).



(a) Original image (b) watermarking

Picture 5 original images and watermarking

(a) Image with watermarking (b) watermarking

Picture 6 abstracted watermarking

6. CONCLUSIONS

The image wavelet analysis is an analysis way of multiple dimensions and multiple resolutions, this character makes the excellent application of wavelet transform in watermarking probably. A adaptive digital watermarking algorithm based on wavelet transform that suggest embed the low frequency and high frequency part of watermark into the low frequency and high frequency part of image respectively proposed in this paper,. In order to amplify the ability of attack-resistance, scramble the low frequency wavelet coefficient matrix of watermark before it had been embedded. The embedding intensity of watermark is the R, G, B proportion at this point. The experiment proves that a satisfying effect had been achieved.

7. REFERENCE

[1] Kundur D. and Hatzinakos D. Digital watermarking for telltale tamper proofing and authentication. Proceeding of The IEEE,1999.7,87(7):1167-1180

[2] Kundur D. and Hatzinakos D. Towards a telltale watermarking technique for tamper proofing.Proc of ICIP'98,2:409-413

[3] Swanson M.,Zhu B.and Tewfik A.Mutliresolution video watermarking using perceptual models and scene segmentation. Proc. IEEE Int. Conf. Image Processing, Santa Barbara,CA,1997.10,2:558-56

[4] Daubechies I. Ten Lectures on Wavelets. CBMS Conference Series in Applied Mathematics, SIAM, Philadephia,1992

[5] Vetterli M. and Kovacevic J. Wavelets and subband coding. Prentice Hall ,Englewood Cliffs,NJ,1996

[6] Miller J.T. and Li C.C. Adaptive multiwavelet initialization. IEEE Trans. On Signal Processing,1998,46(12):3282-3219

[7] Calderbank R., Daubechies I.and et al.Wavelet transform that map integers to integers. Appl. Comut.Harmon.Anal.,1998,5(3):332-369

[8] Donoho D.L. Interpolating wavelet trandform. Technical report, Dept. of Statistics Stanford Univ., 1992.10

Guo Zhiqiang is a lecture of school of information engineering of Wuhan University of Technology. He graduated from Wuhan University of technology and got master degree in 2003. His research interests are image processing, data fusion, image fusion and information hiding etc.

Research on an Agile Protocol for E-Commerce Security

Wang Yong [1, 2] Xiong Qianxing [1]
1. School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei 430063, China
2. School of Management, Wuhan University of Science and Technology
Wuhan, Hubei 430081, China
Email: witsbank@sohu.com **Tel.:**＋86 (0)27 8632 2322

ABSTRACT

A kind of barriers to restrict E-Commerce development is the problem of information security. Though some organizations and scholars put forward some security protocols, such as SSL, SET, and Netbill and so on, after research, it is not hard to find that these protocols are too simple or complicated, not secure and integrate enough and lower efficiency. This paper brings up a new security protocol. It uses object-oriented idea and advanced database technology, researches through security system architecture, agile mechanism, object model and protocol specification etc. we have solved the problem of security and agile property. It meets with application requirements of modern E-Commerce.

Keywords: E-Commerce, Information Security, Protocol, Agile Mechanism, E-P Transaction Model.

1. INTRODUCTION

Higher efficiency and lower cost of E-Commerce makes more and more companies take part in this business mode. However, one of barriers is the problem of information security to restrict E-Commerce development. Information security is becoming an increasingly issue in E-Commerce, due to the incredible expansion of current and distributed system such as database, worldwide web, distributed computing, etc. E-Commerce security depends on protocol. Though some organizations and scholars put forward some security protocols, such as SSL, SET, and Netbill and so on, these protocols have some defects. Some of them are too simple to meet special application, for example, Netbill only can be used to sale digital goods. Some of them are so complicated that it appears low efficiency and larger system expenses, protocol of SET has more than 3000 lines grammar definition, 28 steps transactions, every transaction needs 6 times RSA operations. Some of them have security loopholes, it only considers payment online and anti-negation, and it has not considered other security facts like credit. Some of them are not integrate enough, such as SSL is a security communication protocol based on transmitting layer，it has not considered security of business layer, like electronic negotiation, time limit, agile service etc. They are becoming the development tendency of E-Commerce. The security protocol must meet application needs of modern E-Commerce; it develops to high security, flexible and integrity direction.

This paper presents a new kind of security protocol; it researches the security system architecture, agile mechanism and protocol specification.

2. A NEW CONCEPTION TO SECURITY

Security is the one of design principle of E-Commerce protocol. Usually, people think security of E-Commerce is physical security, memory security, access security and transmitting security. However, there are many facts to influence E-Commerce success or failure, such as service ability, payment method, and credit and so on, these belong to commercial process. Indeed, E-Commerce security not only includes physical security, but also includes information security. Figure 1 shows E-Commerce security system.

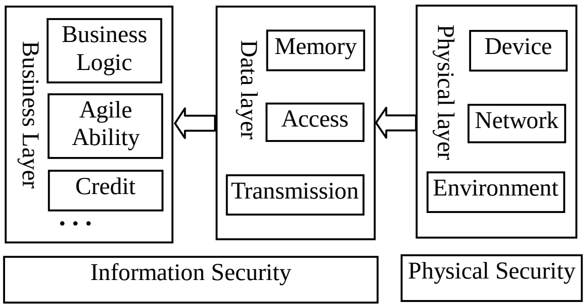


Figure 1 E-Commerce security system

The security system is divided into two parts of physical security and information security; furthermore, it is classified three layers of physical layer, data layer and business layer. Physical security mainly means the influence of device defect, system reliability, environment and non-resistance facts, it can be controlled by increasing system reliability and improving the environment, using transaction dispatch tactics to ensure the security. Information security includes data security and business security. Data security means the security of memory, access and transmission, it can be guaranteed through password, authorization, identification and backup etc. Business security mainly means integrity, consistency and validity of business logic, agile service ability, and credit and so on. We use digital signature, dual signature and logic security verification to ensure the security of business layer.

Presently, there has the ripe technology to control physical security and data security. But business layer is weak; it is waited to be increased. The property of agile service will turn traditional E-Commerce into active E-Commerce.

3. AGILE MECHANISM

Conception of Agility

Let's see an instance of E-Commerce, when a customer accesses merchant web site to buy something, being out of stock, it can not meets customer this time, when the customer comes here next time, it is maybe short yet, but between the two times, perhaps the commodities are in there and sold out again. In this case, it is possible that customer loses the chance not to obtain goods he needs in the end, as a result, customer waste a lot of time, merchant loses business chance too. If merchant can inform customer actively in replenishing stock, or customer can sets up the query time, this problem can be solved better. This case demands E-Commerce system have active service ability, that is to say, it is called agility. Agility means it executes transaction automatically in coming event [1].

E-P Model

The document [2] gives a S-A executing model for advanced database [2]. On base of it, we put forward following executing model.

Definition 1: Events causing agile service consist of different type event.

$E=\{p_t(o)\in(E_t,E_e,E_c,E_u)|o\in D_t\wedge p\in p(o)\}$

Where *E* stands for event set relating to transaction *t*, *pt(o)* stands for event which t executes operating *p* on object *o*, D_t is data set of *t*, *p(o)* is operating set of *o*, E_t, *Ee, Ec, Eu* respectively stands for time event, external event, complex event and user-defined event.

Definition 2: Transaction description. Transaction set deposits a series trigged transaction, which corresponded with different event.

$T= \{p\ (o)|\exists o\exists p\ (o)\ \in(Ts, Tc, Tm,)\}$

Where *Ts*, *Tc* and *Tm* respectively represent system transaction, customer transaction and merchant transaction.

Definition 3: Execution of a transaction management procedure is called a transaction event.

$TE= \{p_t|\exists p\ (p\in TS\wedge p\ (t)\ \in T)\}$

TE stands for a set of transaction event p_t relating to transaction t, *TS* is all possible operating set of transaction management in special transaction model, *p (t)* is operating p of transaction management on *t*.

Definition 4: Rule description. Trigged transaction corresponds with a series of rules; the rule management kit includes condition and trigged transaction.

$R=\{C, T\}$

Where *C*, *T* is separately conditions and trigged transactions.

Definition 5: Processing mode description. An event is related to an executing process of transaction.

$P= \{CD, R\}$

Where *CD*, *R* respectively stands for condition detector and rule.

Inference: Agile executing mechanism is a mate of event and process mode.

$A= \{<e, p>|e\in E, p\in P\}$

This is E-P model.

Agile Service Business System

How to create an agile business system? In E-Commerce, Transaction between customer and merchant proceeds according to protocol. When merchant system can not meet the customer's requirement, the merchant system embedded E-P mechanism allows customer define events and monitor it in real time. Once the defined event appears, condition detector sends message to rule management program, it evaluates the condition and creates a request for a transaction and executes it, at the same time, and agile service protocol starts, so that it can realize the agile service function of merchant system [3]. Figure 2 is the business system with agile ability.

4. PROTOCOL DESCRIPTION

Business Process

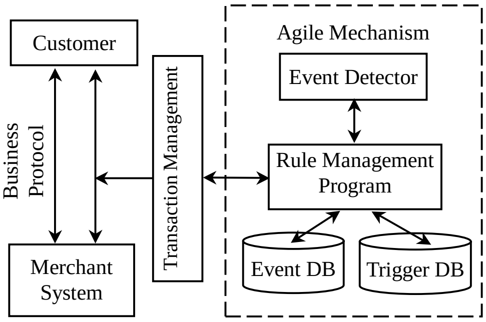


Figure 2 Agile service business systems

According to the business processes descriptions of SET protocol[4], E-Commerce businesses contain certificate authority, merchant web site marketing, customer request, information transmission, identification、payment online, delivery of goods etc. It relates to five major objects, customer, merchant, CA, payment gateway and bank. Banks of merchant and customer are linked with finance network; other objects are connected by Internet. The business process is given in figure 3.

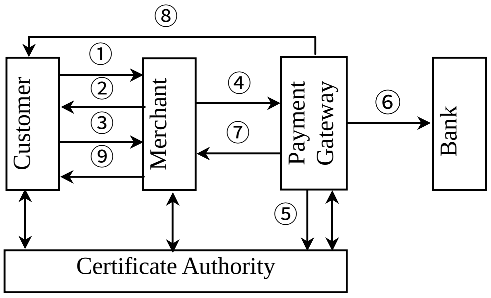


Figure 3 Business process of E-Commerce

Protocol objects

We regard the five major entities as five objects, message transmitting each other as a set of events. In here, we only define three objects of customer, merchant and payment gateway of the protocol, their attributions and deeds given below.

Major attributions of Customer:

CID：Anonymous number of customer. CID=Hash(R_C)，R_C is given randomly by customer.

CCert：Customer certificate.

CCredit：Customer credit rank. It is evaluated by special organization according to trade times, volume of trade, fund information and so on.

CFund：Customer fund information, such as bank (customer opening bank), account and payment method.

CPriKey：Customer private key.

CPubKey：Customer public key.

Customer has deeds of browse, order, submit, request etc.

Major attribution of Merchant:

MID：Basic information of merchant, like name, address, URL, telephone, faxes. etc.

MCert：Merchant certificate.

MCredit：Merchant credit rank.

MBank：Merchant account with the bank for accepting cardholder or digital cash from customer.

Merchant possesses functions of marketing, accept order form, submit message to payment gateway, encrypt and delivery of goods.

Payment gateway has functions to authenticate certificate of merchant and customer, encrypt the message from merchant, pay and return message to merchant and customer.

PGPubKey Payment gateway

TMG：Transaction management procedure.

Protocol Design

This protocol consists of two sub protocol of Business protocol and agility protocol [5]. It is called Business and Agility. Usually, it executes the first protocol, only when it doesn't meet customer, it begins to execute the second protocol. Suppose Business protocol below.

1) Customer, merchant and payment gateway have passed CA and holt the certificates.
2) The identification of customer and merchant authenticated by payment gateway.
3) Before the protocol begins, customer agrees to the goods information of merchant.
4) Each class CA can distribute certificate separately.

The two protocols descriptions are given below.

Transaction
!) Merchant**.** Marketing (goods info.). Merchant spreads goods on its web site.
2) Customer**.** Browse (merchant web). Customer browses merchant web site, if it has no goods he needs, agility protocol begins to execute, or turns to next step.
3) Customer. Order (OI). Customer selects goods and orders. OI is the goods information.
4) Merchant**.** Reply (OI). Merchant returns ordered goods information and ask customer affirming.
5) If the goods quantities or other information does not conform to customer, agility protocol begins to proceed, otherwise turns to next step.
6) Customer**.** Submit(OI,CPriKey(H(H(OI)+H(PI))),PGPubKey (PI, H(OI), CPriKey(H(H(OI)+H(PI)))). Customer submits order instruction and payment instruction through digital signature and dual signature to merchant. PI is the order instruction; H is a Hash function to create the abstract of OI and PI.
7) Merchant**.** Encrypt. Merchant encrypts the message of Pricey (H (H (OI) +H (PI))) with customer public key, verifies integrity and consistency of message.
8) Merchant**.**Submit(PGPubKey(PI, H(OI), CPriKey(H(H(OI)+ H (PI)). Merchant transfers payment instruction to payment gateway.
9) PaymentGateway**.**Encrypt. Payment gateway encrypts the payment instruction with its private key, confirms its integrity and consistency.
10) PaymentGateway**.**Authenticate (CCert, MCert). Payment gateway authenticates identification of merchant and customer from CA.
11) PaymentGateway**.**Pay (CFund, MBank). Payment gateway carries out closing an account through bank network.
12) PaymentGateway**.**Return (MID, CID). Payment gateway returns successful message to merchant and customer.
13) Merchant**.** Send (goods).Merchant deliveries goods to customer.

Agility
1) Customer. Request (goods). Customer fills out an order form, it contains information of goods name, quantities, time of delivery of goods, merchant software system creates user defined event Eu and condition C.
2) CD.Scan (database, system time). Start condition detector, scan database, compare to system time. CD means condition detector.
3) CD.Send(R). Once it meets the condition, CD sends message to rule management program, and proceeds to evaluate the condition.
4) Rule. Activate (T). The rule management program activates the transaction management procedure, creates corresponding transaction to execute.
5) TMG. Inform (CID). TMG Transfers message actively, it informs customer through QQ, customer information system or other communication tools.

5. CONCLUSION

Through research to security requirement and agile mechanism of E-Commerce, this paper introduces an agile protocol for E-Commerce security, it can tackles actively some customer request, makes system respond to customer in time, turns negative purchase into positive request, it brings to customer and merchant more chances. However, this protocol is only a part of research for E-Commerce security protocol, in real world, E-Commerce must meets people's various needs, so there are many respects need to be considered, such as mobile agent purchase, electronic negotiation, protocol security verification and Genetic algorithm in E-Commerce etc. it is worth to research further.

6. REFERENCES

[1] Wang Yong, "Research on Trigger Mechanism of Agile Information System", Computer Era, Vol.2, No.2, February 2002, pp.1~2.

[2] Liu Yunsheng, Advanced Database Technology, Chang Sha: National Defenses Industry Pub., 2001.

[3] R. Yoram, K. Suresh, D. Allen, "Building Agility for Developing Agile Design Information Systems", Research in Engineering Design, Vol.11, No.2, February 1999, pp.67~84.

[4] SET Specification Book1,2,3 Version 1.0, MasterCard &Visa: 1997.

[5] Wang Tian, "A Control Model for Process Security of Electronic Commerce Based on Security of Events", Computer Application Research, May 2001, pp21~23.

Wang Yong is an Associate Professor in Information Management Department, School of Management, Wuhan University of Science and Technology, a doctor candidate in Wuhan University of Technology. He graduated from Wuhan University of Iron and Steel in 1991, from Wuhan University of Science and Technology in 1998. He is engaged in teaching and research of two major courses of information system and e-commerce. He has the honor to win the gold award of the second national commerce plan of college students (2000). He has finished three research projects, edited one book, over 15 Journal papers. His research interests are in advanced database, e-commerce security, intelligent agent and information system.

Grid Security and Relevant Technology

Tian Junfeng Ma Yankun
College of Mathematic and Computer, Hebei University
Baoding, Hebei Province, China
Email: hdmykun@sohu.com **Tel:** 13011923251

ABSTRACT

Grid will be adopted in the mainstream computing and the existing infrastructure. And how fast this curse will be driven by how soon we can solve the security challenges imposed by grid. Security has therefore always been a primary concern within the grid since its origin in the high performance computing community. This paper first analyzes the unique security requirements of grid computing. Then it explores advances in grid security that is beginning to address critical security issues by introducing some advanced technologies applied in the business.

Keywords: Grid, security, PKI, authentication, authorization, SHARP, sandbox

1. INTRODUCTION

The grid is a ubiquitous computing infrastructure that allows flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources. It is this ability that forms these virtual organizations, composed from individuals in real organizations with their own resources that characterize the paradigm shift that is now occurring within many communities dependent on computational resources.

The virtualization of resources offered by the emerging grid middleware enables resources to be shared securely within an enterprise as well as between enterprises to enable new innovative business models. The ultimate vision of the grid is a ubiquitous computing infrastructure providing a utility capability to a worldwide community analogous to the well-established national electrical power grids that underpin our daily lives.

Grid, in the goal of a ubiquitous computing, is trying to provide seamless fabric for sharing of resources over the Internet. Grid computing soon could be central to networking strategy. But what about its security? In this paper, we’ll explore advantages in grid security which are beginning to address critical security issues.

As researchers and corporate administrators get their hands dirty with large-scale grid implementations, they're developing a new generation of grid computing security approaches that stand to meet enterprise needs.

So far, in general, most grid standards that have evolved have not done an extensive job of addressing security at all. But as we go forward they will start doing so, because it's a critical requirement for business.

The rest of this thesis is organized as follows: Part 2 discusses the security requirements of grid computing. Part 3 provides some information on the security issues and introduce some advanced technologies applied in the business, specially highlighted the Globus Toolkit. And finally we’ll discuss the future direction in Part 4.

2. SECURITY REQUIREMENTS

Grid systems and applications may require any or all of the standard security functions, including authentication, access control, integrity, privacy, and nonrepudiation. In this paper, we focus primarily on authentication and access control. Specifically, we should seek to (1) provide authentication solutions that allow a user, the processes that comprise a user's computation, and the resources used by those processes, to verify each other's identity; and (2) allow local access control mechanisms to be applied without change, whenever possible. Authentication forms the foundation of a security policy that enables diverse local security policies to be integrated into a global framework.

In developing a security architecture that meets these requirements, we also choose to satisfy the following constraints derived from the characteristics of the grid environment and grid applications: ***Single sign-on:*** A user should be able to authenticate once (e.g., when starting a computation) and initiate computations that acquire resources, use resources, release resources, and communicate internally, without further authentication of the user. ***Protection of credentials:*** User credentials (passwords, private keys, etc.) must be protected. ***Interoperability with local security solutions:*** While our security solutions may provide interdomain access mechanisms, access to local resources will typically be determined by a local security policy that is enforced by a local security mechanism. It is impractical to modify every local resource to accommodate interdomain access; instead, one or more entities in a domain (e.g., interdomain security servers) must act as agents of remote clients/users for local resources. ***Exportability:*** We require that the code be (a) exportable and (b) executable in multinational testbeds. The exportability means that our security policy cannot directly or indirectly require the use of bulk encryption. ***Uniform credentials/certification infrastructure:*** Interdomain access requires, at a minimum, a common way of expressing the identity of a *security principal* such as an actual user or a resource. Hence, it is imperative to employ a standard (such as X.509v3) for encoding credentials for security principals. ***Support for secure group communication:*** A computation can comprise a number of processes that will need to coordinate their activities as a group. The composition of a process group can and will change during the lifetime of a computation. Hence, support is needed for secure (in this context, authenticated) communication for dynamic groups. No current security solution supports this feature; even GSS-API has no provisions for group security contexts. ***Support for multiple implementations:*** The security policy should not dictate a specific implementation technology. Rather, it should be possible to implement the security policy with a range of security technologies, based on both public and sharing key cryptography.

3. SECURING THE GRID

As we look to sharing resources within our organizations with others, our primary concern has to be to maintain the integrity of these resources. Security has therefore always been a primary concern within the grid since its origin in the high performance computing community in the mid 1990's where national supercomputering resources in the USA where linked to form the first computational grid. Security issue in the grid is usually broken down into three distinct areas: authentication (verifying the identity of an individual), authorization (ensuring they are permitted to access a resource) and accounting (monitoring an individual use of a resource).

Past generations of grid security have built more or less directly on well-known identity management and access control technologies, notably Public Key Infrastructure-based security tools.

Today, emerging grid security efforts are also beginning to address application and infrastructure security issues, including application protection and node-to-node communications. Among other advances, emerging grid security approaches are integrating Kerberos security with PKI/X.509 mechanisms, securing peer connections between network nodes and better protecting grid users and apps from malicious or badly formed code.

3.1 The Globus Toolkit

One of the best-known security approaches for Grid computing can be found within the Globus Toolkit, a widely-used set of components used for building grids.

Within the Globus Toolkit, the defacto standard grid middleware, the Grid Security Infrastructure (GSI) uses an established Public Key Infrastructure (PKI) to build a security framework for distributed computing. This framework is based upon a verifiable certificate (similar to a passport) that may be presented to a resource to authenticate an individual.

The GSI also uses specifically public/private keys and X.509 certificate (named after the standard specifying their format), as the basis for creating secure grids.

Among the GSI’ s key purposes are to provide a single sign-on for multiple grid systems and applications; to offer security technology that can be implemented across varied organizations without requiring a central managing authority; and to offer secure communication between varied elements within a grid.

GSI Authentication
GSI offers users a secure authentication option by creating a time-stamped proxy based on the user's private key. Users can't submit jobs to run or transfer data without creating the proxy. Once created, the proxy is used to grant -- or deny -- access to resources found throughout the grid. Because the proxy is used across the system, this gives to the end user the ability to sign on only once.

One alternative to this authentication approach is GSI-enabled OpenSSH(Secure SHell), which uses the same authentication mechanism. But this lays outside of the core Globus Toolkit functions.

GSI Authorization
The GSI handles user authorization by mapping the user to a local user on the system being accessed. In a GSI-enabled grid, the system receiving the request reads the user's name from the proxy, and then accesses a local file to map that name to a local user.

To avoid creating scores of extra user IDs on varied grid systems, administrators can assign users to virtual groups. All users from a particular domain can be mapped to a single, common user ID when accessing a given grid resource. GSI is designed this way to help administrators separate outside users running grid computations from local users in need of local administration and support.

GSI Confidential Communication
By default, GSI secures communications using digital certificates for mutual authentication and SSL/TLS (Secure Sockets Layer/ Transport Layer Security) for data encryption. The Toolkit contains OpenSSL, which is used to create an encrypted tunnel between grid clients and servers.

For secure remote access to the grid -- for users who cannot physically access their grid client or server -- Globus suggests using GSI-Enabled OpenSSH. Secure Shell (SSH) establishes an encrypted session between the user's client and the grid server.

3.2 Other Technology

The Globus approach outlined above has been enough to kick off a substantial initial round of grid implementations, notably in the financial services and pharmaceutical industries. However, it does leave some organizations out, such as, Kerberos authentication, SHARP, and the sandbox.

Integrating Kerberos
Kerberos authentication is a widely-used authentication method whose mechanism differs from the existing X.509 mechanism and GSI is not directly compatible with it.

One emerging grid security effort attempts to sidestep this problem by blending Kerberos infrastructure and X.509 certification. KX.509, developed at the University of Michigan, is designed to provide a bridge between Kerberos and PKI.

KX.509 is a "Kerberized" client-side program that acquires an X.509 certificate using existing Kerberos tickets. Users who have a valid login to a Kerberos realm, Kerberos client software, and system access to a Kerberos Certificate Authority server can participate. It does this by creating X.509 credentials -- the certificate and private key -- using a user's existing Kerberos ticket. These credentials are then used to generate the Globus proxy certificate. Systems running the Globus software (the standard Globus client or Condor-G) can be set to recognize the KX.509 certificate as though it were a standard X.509 transaction.

The certificate and private key generated by KX.509 are normally stored in the same cache alongside the Kerberos credentials. Netscape or Internet Explorer then loaded a PKCS11 library to secure any Web activity.

SHARP: Secure resource sharing

Another group of emerging grid security efforts, meanwhile, is focused on secure grid resource management, in an effort to share systems more efficiently.

One example is a research project known as SHARP (secure highly available resource peering), a Duke University-based project which tries to define new ways to share grid resources and delegate authority for using those resources.

SHARP proposes a new type of grid security infrastructure, a 'policy server', which controls when, where, and to what extent users can access grid resources. These policy servers give users a ticket which proves to the owner of a resource that this authorized policy server has granted access to it.

Our approach to security in the past is inadequate to grids. For one thing, it doesn't provide means to control resources. The new model is fluid organizations with users all over the world -- and the policy for who accesses what resource is widely distributed.

One of the key features of SHARP is its method for making secure sharing possible without creating a central authority to manage resource requests. Valid principals within a SHARP grid obtain 'claims' to control a share of grid resources; varied principals can exchange claims in the same manner that ISPs do in exchanging network bandwidth for routing.

Within the SHARP model, each site acts as a central authority to certify keys, validate signatures, and detect conflicts for claims on its local resources. Claims are cryptographically signed to make them unforgeable, non-repudiable, and independently verifiable by third parties. Once established, the claims are managed by 'agents', pluggable modules which subdivide the claims and allocate them to their clients. These agents are designed to make the resource claim process more efficient.

To avoid tying up excess system resources, these claims are timed, and would expire after a specified period, so the system can recover the resource if the claim holder doesn't exercise their option. In some situations, agents may oversubscribe resources with extra claims, a method which makes sure that the resource pool is fully used even with some claims failing to materialize or timing out.

The ultimate aim of these technologies is to make on demand computing a reality. Though grid technology is important to the researchers, SHARP's developers are focused on utility computing. The grid is now regarded as a means to address that goal.

Building the sandbox

Yet another class of emerging grid security efforts is designed to help companies protect grids against malicious code, viruses, and other deliberate or inadvertent attacks on the grid infrastructure.

One example comes from commercial grid computing firm United Devices of Austin, TX. United Devices has worked to create a secure shared environment which protects both a user's device from the grid agent and grid devices from user applications. The idea is to make sure that both user and grid applications are only able to access the appropriate set of underlying system resources.

It puts the grid application running on the end user's environment in a 'sandbox,' a protected wrapper of code which sits around the grid application environment and traps all of the calls. The sandbox functionality is typically part of the grid agent, which sits on the user machine.

The sandbox is designed to make sure only safe calls are executed into a host/user environment. The grid system intercepts every file system call and examines it within the sandbox. If the call is not permissible, the grid security system won't let that call proceed. This protects grid users against malicious attacks or badly formed code within applications.

UD also takes steps to secure data being used by the grid application. This aspect of their technology is designed to make sure that grid application users can't see or modify grid data before it can be shunted back to the grid's central server.

UD uses an array of standard data protection techniques, such as encrypting the data after it is written out to a disk and PKI-based authentication to vouch for the parties doing the data exchange. The system verifies both grid users and applications using PKI and X.509 certificates for digital signatures and authentication.

As part of installing the UD system, customers create a signature on all grid applications, using their own existing set of keys. When the agent starts up, it communicates to central services and asks for work. As part of that process, the agent is authenticated to make sure that the agent is a valid part of the grid. The user's system, in turn, then checks whether the grid application is granted by checking its digital signature.

4. FUTURE DIRECTIONS

Grid Security and its associated aspects of authentication, authorization and accounting have always been a priority to grid middleware developers. Going forward, grid security efforts should embrace technologies rapidly, while they're still at the cutting edge of mainstream corporate development.

The current Grid Security Infrastructure is undergoing standardization though bodies such as the Global Grid Forum and the Internet Engineering Task Force. While the recent adoption of the Open Grid Services Architecture (OGSA) by the grid community will combine existing and future Web Service based security standards with the collaborative demands of virtual organizations, it is these commercially proven security infrastructures for managing Web Services from major industrial vendors that will form the basis for future grid middleware deployments.

Members of the OGSA security group plan to realize OGSA security using the WS-Security standard backed by IBM, Microsoft, and VeriSign Inc. Among other features, WS-Security offers security enhancements for SOAP messaging and methods for encoding X.509 certificates and Kerberos tickets. With critical technologies like Web services being securely grid-enabled, grid technology should soon be central to just about any enterprise's networking strategy.

5. REFERENCES

[1] Overview of the Grid Security Infrastructure, http://www.globus.org/security/overview.html

[2] Anne Zieger, Grid security: state of the art--Expanded grid security approaches emerge, http://www-900.ibm.com/developerWorks/cn/grid/gr-security/index_eng.shtml

[3] Ian Foster, Carl Kesselman, Gene Tsudik, Steven Tuecke, A Security Architecture for Computational Grids, ftp://ftp.globus.org/pub/globus/papers/security.pdf

[4] Dr Steven Newhouse, Security in the Grid, http://www.lesc.ic.ac.uk/admin/security.html

[5] Vinay Bansal, Policy Based Firewall for GRID Security, http://www.cs.duke.edu/~vkb/security/grid-firewall/report/GRID_Firewall.pdf

Tian Junfeng, full professor and head of the Parallel Processing and Distributed Computing Lab, is currently serving as Secretary of CPC’s committee in the College of Mathematic and Computer of Hebei University. And now he is a candidate of Doctor’s Degree in National University of Defense Technology. Professor Tian has been working on teaching and scientific research for many years and has taken change of many Projects, such as Hebei Province Natural Sciences Foundation Project and many other important transverse projects. He has published several books, over 30 Journal papers and now is the vice-secretary-general of Professional Committee of Open System of China Computer Federation. His current research interests include network technology, system architecture, grid computing and distributed computing.

Ma Yankun, candidate of master’s degree in the College of Mathematic and Computer of Hebei University. She is studying on the issue of distributed parallel processing and grid computing in the Parallel Processing and Distributed Computing Lab.

Security Design Model of E-Government Collaborative Platform

Su Jindian Guo Heqing Yu Shanshan
Dept. of Computer Science and Engineering, South China University of Technology
Guangzhou 510641, China
Email: Brisk_su@hotmail.com susyu@tom.com **Tel.:** (020) 13711126764

ABSTRACT

Based on many security problems existing in current e-government collaborative platform construction and incorporating some popular security technologies such as Public Key Infrastructure，Digital Signature and Electronic Seal，this paper tries to do some researches on how to build secure and dependable e-government platforms according to the different security requirements in Internet, extranet and intranet respectively. The main goal is to provide a reference to the security design of e-government systems.

Keywords: E-Government, PKI, Internet, Extranet, Internet, Document Flow

1. INTRODUCTION

Nowadays, with the rapid development of e-government collaborative platform construction in the right direction, how to build a secure, dependable and trustworthy e-government platform becomes a demanding task for software developers and government workers.

E-government organizations in China can be divided into two main types: “Vertical-align” and “Horizontal-align”, and government institutes at different levels have different demands on the security level of their information and systems. For the superior government departments, more functions of decision-making management they have, higher demands on information secrecy, security, usability, integrity, non-repudiation and auditability they need [1]. So, system security design models of electronic government collaborative platform must adopt corresponding security methods according to different security requirements. (Logic System Architecture of E-Government is shown in Figure 1 [1].)

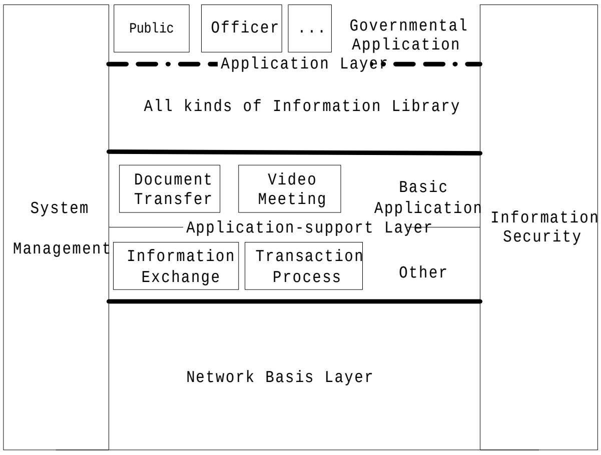


Fig.1 Logic System Architecture of E-Government

The information security part of Chinese E-Government Standardization Specifications points out clearly that security technologies should be widely incorporated in network infrastructure layer, application-support layer and application layer in e-government systems. Especially for those electronic government collaborative platforms based on network, they should guarantee multi-level security and secrecy of systems and information. Besides considering the security of governmental intranet, system designs should also take the security of Internet and extranet into account.

2. DESIGN MODELS OF SYSTEM SECURITY

Electronic government networks can be divided into three types: Internet, extranet and intranet. Different types of networks mean the risks that e-government systems will face have a lot of differences and their system design models must use different security technologies according to concrete network conditions. These technologies should be secure enough and difficult to be cracked, at the same time they must be flexible and convenient for users to master and implement.

2.1. Internet

Internet also calls Public Network, whose main purposes are to enable the public to know various kinds of governmental information and services. Usually, some web portals are provided as the windows of the government departments and for public users to browse.

As these e-government portals must expose themselves to Internet users all the time, so the servers subject to be the objects of being attacked or breached; especially they often become the targets of malicious hackers from enemy countries. But for public network, it is not suitable to use too strict identity authentication mechanisms and authority controls. In order to guarantee the security of servers by all means, the platform can use some effective methods such as hardware firewalls, software firewalls, anti-virus software and hole-scan software to strengthen the security of web servers and database servers. Meanwhile, the servers providing public services should seek full possibility to be separated from those servers providing special business services and inner office systems in order to deploy different security mechanisms and thus decrease the risk of being attacked. (Internet Security Design Model of E-government is shown in Figure 2.)

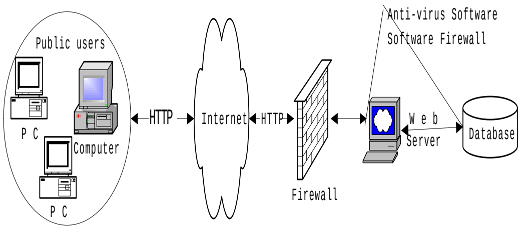


Fig.2 Internet Security Design Model of E-government

2.2. Extranet

Extranet is the professional business network of government, which mainly provides those professional services facing society and businesses that are not suitable to run in intranet. Implementation of handling official business work via network remotely is one of the most important goals of e-government systems. According to the requirements of "Three Network and One Library ", extranet's main purposes are to establish a communication platform among government departments. But because extranet often involves some professional services and sensitive information, its security designs should be stricter than those of normal public network. (Extranet Security Design Model of E-government is shown in Figure 3)

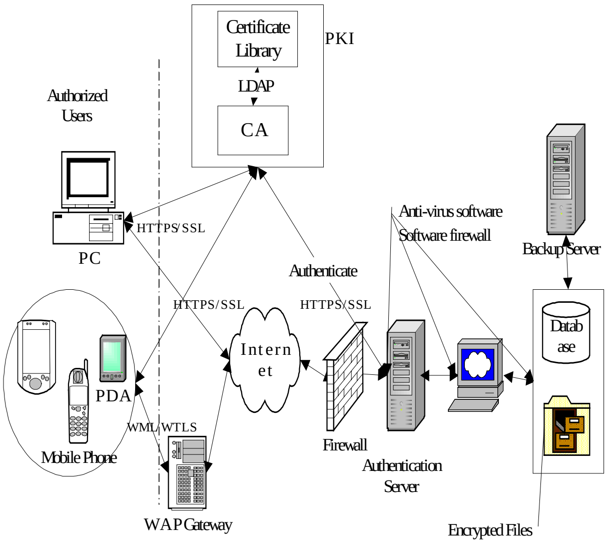


Fig. 3 Extranet Security Design Model of E-government

When users access business office systems or internal data using personal computers, PDA, mobile phones or mobile computers, it must strictly confirm the dependability of the connections and the security of information. At the present time, we can use PKI (Public Key Infrastructure) as a satisfied solution to take charge of users' login and identity validation authentication, and solve those security problems such as integrity, secrecy and non-repudiation of electronic information.

For example, when a user wants to login, he can send a certificate to the server and the server will ask a CA (Certificate Authority) for authenticating the certificate. After the CA center validates the user, the user can establish a connection with the server using HTTP or SSL and encrypt information being transferred. Secret information or inner data should be encrypted or deny network access according to the requirements of security levels.

A noticed problem is that wireless network is beginning to be widely used. Besides a lot of mobile applications wireless network has brought to the public, it also produce many security problems; especially when mobile office becomes an important part of e-government, which enables government officers to work remotely via wireless network. So the security of wireless communication is also a key part that should be paid special attention to during the process of e-government collaborative platform security design.

In the routine system management, many administrators' security consciousness is so weak that they often take it for granted that application servers and database servers will not be attacked once they have used firewalls. Although, facts prove that firewalls can resist most of network attacks, it is useless when facing some new attack methods. So, besides using traditional hardware firewalls, application servers should also combine software firewalls, real-time anti-virus software and hole-scan and repair software to guarantee the security of systems and data. At the same time, data in the database should backup timely to improve their ability of fault tolerance and recovery, and confidential information should also be encrypted.

2.3. Intranet

E-government intranet mainly refers to the physical disconnection between networks above vice-province-level governmental departments and those networks below province-level, and the scientific management according to secrecy and authorization. Because of the physical disconnection, most of the intranet administrators tend to believe that their systems in the intranet are secure enough and thus neglect the supervision of workers in the intranet. According to a survey, almost 80 percent of information leaks are done by inner workers. Compared with outer assaulters, inner workers usually have a deeper understanding with the systems and their attack effects are more serious. So, how to prevent the attacks from office workers and establish an integrated security mechanism is also a key problem that should be solved with great emphasis.

In Chinese government departments, many employees have not enough consciousness of the importance of security and their passwords are very simple and easy to be guessed. Messages transferred in the intranet do not have any encryption or just have simple encryption. Once the passwords and messages are spied, they are easy to be cracked; especially for leaders in offices, the consequences will be very serious once their identities are misused by some malicious workers. So, it is necessary to implement a strict control on users' identities and authorities to guarantee that all accesses to resources are authenticated and authorized first; or after systems are attacked, it is able to find out the attacker's identity according to the evidences recorded by the systems during the process of attack [2]. Those confidential files and data should be encrypted and databases should backup timely. (Intranet Security Design Model of E-government is shown in Figure 4)

Firstly, the system can require identity authentication when users login the system, and then adopt limited authorization mechanism. As the range involved in intranet is relatively small, it is not suitable to use PKI. But it can combine security authentication mechanisms of operation systems and web servers to strengthen security. For example, windows platform can use interactive login that means using a user's local computer or active directory account to validate the user; or using network identity authentication to verify the user according to the network services the user try to access. Only those users who are permitted can access authorized resources inside intranet. Confidential data such as users' passwords and secret messages being transferred via intranet must be encrypted so as to avoid being spied and assure the integrity of information. At the same time, systems should provide an

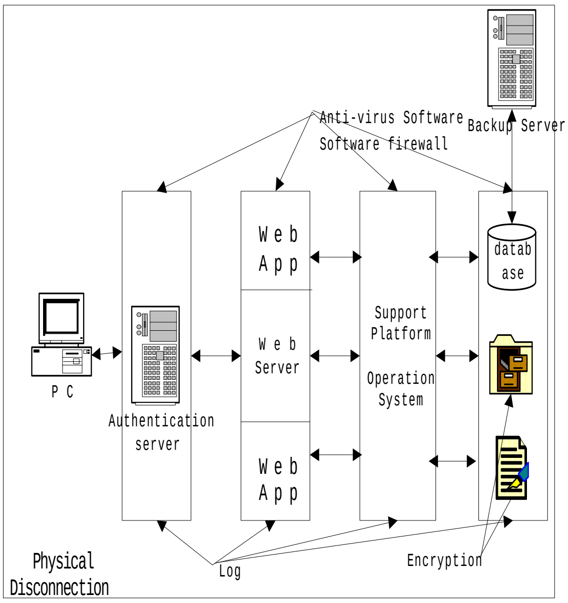


Fig. 4 Intranet Security Design Model of E-government

integrated set of authority management mechanisms, which makes it possible to have concrete controls such as read-write, modification and read-only authority to each file. For some confidential files' modification and deletion, systems can provide detailed log traces.

Another problem that should be noticed inside intranet is network cables should meet the requirements of country's secrecy. For some important fields involving national security, stricter prevention measures ought to be taken and network equipments should use dependable electromagnetism shield or anti-jamming facilities to prevent from being spied and improve the ability of anti-jamming.

3. DOCUMENT FLOW

Document flow, as one of the most important parts of electronic government systems, often requires many positions and operators to deal with the same file. Some files might include sensitive information about offices or country, and different workers have different process authorities at different time. So, it is necessary to judge whether some user has access or modification authorities to a file at some time, and assure the exactness and high-performance of document process and security of document transfer. Reading, transferring, keeping in the archives and auditing the files during the whole process of document flow must have real-time traces and information's modification and examination and sanction should be recorded. In order to accomplish this, system designs can incorporate multi-level security mechanisms such as identity authentication, data access control and authorizations to assure documents' security and secrecy. (Security Design Model of Document Flow is shown in Figure 5)

Before accessing documents, every user must pass through identity authentication and authority examination; especially for some confidential files, usually only a few persons are allowed to view or audit them. After dealing with a document, office leaders need to stamp on or sign in the document. As a result, it is necessary to make sure the validation and non-reputation of the seal or signature, and provide the authentication inquiry of all signatures or seals. During the time of providing electronic seal, it should try to deploy secure and dependable electronic seal encryption algorithms and reasonable seal management to prevent documents from being processed by some unauthorized people.

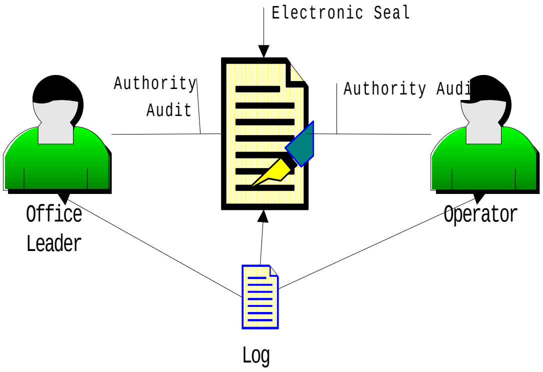


Fig.5 Security Design Model of Document Flow

During the process of document flow, it is very important to try to avoid the interruption of the document flow. Systems should provide error detection mechanisms and security recovery functions. When errors occur, systems can rollback to the original state and prevent the disappearance of documents before finish.

4. CONCLUSIONS

Security design of e-government systems is a very complex and systemic project, which includes a lot of aspects, i.e. software, hardware and networks. Besides using all kinds of dependable security technologies, the integrity of security also depends on the establishment and strict execution of security rules and management specifications. But the key is to improve the security consciousness of systems users; especially for those workers and system administrators in government offices, they should be required to have a high concern of the importance of e-government from the point of view of national security and social stability. Only in this way can it adapt to the development of government informization [3].

5. REFERENCES

[1] Electronic Government Standardization Guide -Sixth Parts: Information Security Pages 4-5, 2003

[2] Thomas A Wadlow, The Implement of Network Security, Beijing, Post & Telecom Press, 2000

[3] Yue Shen-qin, Electronic Government Should Pass Security Gate First, 2002

[4] Zhang Huan-guo Tan Zhong-ping, Chinese Congress on Information and Communication Security—CCICS'2003, Science Publishing House

Su Jindian is presently pursuing a PhD degree in Computer Science Department at South China Univ. of Tech. His current research interest includes Internet security，E-Government，PKI，Design Pattern and trust models.

Guo He-qing (1936-), female, professor, South China University of Technology, director of software engineering lab, research direction: Integration of Network Information System, Intelligent Information System, Software Engineering and Web Software Factory. E-mail: Guozhou@scut.edu.cn

Study of Trust Model for Grid Security*

Zhu Dawei, Zhou Zude, Liu Quan
School of Information Engineering, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: zdw@mail.whut.edu.cn; **Tel:** 13349951639

ABSTRACT

Grid security mainly bases on the trust relationship establishment. In this article, we focus on this issue and assess major exiting PKI trust propagation models. Based on analysis and assessment, we develop a BCA-based hybrid trust model with which dynamic and distributed Grid resources requirement are matched. Also we study some key techniques to achieve the Grid trust model such as trust path construction and middleware for trust management. Additionally, we provide a solution to trust in Grid environment.

Keywords: Grid, trust model, PKI, Bridge-CA, middleware

1. INTRODUCTION

Grid is an important information technology emerged recent years. It aims at uniting compute systems, storage systems, pools of servers, and networks into one big virtual computing system in “virtual organizations” (VOs) so that non-trivial qualities of service can be delivered to end user or application. Because of the expansible, dynamic, distributed and multi-institutional nature of these environments, how to solve this problem safely is the crucial point for technical approaches. [1]

As described by the Open Grid Services Architecture (OGSA): the security challenges faced in a Grid environment can be grouped into three categories: integration, interoperability and trust. And solutions to define, manage and enforce trust policies within a dynamic Grid environment are very important to Grid security.

Because of the dynamic nature of Grids, the trust relationship establishment needs to use the trust proxies as intermediaries dynamically. Trust can be established and enforced on basis of trust policies which can be defined a-priori or dynamically. After a trust model is defined, it will play a role in defining how trust assertions are to be consumed by a service provider or a requester as the case may be. Single logon achievement derived from trust of asserting authority or trust on requesting member of a VO will be also satisfied with the basis the formed model.

Due to the user-controlled deployment, dynamic nature and management of transient services, the trust relationship establishment is more difficult in a Grid environment. Such transient services created by end users are used to achieve request-specific tasks, in which user code can be executed. For example, in a distributed data mining scenario, transient services may be created at various locations both to extract information from remote databases and to synthesize summary information. User-created transient services lead to the problems: identity and authorization, policy enforcement, assurance level discovery, policy composition, delegation.

To address these challenges, this paper analyzes and compares exiting PKI trust models. Also it sets up a trust model to meet the dynamic nature of Grid resource and discusses the key techniques. [2, 3]

2. TRUST MODEL FOR GRID

2.1Analysis of PKI Trust Model

In a Grid environment, applied resources are involved in different PKI domains. Because of the distributed nature, trust delegation and trust propagation are required in this environment. Several PKI models will be taken into account: Subordinated Hierarchical Trust Model, Equity Trust Model, Cross-certified Meshes Trust Model, CA Bridge Trust Model and Trust Lists model. [1, 4]

- The subordinated hierarchical model involves a tree structure of certification authorities, with each node certifying nodes at the level below. Each end entity has only one certification path, which starts from the root certification authority. Within this model, each participant must have knowledge of the root CA’s public key, which forms the fundamental trust anchor for all participants.

- The hybrid PKI uses bilateral cross-certification between hierarchical and mesh PKIs. The properties of this model as following: multiple root CAs exist; all non-root CAs are certified within a root CA’s hierarchy, with paths certified both “downward” from the root and “upward” towards it; root CAs establish a cross-certified mesh among themselves, so each hierarchy can reach every other hierarchy via a single cross-certificate at the root level.

- The Cross-certified Mesh Trust Model is built with cross-certified CAs with a single level of subordinate entities. The relationships of certificate users are not superior-to-subordinate but peer-to-peer, and CAs issue cross-certificate to each other.

- The Bridge CA model embodies a central cross-certification authority, whose purpose is to provide cross-certificates rather than acting as the root of certification paths. As in the mesh model, within the Bridge CA model, a participant is preconfigured with the public key of its local CA to act as a trust anchor; configured knowledge of the Bridge CA’s own public key is not required.

- The PKI Trust List model uses a set of trust lists to enable remote entities to discover the trust path driven by verifiers. In this model, client systems (or their delegated verifiers) are provided with the public keys of a set of trusted roots. To be validated successfully, a certificate must chain to one of these trusted roots, sometimes directly and without

* Supported by: national key scientific and technological project (No:2003BA103C)

intervening CAs.

Table 1 Trust Models for Grid Security

	Subordinated Hierarchy	Cross-Cert. Mesh	Hybrid	Bridge CA	Trust Lists
Interoperability over Multiple PKI Domains	Weak beyond common root	Good Through moderate numbers of enterprises	Good through moderate numbers of enterprises	Very good through large scale	Fair: May require Intensive management
Trust Path Complexity	Simple within local hierarchy: down from root only	Hard: may be multiple routes to source, requiring iteration	Medium: multiple routes may exist, but simple path known	Simple: all non-local paths traverse. bridge	Simple but limited: all available paths begin within local trust list
Service Growth Scalability	Medium with a top-down growth	Low with a pairwise growth	Medium with a top-down pairwise growth	Very high with bridge CAs	Medium with recognition by verifiers
Robustness and Directory Dependence	Low due to one or fewer roots	Robust but high dependence on the directory	Medium with a limited number of PKI domains	Very robust with a cluster-powered architecture	Robust and low directory dependence

2.2 Comparison among Trust Models

When we evaluate PKI models in a Grid environment, the trust propagation is an important factor which can be considered from four aspects: interoperability over multiple PKI domains, efficiency in trust path construction, service growth scalability and fault tolerance and dependence on database directory. We assess these models in Table 1in above four aspects. [2]

Because of its low interoperability and low fault tolerance compared with others, the subordinated hierarchical PKI model can only meet the requirement of localized Grid computing. The mesh PKI model would be limited by the scalability when the number of Grid resource sites increases rapidly. The hybrid PKI model inherits from subordinated hierarchy and mesh models' advantages and not much shortcomings. Although the trust PKI model is strong in robustness and less relay on the database directory, it is weak in interoperability. So it cannot satisfy the request of Grid resource's nature. [3]

3. BRIDGE CA-BASED HYBRID TRUST MODEL FOR GRID

3.1 Trust Model

Usually service requests in a certain Grid can span several security domains. So under the circumstances, trust relationships of security domains are important for those end-to-end traversals. For example, a server needs to make the clients know its requirements so that the clients can securely request an access. If trust between end points relies on topological assumptions (e.g. they are parts of a VPN), it can be presumed. If trust is specified as policies and enforced through exchange of some trust-forming credentials, it can explicit. As a consequence of the dynamic and distributed nature of VO relationship, presumed trust is unfeasible. In some circumstances, trust may be established once for every session or evaluated on every request dynamically. In particular circumstances, with the dynamic nature of the Grid, the trust relationship can be established among sites before application executing. It is necessary to realize that the participating domains may have different security infrastructures (e.g., Kerberos, PKI), so we need to build trust

relationships which through some form of federation among the security mechanisms. [6]

Based on the evaluation of 5 PKI models, the bridge CA model can satisfy the request of large-scale grid application. In the bridge CA model, many different PKI domains can be bridged together through a central authority for cross certification; in this view the model has the strongest scalability and interoperability. And the bridge CA model scales very well with the demands of services over wired PKI and wireless PKI domains. And a cross certificate with the central bridge CA is built by each participating PKI domain.

But using only one trust model for Grid must be limitation, because of complexity of Grid resource. In this case, we introduce Bridge CA-based hybrid trust model, and discuss trust path construction, middleware for trust management and other key technologies of this trust model.

Base on using trust model in existing PKI domains, BCA-based hybrid trust model interconnects various PKI domains through which central Bridge CA sets up a cross-certificate with each PKI domain. The participating PKI domains can assume different PKI structures such as the hierarchical, trust lists, and mesh-structured domains. The inter-domain trust path must traverse through the bridge CA with cross-certificates carrying the policy mapping, path constraints, etc. This can implement PKI trust propagation in Grid resource.

Both server and client side application will be satisfied with the high scalability and availability of the BCA-based hybrid trust mode. Using the BCA-based hybrid trust model, the process to export, import, or update the policy configuration can be more efficiently carried out without affecting the client side applications. And the administrator can add or remove redundant cluster nodes to get higher performance. So this model well meets the need of the dynamic nature of Grid resources. [3, 4]

As illustrated in Figure 1, BCA-based hybrid trust model scales well with the demand of services over various PKI

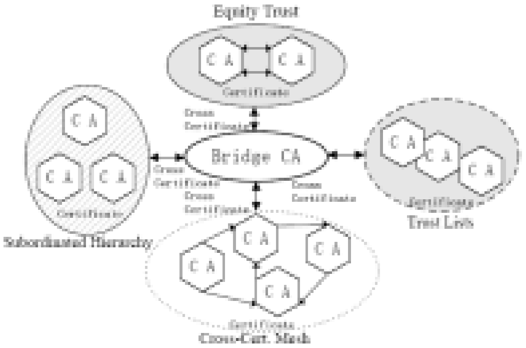


Figure 1 Trust Propagation Using BCA-based Hybrid Trust Model.

domains. Each PKI domain sets up a cross-certificate with the central Bridge CA. By setting up the central trust bridge, the length of trust path is effectively shortened, thus lower trust propagation overhead. The trust relationship between BCA and participating CA is peer to peer. To interconnect ***n*** PKI domains, 2***n*** cross-certificates are required. The number of certificate is small, and management of certificate is convenience. And this model can interconnects PKI domains with heterogeneous structures, protocols, and devices, which is exactly the situation facing all heterogeneous Grid construction.

In sum, in grid applications the BCA-based hybrid trust model has the following distinct advantages.

- Support heterogeneous certificates: model can support X.509 certificates and their variants such as attribute certificates, thus it offers high application potential.
- Scalability and Manageability: The Bridge CA cluster can scale freely in establishing the trust relationship among multiple CA's associated with different PKI domains.
- Low implementation cost: The certificate path discovery and interoperability support grow together in a peer-to-peer fashion, thus reducing the implementing costs.
- Robustness and Availability: The BCA cluster architecture has failover and recovering capability after the failure of any processing node in the bridge CA cluster.

So this model fully meets the need of heterogeneous Grid resources, lowers difficulty of trust path construction, and heightens credibility of trust relation.

3.2 Key Techniques For Trust Model

3.2.1 Trust Path Construction：How to establish trust path between Grid resource domains is very important for trust model in Grid environment. Figure 2 shows how to establish the trust path between a client in one Grid resource domain (GRD) and a server in another GRD. In this process, we assume that trust relationship, trust agents, CA and BCA agreement are all in the security database of each GRD.

As illustrated in figure 2, a trust path consists of a sequence of intermediate CAs which use cross certificates to connect each other, and cross certificate accelerates the path construction. In case of multiple trust paths existing in the process of path construction, we should choose one which with minimum latency, minimum cost, highest reliability, assured security and privacy protection. From Table 1, we can find bridge CA model has added only moderate cost in handling cross certificates, and bridge CA models have the lowest path construction cost.

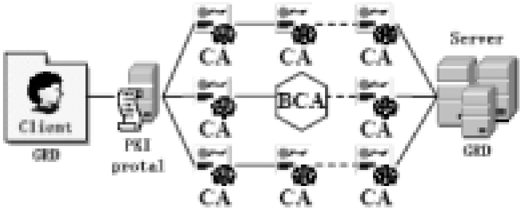


Figure 2 Trust Path Constructions between Grid Resource Domains.
(GRD: Grid resource domain, CA: certificate authority
BCA: bridge CA)

We find that BCA-based hybrid trust model is best implemented by using multi-server cluster architecture. In the BCA-based hybrid trust model, on one hand, each principal CA issues a cross certificate to the BCA, on the other hand, each principal CA asserts its own issuer domain certificate policy to map the BCA policies. Trust propagation is carried out by each sever node being in change of one specific BCA function. In Figure 3, we demonstrate how to use middleware to implement trust management; middleware with four bridge CA functions is running on four server engines.

3.2.2 Middleware for Trust Management

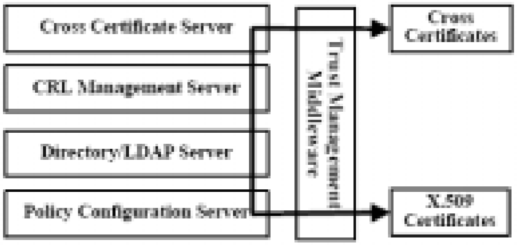


Figure 3 Middleware for Trust Management on A Multi-server Bridge CA Cluster

In the Grids, cross certificates issued from the root CA of various PKI domains, we use the middleware to manage these cross certificates. The middleware consists of four processing engines. Each processing engine is specified in the left four boxes: cross certificate server, CRL management server, directory/LDAP server and policy configuration server. These are for handling cross certificates, maintaining the certificate revocation list (CRL), checking the database directory LDAP (light-weight directory access protocol), and security policy reconfiguration. Because the middleware based on the clustered BCA design, it must be scalable with adding more server engines, in case the grid coverage increases or the Grid needs new server. So the design of middleware is expansibility and meets the need of Grid development.

In the above design, the middleware glues all four severs together to carry out the bridge PKI services that include trust propagation and delegation of trust throughout the different Grid resources. It is worth explaining here that the possibility to for middleware to include the integration of distributed direction depends on the special requirement in the grid applications. But the single sing-on authentication across various PKI domains is still desired in the Grid.

4. CONCLUSION

The trust relationship establishment is the foundation of Grid security. Any solution for federating credentials to achieve interoperability will be depended on the trust models defined within the participating domains and the level of integration of the services within a domain. This paper provides BCA-based hybrid trust model. This model is fully adapted to the characteristic of the dynamic and distributed nature of Grid resources and basically satisfies the needs of different security mechanisms and security policy in the existing Grid environment. Without doubt, it will enable businesses and organizations more rapidly to develop secure, interoperable Grid services. Of course, it should be stressed here that the building of trust model is only one basic links for Grid security. The basic OGSA security model includes the following security disciplines: single logon, message integrity, secure logging, firewall traversal and so on. These are directions that will be studied in the future.

5. REFERENCES

[1] S.Tuecke, “Grid Security Infrastructure (GSI) Roadmap”, Internet Draft, October 2000.

[2] R. Perlman, “An Overview of PKI Trust Models” IEEE Network, Dec. 1999, pp.38-43.

[3] W. T. Polk and N. E. Hastings, “Bridge Certification Authorities: Connecting B2B Public-Key Infrastructures”, Technical Report, National Institute of Standards and Technology, 2001.

[4] J. Linn, “Trust Models and Management in PKI”, Technical Report, RSA Security Laboratories, Nov.6, 2000.

[5] Perlman R. “An overview of PKI trust models” [EB/ OL]. netlab. cs. ts2, August, 2003.

[6] I. Foster, C. Kesselman, G. Tsudik, S. Tuecke. “A Security Architecture for Computational Grids”, Proc. 5th ACM Conference on Computer and Communications Security Conference, 1998, pp. 83-92.

[7] 2 I. Foster, C. Kesselman, S. Tuecke “The Anatomy of the Grid: Enabling Scalable Virtual Organizations”, International J. Supercomputer Applications, 15(3), 2001.

[8] 3 I. Foster, C. Kesselman, J. Nick, S. Tuecke, “The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration”, January, 2002.z

Dawei Zhu, male, was born in Jan. 1980, is a postgraduate in the School of Information Engineering, Wuhan University of Technology. His research direction is grid computing and network security.

ZuDe Zhou, male, professor, tutor of doctor, is the President of Wuhan University of Technology, his research interests are CNC Theory and technology, Intelligent Control, Digital Manufacturing, Reliability and Fault Diagnosis of the Modern Manufacturing Systems and etc.

Research on Secure Gateway Based on Real-time Embedded Systems

Yufeng Wu, Quan Liu, Fangmin Li
School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China
Email: ahthwyf@tom.com**,** qliu@public.wh.hb.cn, lifangmin@mail.whut.edu.cn
Tel: 13545250928, 13908659571, 13973255023

ABSTRACT

Based on two typical modern encryption (DES algorithm and RSA algorithm), this paper proposes the encryption algorithm which combines DES algorithm and RSA algorithm and implements it with hardware, thus the whole performance of the secure gateway gets improvement, then we design the secure gateway in detail, present the hardware frame and RSA algorithm module. At last, we apply the secure gateway to VPN, and obtain good effect.

Keywords: FPGA, DES-RSA, VPN, Encrypt, Real-time Embedded Systems, Secure Gateway

1. INTRODUCTION

With the development of computer network, the opening and sharing of the network build up, the degree of interconnecting is continually expanding, many new operations begin spring up(such as E-business 、digital currency 、Internet bank and so on), however, the present network protocol is IPv4, and it doesn't provide security function, so the network security is becoming more and more important. Some military affairs 、finance organizations and companies which own secret data have to lay special network 、 special line or hire special line in order to ensure secure communication, and the reduplicate construction induces much waste, at the same time, different companies and finance organizations need found more and more frequent business relationships, and the special line and special network isn't flexible. Secure gateway is a special network security device, which fully uses the network resource and can filtrate 、authenticate and encrypt to the information which is transmitted through the network, at the same time, it can be used to build high secure and powerful function's Virtual Private Network (VPN). Therefore, it is very significative to research and design a high performance secure gateway.

At present, the international standard of building VPN is IPSec and L2TP. L2TP is protocol of Virtual Private dial-up Network, and it is still in the stage of the draft at present. IPSec, which is the opening IP security standard based on IP network, was instituted formally by IETF, and it is the base of VPN and is quite mature and credible [1]. Therefore, we can analyze the working principle of IPSec in order to illuminate the VPN's. IPSec security protocol includes Authentication Header (AH) protocol and Encapsulating Security Payload (ESP). AH is used to ensure the user and data and ESP strictly protects the transferred data. Both AH and ESP support two modes: tunnel mode and transfers mode. The tunnel mode needs to adopt new IP header to encapsulate, while transfers mode doesn't need to [2]. The VPN systems designed by us substitute secure gateway for the secure transmit work handled by user's computer, since the tunnel's beginning and ending are both in secure gateway, so it is necessary to encapsulate new IP header in the out of the original IP header, and the new IP header uses the IP address of secure gateway as the source IP address and destination IP address. The framework of VPN, which is composed of secure gateway based on tunnel mode, is shown in figure1.

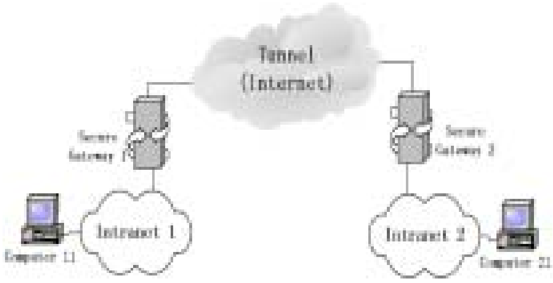


Figure 1. VPN frame consists of secure gateway based on tunnel mode

2. RESEARCH ON REAL-TIME SYSTEMS

Since the Real-time characteristic is very important in secure gateway, we research and choose the appropriate Real-time systems for the secure gateway in the following content.

2.1 Merits of Real-time Embedded Systems

Embedded systems are special systems, which are based on computer technology, whose software and hardware are configurable and have strict constraint to function 、reliability 、cost 、volume and power [3]. Real-time embedded systems are operating systems which can respond the request of the outside incident in time and can accomplish the disposal to the incident in the stated time, at the same time, it can control Real-time task running in phase. The application of Real-time embedded systems is extensive, including experiment control 、process control device 、robot 、traffic management of air 、tekecommunications 、military command and control systems, next generation systems will include automobile 、robot controller which owns elasticity arthrosis 、space station 、benthal reconnoiter and so on [4]. The characteristics of Real-time embedded systems are as follows:

(1) Purpose or pertinency, namely embedded systems are designed for the purpose of implementing special function or applying to the special situation. This is the difference between Real-time embedded systems and general control systems.

(2) The real-time quality respond to the outside incident, namely it can accomplish the disposal to the incident in the stated time.

(3) Satisfy certain peak value load request, namely some key tasks can be satisfied in the situation that load is heavy, even overload.

(4) Assure the reliability of system running. The design of Real-time embedded systems general adopts fault-tolerant and anti-interference technology.

Diminutive embedded systems can adopt process-oriented development means to develop sequence program according to the time sequence of system running and function demand. The embedded systems' real-time quality is high and its structure is simple, generally, which can be accomplished by hardware technician. The design and developing of embedded systems whose function is complicated, more and more, inclines to adopt object-oriented means, the whole course is divided into four stages---analyse 、design 、transform and test. Analyse stage confirms the basic characters of the right solution; design stage adds some elements to the analyse stage's result and defines specially solution according to optimized rules; transform stage builds the design which is executable and disposable; test stage checks if the result of transform stage is accord to the design and if the software satisfies all the validity rules of the design stage.

From the above analyse, we can draw a conclusion that Real-time embedded systems are very suitable for the application in the secure gateway, since systems' function demands powerful, we decide to adopt the object-oriented development means.

2.2 Choice of Real-time Operating System

VxWorks operating system is a high performance embedded Real-time operating system developed by WinRiver Company of U.S.A in 1983. It is a multitask operating system and its main characteristics are as follows: highly cut out microkernel structure; efficient multitask scheduling; supports interrupt drive preferential competition scheduling and timeslice cycle scheduling; owns determinate and quick context switch ability; flexible communication means among tasks; communication among processes and mutex means such as signal lamp 、message queue 、pipeline 、signal and socket and so on; definite microsecond interrupt delay time; TCP/IP Protocol Suite and common application protocol based on TCP/IP Protocol; abundant board support package (BSP), supports many processor boards; supports POSIX1003.1b; clipping and flexible I/O systems; supports many files systems; flexible startup ways, supports startup from ROM 、local disk and network. Since VxWorks owns so many merits, we decide to adopt VxWorks as the operating system of the secure gateway.

3. ANALYSE OF ENCRYPTION ALGORITHM AND HARDWARE IMPLEMENT

The main purpose of secure gateway is to protect transferred data, and one good way to implement is to encrypt the data. Which encryption algorithm should we choose and how to implement it? The two question will be solved through the following analyse and research.

3.1 Analyse of Encryption Algorithm

We will analyze the modern encryption, and decide to adopt which encryption algorithm. There are two encryptions---one is symmetric encryption, the other is asymmetric encryption. Symmetric encryption demand the same key for encryption and decryption, so sender and recipient must hold identical key, and merits of these algorithms are generally fast and compact. Asymmetric encryption uses different keys for encryption and decryption. The key for encryption is public, that is anyone can know it and can use it to encrypt data, while the key for decryption is secret, that is only the intended recipient may read the message, in that the knowledge of public encryption key does not permit nor assist decryption. The demerits of these algorithms are slow response and taking up too much resource.

Key management deals with the secure generation, distribution, and storage of keys. Symmetric ciphers require that the same key is available at both sides, and key is often the most vulnerable aspect of a security system, so it is vital of data security system to distribute these keys in a secure manner.

DES(Data Encryption Standard) algorithm was originally developed by IBM in the early 1970s, and it is adopted by NIST as the "Data Encryption Standard" in 1977. After that, it becomes the most popular encryption standard in the world. DES algorithm adopts 64-bit data block encryption considering its speed. Encryption and decryption operations are almost identical and easily to implement in hardware. The detailed DES algorithm be shown in reference [5].

RSA algorithm was invented by Ron Rivest, Adi Shamir and Leonard Adleman in 1977, which is a asymmetric encryption algorithm based on large prime that is difficult to be decompounded, and it is the basis for public key cryptography [6]. However, it is at least 1000 times slower than DES or AES, so it is generally used for low bandwidth key exchange only and can't be applied in wirespeed data encryption. Here's the relatively easy to understand math behind RSA public key encryption.

(1) Find P and Q, two large (e.g., 1024-bit) prime numbers.
(2) Choose E such that E is greater than 1, E is less than PQ, and E and (P-1)(Q-1) are relatively prime, which means they have no prime factors in common. E doesn't have to be prime numbers, but it must be odd. (P-1)(Q-1) can't be prime numbers because it's an even number.
(3) Compute D such that DE-1 is evenly divisible by (P-1)(Q-1). Mathematicians write this as DE = 1 (mod (P-1)(Q-1)), and they call D the multiplicative inverse of E. This is easy to do- -- simply find an integer X which causes D = (X(P-1)(Q-1) + 1)/E to be an integer, then use that value of D.
(4) The encryption function is C = (T^E) mod PQ, where C is the ciphertext (a positive integer), T is the plaintext (a positive integer), and ^ indicates exponentiation. The message being encrypted, T must be less than the modulus--- PQ.
(5) The decryption function is T = (C^D) mod PQ, where C is the ciphertext (a positive integer), T is the plaintext (a positive integer), and ^ indicates exponentiation.

The public key is the pair (PQ, E). The private key is the number D (reveal it to no one). The product PQ is the modulus (often called N in the literature). E is the public exponent. D is the secret exponent.

We can publish public key freely, because there are no known easy methods of calculating D, P, or Q given only (PQ, E) (the public key). If P and Q are each 1024 bits long, the sun will burn out before the most powerful computers presently in existence can decompound modulus into P and Q.

DES algorithm and RSA algorithm are respectively the typical representations of symmetric encryption and asymmetric encryption because of different mechanisms and algorithms, which make them have merits and demerits of their owns. We will compare them in the following.

(1) DES algorithm is superior to RSA algorithm in the aspect of encryption speed. Since the length of DES key is only 56 bits, its implement speed is very fast; while RSA algorithm need carry through disposing power of n-bit integer and module, so its disposing speed is obvious slower than DES algorithm. With the increasing network congesting, RSA algorithm isn't fit for the encryption of long plaintext.
(2) RSA algorithm is superior to DES algorithm in the aspect of key management. Since RSA algorithm publish its public key, it is very easy for it to update the key, receiver only need keep his own private key secret for different communication objects; however, DES algorithm demands distributing key in secret before communication, it is difficult to replace the key, at the same time, DES need generate and keep different keys according to different communication objects, moreover, it is different to ensure the security of transferring.

(3) DES algorithm and RSA algorithm are both good in the aspect of security. There is still not efficient way for decryption the ciphertext which is encrypted by DES algorithm or RSA algorithm at present. From the above comparing, we can see that both DES algorithm and RSA algorithm have merits, so we can develop a new real-time embedded encryption scheme which integrates the merits and avoid of the demerits of the two algorithms. The fundamental is that we encrypt plaintext using DES algorithm, then encrypt the key of DES algorithm using RSA algorithm, finally sent the two parts to receiver (shown in figure 2). The particular process is as followed:

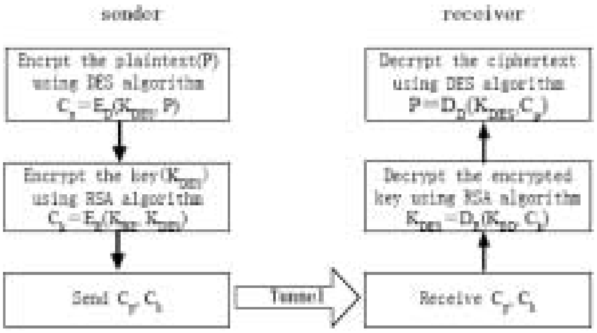

sender
receiver
Encrpt the plaintext(P) using DES algorithm $C_P=E_D(K_{DES},P)$
Encrypt the key(K_{DES}) using RSA algorithm $C_K=E_R(K_{BE},K_{DES})$
Send C_P, C_K
Tunnel
Receive C_P, C_K
Decrypt the encrypted key using RSA algorithm $K_{DES}=D_R(K_{BD},C_K)$
Decrypt the ciphertext using DES algorithm $P=D_D(K_{DES},C_P)$

Figure 2. DES-RSA encryption/decryption process

(1) The sender generates the encryption key (k_{DES}) of DES algorithm, then encrypts the plaintext (P) using the k_{DES}, and generates cryptograph (C_P).
(2) The sender gets the public key (K_{BE}) of the receiver, and encrypts the k_{DES} using K_{BE} and obtains the cryptograph C_K.
(3) The sender sends the C_P and C_K to the receiver through the Internet.
(4) The receiver decrypts C_K using the private key (K_{BD}) and obtains the sender's key ---k_{DES}.
(5) The receiver decrypts C_P using the k_{DES}, and obtains the plaintext---P.

3.2 Selection of Encryption Hardware

Since IPSec implements the network security in VPN, it needs the support of control channel and data channel. Control channel and the key exchange can be implemented using software in RISC engine (such as PowerPC processor). Data channel need implement the encryption and decryption algorithm of AES and RSA in hundred of million bits percent second (even Gbps), but it is impossible for software to achieve it. The standard commercial CPU and DSP can't satisfy the demand of data encryption algorithm. Network processor (NP) can't satisfy the high-powered network security, at the same time, several ASSP in market also doesn't bring the hoped performance. However, FPGA just can satisfy the demand because of its unique performance

FPGA is a half-customization's ASIC, and it has many merits---ultimate programmable ability, design flexibility, convenient upgrade, low power expense and quickly marketability. For example, FPGA special algorithm can achieve 990Mbps originality throughout in 66MHz frequency. New platform FPGA production---Xilinx Virtex-II Pro device can achieve deal with ability with speed of line in the same chip, which also achieves network process function, and accomplishes flexible control process of network security [7]. The FPGA devices comprise embedded PowerPC processor and programmable array structure adapted to requisite algorithm, which is the key factor of realizing high performance security process. FPGA and IP provider provided the DES algorithm and RSA algorithm, which makes the design easier.

4. CONCRETELY IMPLEMENT OF SECURE GATEWAY BASED ON REAL-TIME EMBEDDED SYSTEMS

Concretely implement of secure gateway based on Real-time embedded systems conclude hardware implement 、FPGA software implement 、FPGA driver implement 、VxWorks application implement and so on. We just introduce the hardware implement and parts of FPGA encryption module because of restrict of paper length.

4.1 Hardware Implement

We designed a secure gateway based on Real-time embedded systems, whose main hardware includes Intel IXP425, Xilinx Virtex-II Pro, SDRAM, FLASH and two LXT972 chips. The hardware frame is shown in figure 3. The SDRAM connected with IXP425's SDRAM controller interface; FLASH connected with IXP425's extended bus controller interface; FPGA connected with IXP425's PCI slot; two LXT972 connected with MII.

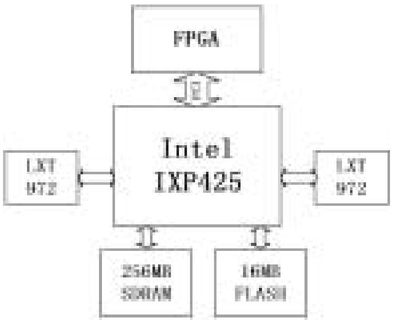


Figure 3. Hardware frame of secure gateway based on Real-time embedded systems

The Intel IXP425 network processor family is designed to meet the needs of broadband and embedded network products such as high-end residential gateway; The Intel IXP425 network processors deliver wirespeed performance and sufficient "processing headroom" for manufacturers to add a variety of rich software services to support their applications. These are highly integrated network processors that support multiple WAN and LAN technologies giving customers a common architecture for multiple applications. This network processor family offers the choice of multiple clock speeds at 266, 400, and 533 MHz, with both Commercial (0° to 70° C) and Extended (-40° to 85° C) temperature options [8]. We can select Xilinx Virtex-II Pro family as FPGA, whose main purpose is to implement the function of encryption and decryption of the data. LXT972 is an interface chip of secure gateway and Internet or Intranet.

4.2 Encryption Software Module

Because DES-RSA encryption/decryption algorithm is implemented by FPGA, the program language need use hardware description language (e.g., VHDL or Verilog HDL). The encryption and decryption of DES algorithm and RSA algorithm can use the same module because of their particularities, and the only difference lies in the value of control parameter. Taking the RSA algorithm module as an example, we introduce the encryption module (it also can be used as decryption module). RSA encryption and decryption process can be implemented using two modules (main module and modular exponentiation module), which are shown in figure 4 and figure 5.

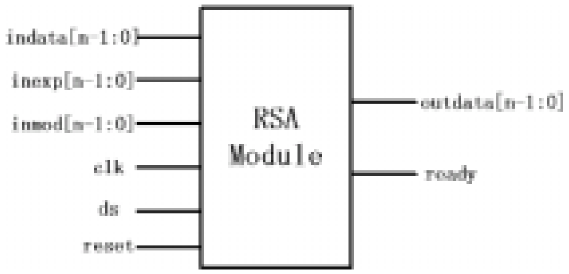


Figure 4. Main module of RSA algorithm

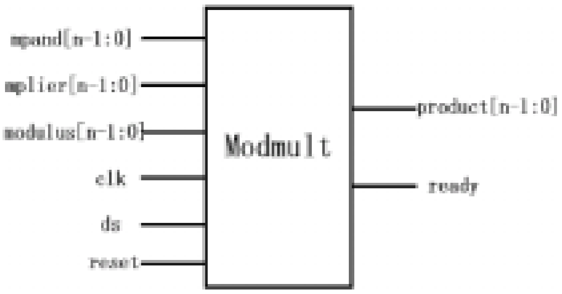


Figure 5. Modular exponentiation module of RSA algorithm

N indicates the length of key of RSA algorithm (e.g., 512 、1024 or 2048) in figure 4 and figure5. Indata indicates the encrypted data; inexp indicates the key; inmod indicates the module (it is the PQ in RSA algorithm above mentioned); outdata is the data which has been encrypted; clk is clock; ds is clear signal; reset is reset signal; mpand is multiplicand; mplier is multiplier; modulus is model; product is the result of the algorithm.

4.3 Typical Application

With the expanding of enterprise operations, the Intranet is also expanding, for example some enterprise LAN becomes WAN even transnational network. Using the secure gateway designed by us can adequately utilize the Internet resource and build enterprise's VPN, thus, enterprises can both save cost and achieve the advantages of security 、easily manage and easily expand. Figure 6 shows how to connect the secure gateway to enterprise VPN.

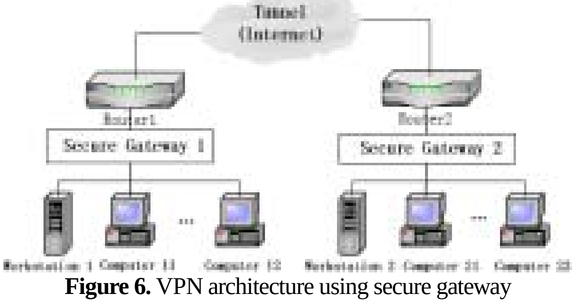


Figure 6. VPN architecture using secure gateway based on Real-time embedded systems

5. CONCLUSIONS

Through the analyse and research about the VPN security protocol 、Real-time embedded systems and encryption algorithm, we designed a secure gateway based on Real-time embedded systems, which adopts the encryption algorithm combined DES algorithm and RSA algorithm, and realizes it using FPGA, consequently, reduces the burden of CPU and improves the whole performance of systems.

Of course, because of the complexity of DES algorithm and RSA algorithm, the encryption speed is the bottleneck all the time, with the further research of encryption algorithm, more effective encryption algorithm will continually come forth. However, since the encryption modules are realized through FPGA, it can adapt the continual varieties, so the secure gateway based on Real-time embedded systems owns promising prospect.

6. REFERENCES

[1] Xiaobing Peng, Quan Liu, Shufan Yang. Design of a Virtual Private Network based on Ptolemy. Wuhan: Journal of Wuhan University of Technology, 2003(3): 24-27. (In Chinese)

[2] Quan Liu, Yiwen Zhu, Fangmin Li. Information Security Frames based on VPN Technology. Wuhan: Journal of Wuhan University of Technology, 2003(5): 62-64. (In Chinese)

[3] Lee EA. Actor-oriented Design of Embedded Hardware and Software System [J]. Journal of Circuits Systems and Computer, 2002,11(5): 21-25.

[4] Sangiovanni-Vincentelli A, Martion G A. Vision for Embedded Systems; Platform-Based Design and Software Methodocosy [J]. IEEE Design and Test of Conpulers, 2001,18(6): 22-33.

[5] National Institute of Standard and Technology. Data Encryption Standard (DES). Federal Information Processing Standards Publication 46-3 (FIPS PUB 46-3), 1999.

[6] Rivest R L, Shamir A, Adleman L. A method for obtaining Digital Dignatures and Public-key Cryptosystems. Communications of the ACM, 1978, 21(2):120～126.

[7] http://www.xilinx.com.

[8] http://www.intel.com.

Yufeng Wu, male, was born in Nov 1980, is a postgraduate in School of Information Engineering, Wuhan University of Technology, his research direction is System and Signal Processing.

Quan Liu, female, professor, supervisor of doctor, is the dean of the School of Information Engineering, Wuhan University of Technology, her research interests are information security, signal processing, communication technology, grid computing and network security.

Data Encryption Algorithms for Internet-based Real-Time Systems

Li Hongyan [1,2], Shuang H. Yang[1,3], and Tan LianSheng [1]

[1]Department of Computer Science, Central China Normal University, Wuhan, China.

[2] Department of Computer and Electronic Science, Hubei University of Economics, Wuhan, China

[3]Department of Computer Science, Loughborough University, Loughborough, Leicestershire LE113TU, UK

Email: lhywawa@sohu.com, s.h.yang@lboro.ac.uk, 1.tan@ccnu.edu.cn

ABSTRACT

In recent years, the Internet has proved to be a powerful tool for real-time applications. However, security risk of the Internet communication still stops people to bring the real-time application into a reality. Little work has so far been done in developing a data encryption algorithm for Internet-based real-time applications. In order to satisfy the security requirements of Internet-based real-time systems, two hybrid data encryption algorithms are proposed. One is the combination of the Advanced Encryption Standard (AES) and the most popular public-key cryptography (RSA); the other is the combination of the AES and Secure Sockets Layer (SSL). The end-to-end encryption latency of different algorithms is investigated to show the efficiency of the two new algorithms for Internet-based real-time applications.

Keywords: real-time Systems, Encryption Algorithm, AES, RSA, SSL.

1. INTRODUCTION

Dealing with the Internet time delay and security are two major challenges facing in the design of Internet-based real-time systems. How to improve or reduce the time delay of real-time data transmission has been received much attention in the recent years. To date, many literatures focus on developing real-time applications [1], such as developing systematic design methods for the design of Internet-based process control systems [2-4]. Most research on real-time applications emphasized the real-time scheduling strategies [5-6]. The research in the security aspect more focuses on some typical encryption algorithm's reliability and security [7-8]. Few of them take a consideration on the real time data transmission security. In fact, existing data encryption algorithms are not appropriate for Internet-based real-time systems, as they take too long time in data encryption and decryption.

This paper aims to develop data encryption algorithms for Internet-based real-time systems and to demonstrate their efficiency through the testing of the network latency for real-time data transmission. The paper is organized as follows. End-to-end real-time data transmission architecture is specified in Section 2. Several typical data encryption algorithms are briefly introduced in Section 3. Two hybrid encryption algorithms for real-time systems are proposed in Sections 4 and 5. The end-to-end latency of different encryption algorithms is compared in Section 6. Section 7 contains conclusions.

2. END-TO-END REAL-TIME TRANSMISSION ARCHITECTURE

The real-time applications are normally running on the UDP layer due to its outstanding transmission performance. In our previous work [9], the end-to-end real-time transmission architecture absorbs the idea of the existing real-time technologies and mainly implements at the transport layer, as shown in Figure 1. Real time Control Protocol (RCP) is used to solve the reliability and real-time issue for real-time data transmission. The transmission architecture uses TCP to establish the connection instead of UDP because any UDP packet might be blocked by the firewall. The Resource Reservation Protocol (RSVP) allows applications to reserve bandwidth and enables the applications to obtain differing QoS for their data flow [10]. Security is a critical issue in the end-to-end real-time data transmission architecture. The real-time plain data without encryption or security protection could be intercepted by web hackers. IPSEC and SSL are two commonly used security alternatives. However, IPSEC operates at the network layer and must be supported by the operating system and SSL is based on TCP, which is not available to the UDP socket. In order to satisfy the real-time requirements of the communication data, the data encryption must be fast enough and security enough as well. So a new data encryption algorithm for real-time systems is required.

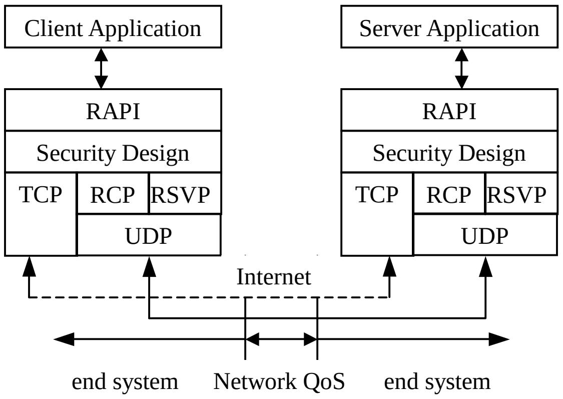


Figure 1: end-to-end real-time transmission architecture

3. TYPICAL DATA ENCRYPTION ALGORITHMS

There are two kinds of cryptosystems: symmetric and asymmetric. Symmetric cryptosystems use the same key to encrypt and decrypt a message, and asymmetric cryptosystems use one key to encrypt a message and another different key to decrypt it. In this section, three typical data encryption algorithms are briefly reviewed. They are the Rijndael (AES Encryption) algorithm, which is used in symmetric cryptosystems, the RSA (Rivest-Shamir-Adleman) algorithm, which is the most popular algorithm of asymmetric cryptosystems, and the SSL (Internet Security Protocal) algorithm, which provides user an authentication mechanism

* Corresponding author, Hong-Yan Li, master in Computer Science Department at Central China Normal University.

and prevents unauthorized access to the network.

Rijndael Algorithm

The Advanced Encryption Standard chose the Rijndael algorithm as its standard algorithm, and is the next-generation defense against malicious persons tampering with confidential data. Concerning the current standards and rates of technological progress, it can be said that the Rijndael algorithm is undecipherable. This algorithm aims to remove the weak link of cryptography from the security equation. The beauty of the Rijndael algorithm is that a system running the algorithm is not necessary to have a great deal of processing power or memory. All of the steps in the data encryption are simple matrix operations. Additionally, each step can easily be reversed, provided that the user has a key. The Rijndael algorithm is able to diffuse the changes that the key makes to the encrypted file, and therefore it is difficult to trace the changes.

RSA Algorithm

The RSA algorithm is named after Ron Rivest, Adi Shamir and Len Adleman, who invented the algorithm in 1977. It is a commonly used public-key cipher algorithm. Because the RSA algorithm uses a public key to encrypt a file and uses a private key to decrypt it, it is easy to manage the RSA cipher. Because of the slow rate of encryption/decryption, RSA is not suitable for abundant data encryption. In this paper the RSA algorithm will be cooperated with a symmetric-key cryptography in encrypting/decrypting data in order to achieve an ideal encryption implementation.

SSL

Secure sockets layer (SSL) is a security protocol securing the socket between the application layer and the transport layer. It provides a user authentication mechanism and encryption using shared session keys on the media streams and a public key cryptography to distribute the session keys. One party uses SSL to authenticate another party's identity and generate the session key, encrypt or decrypt the data with the key, and direct the data through the SSL session connection to another party. However, the SSL is based on TCP and secures only the TCP socket, which is not suitable for transmitting real-time data on UDP.

4. HYBRID ALGORITHM BASED ON AES AND RSA

In order to meet the real-time requirements of the communication data, the encryption must be fast enough and security as well. The AES algorithm secretly assigns the cipher before the real-time data transmission and the AES cipher is t hen sent to the receiver through Internet.

However, in the process of RSA encryption, the private key is saved in the receiver and the public key is transmitted to the sender at the same time. The RSA algorithm adopts the public key to encrypt the data and the private key to decrypt it. It is impossible to determine the private key from the public key. Owing to the private key never be transported, the security of the cipher for the RSA algorithm is much higher than that of AES algorithm apparently. The security of the RSA algorithm comes from the computational difficulty of factoring large numbers, and costs much more time for encryption/decryption data than AES.

Due to the low encryption/decryption rate of the RSA algorithm, it is unsuitable for encrypting abundant data. In contrast, the AES algorithm encrypts data only using simple matrix operations and has a high rate of encryption/decryption, but the AES cipher is difficult. If we can use the AES algorithm to encrypt the real-time data and use the RSA algorithm to encrypt the AES cipher, the hybrid real-time data encryption algorithm might absorb the advantages of AES and RSA algorithms and avoids their disadvantages.

The principle of this hybrid algorithm is as follows: before sending the real-time data, the sender uses the AES algorithm to encrypt these data and uses the RSA algorithm to encrypt the AES cipher. The encrypted AES cipher and the cipher-text are then sent to the receiver together. After receiving the encrypted AES cipher and the cipher-text, the receiver uses the RSA private key to decrypt the encrypted AES cipher, and then uses the AES cipher to decrypt the cipher-text. The information flow is illustrated in Figure 2. The sender endpoint implements the real-time data encryption and the receiver endpoint implements the decryption. To ensure the security of the AES cipher, the receiver should first create a RSA key pair.

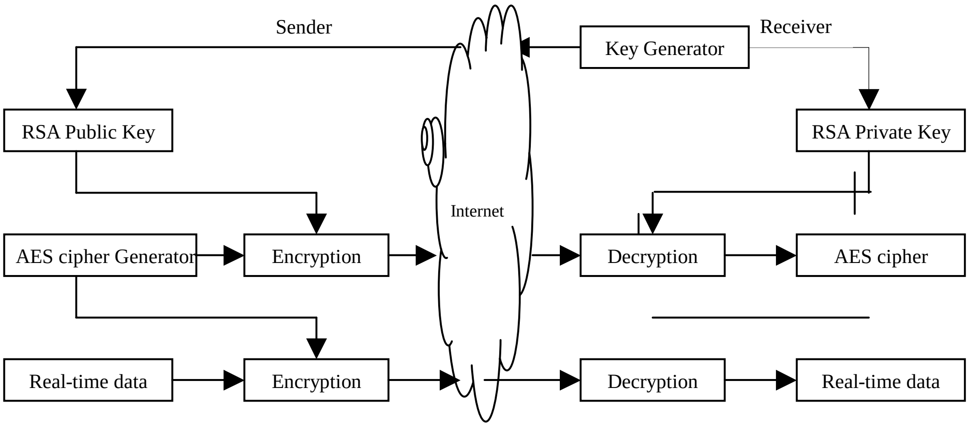


Figure 2: Information flow of the hybrid algorithm using AES and RSA

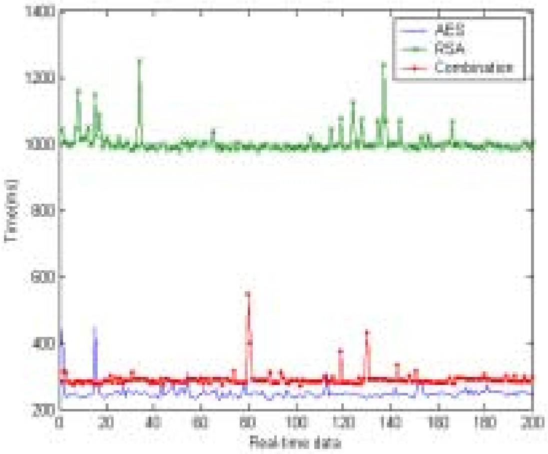


Figure 3: Comparison of the end-to-end latency

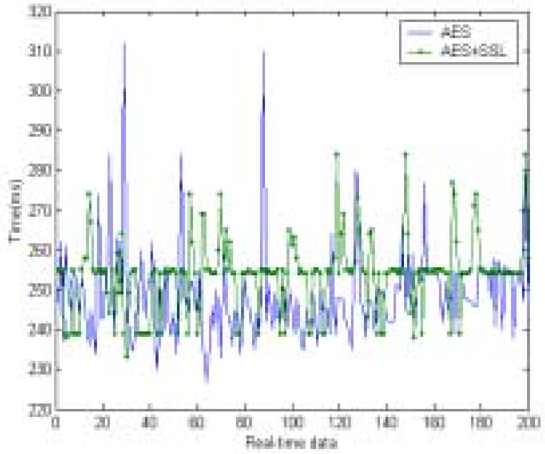


Figure 5: Comparison of end-to-end latency

In order to evaluate the hybrid encryption algorithm, the AES algorithm, the RSA algorithm, and the hybrid encryption algorithm are compared in a similar network environment. The total time delay is composed of three parts: the data encryption time delay, the encrypted data network transmission time delay and the data decryption time delay. If the network environment and the processing power are similar the total time delay will depend on the data encryption/decryption time delays. From Figure 3 it is observed that the total time delay of the hybrid algorithm is close to that of the AES and it is much less than that of the RSA. Furthermore, the AES cipher of the hybrid algorithm is encrypted by the RSA, it is much hard to be deciphered. Therefore, this hybrid algorithm provides a promising way to assume the real-time data encryption/decryption secure and fast enough.

5. HYBRID ALGORITHM USING AES AND SSL

There are two ways to secure the network communication: preventing unauthorized access to the network or encrypting messages sent over the network. The SSL security protocol provides data encryption, server authentication, message integrity, and optional client authentication for a TCP/IP connection. The SSL is used to transfer the AES cipher. Thus, if we use the AES to encrypt the real-time data and the SSL to transfer the AES cipher, the hybrid real-time data encryption algorithm can prevent unauthorized access to the network and encrypt the real-time data as well.

The principle of this hybrid algorithm is as follows: before sending the real-time data, the sender uses the AES algorithm to encrypt the real-time data and uses the SSL to transfer the AES cipher at the same time. The receiver uses the accepted AES cipher to decrypt the encrypted real-time data based on the UDP transmission. The real-time data is securely transmitted on the UDP. The information flow is illustrated in Figure 4. The sender endpoint implements the real-time data encryption and the receiver endpoint implements the decryption. The sender endpoint firstly sends the SSL connection request and establishes the SSL connection after authentication. The AES cipher created by the sender is then transferred to the receiver through the SSL connection. The real-time data is encrypted by the AES cipher and then transferred to the receiver through the Internet. The receiver accepts the SSL connection request after authentication, and gets the AES cipher through the SSL connection. Finally the cipher text is decrypted by the AES cipher and the real-time data is recovered.

This hybrid algorithm involves the server authentication and the client authentication. The AES algorithm and this hybrid algorithm are implemented and compared. The experimental results are illustrated in Figure 5. As Figure 5 shown, this hybrid algorithm and the AES algorithm have a similar end-to-end latency. But the hybrid algorithm can prevent unauthorized access to the network. It is an ideal algorithm for data encryption of Internet-based real-time systems.

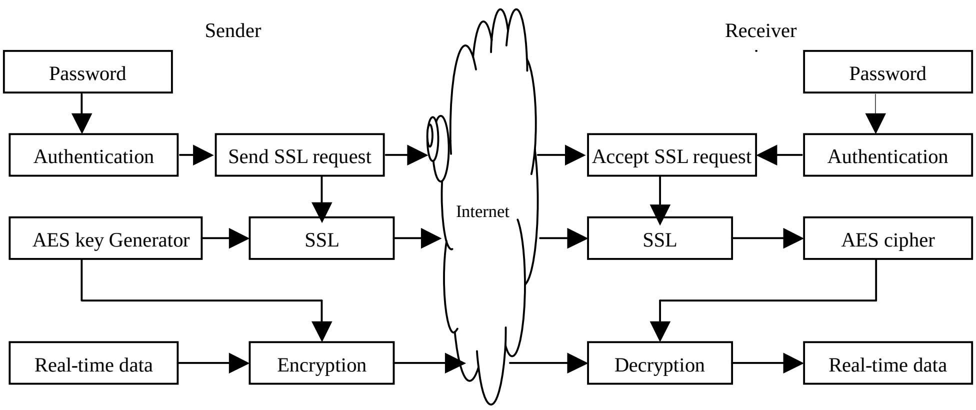


Figure 4: Information flow of the hybrid algorithm using AES and SSL

6. DISCUSSION OF THE EXPERIMENTAL RESULTS

The experimental statistic results are summarized in Table1. The end-to-end transmission latency is composed of the execution time for data encryption, the encrypted data network transmission latency and the execution time for data decryption. The average end-to-end latency indicates the normal operation period. The maximum latency and minimum latency illustrate the existence of the unpredictability of the Internet transmission. Table 1 illustrates the advantages and disadvantages of the four possible data encryption/decryption algorithms. As showed in Table 1, the end-to-end latency of two hybrid algorithms is close to that of the AES algorithm and much shorter than the RSA algorithm, and the lower average latency for these two hybrid algorithms, 288.94 and 255.47, indicates that they are suitable for data encryption/decryption for real-time systems. Considering the two hybrid algorithms, the one using the AES and the SSL algorithms can prevent unauthorized access to the network and bring lower latency as well. But it can’t ensure the AES cipher and cipher-text simultaneously. Both of them use the AES algorithm to encrypt and decrypt the AES cipher.

7. CONCLUSIONS

The Internet has brought a great potential for real-time applications. Security of real-time data transmission over the Internet is one of the obstacles to bring the real-time Internet based applications into a reality. The data encryption and decryption for Internet-based real-time systems must satisfy the real-time requirements and secure enough as well. This paper provides two novel data encryption/decryption algorithms for real-time systems. One is the combination of the AES and RSA algorithms; the other is the combination of the AES and SSL algorithms. The principle and implementation of these two hybrid algorithms are introduced. The experimental results on the end-to-end latency for real-time data encryption/decryption show that the two hybrid algorithms have a great potential to assume the security and satisfy the real-time requirements as well.

8. REFERENCES

[1] S. Rudkin, A. Grace and M. W. Whybray, “Real-time applications on the Internet”, BT Technol. J., Vol. 15, No. 2, 1997, pp. 209-225.

[2] S.H. Yang, X. Chen, and J.L. Alty, “Design issues and implementation of Internet-based process control systems,” Control Engineering Practice, No.11, 2003, pp. 709-720.

[3] S.H. Yang, J.L. Alty, ”Development of a Distributed Simulator for Control Experiments through the Internet”, Future Generation Computer System, Vol. 18, No.5, 2002, pp.595-611.

[4] S.H. Yang, X. Chen, D.W. Edwards, and J. L. Alty, “Designing Internet-based control systems for process plants”, The 4th Asian Control Conference, Singapore, September 2002.

[5] D. C. Schmidt, R. Bector, D. L. Levine, S. Mungee, and G. Parulkar, “An ORB Endsystem Architecture for Statically Scheduled Real-time Applications,” in Proceedings of the Workshop on Middleware for Real-Time Systems and Services, (San Francisco, CA), IEEE, December 1997.

[6] D. C. Schmidt, D. L. Levine, and S. Mungee, “The Design and Performance of Real-Time Object Request Brokers,” Computer Communications, Vol. 21, Apr. 1998, pp. 294–324.

[7] J. Daemen, and V. Rijmen, AES Proposal: Rijndael, 1999. http://csrc.nist.gov/CryptoToolkit/aes/rijndael/Rijndael-ammended.pdf

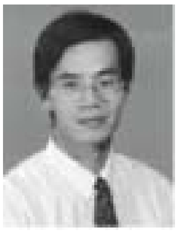
[8] RSA Algorithm, http://www.di-mgt.com.au/rsa_alg.html

[9] X. Chen and S.H. Yang, Control Perspective for Virtual Process Plants, Proceedings of the 5th Asia-Pacific Conference on Control & Measurement, 8-12 July, 2002.Dali, China, pp. 256-261.

[10] R. Braden, “Resource Reservation Protocol (RSVP)-version 1 functional specification”, IETF, RSVP working group work in progress (draft-ietf-rsvp-spec-14)(November 1996).

Table1: Comparison of the experimental results

Algorithms	End-to-end real-time transmission latency			Advantages	Disadvantages
	Average latency (ms)	Maximum latency (ms)	Minimum latency (ms)		
RSA	1003.46	1250	978	Cipher more safety	High latency
AES	249.57	446	227	Low latency	Cipher not safety enough
Combination of AES and RSA	288.94	544	270	Cipher more safety; low latency; AES cipher and cipher-text synchronous	No authentication
Combination of AES and SSL	255.47	284	236	Cipher more safety; low latency; access by authentication	Can’t ensure AES cipher and cipher-text synchronous



Shuang H. Yang is a lecturer in Computer Science Department at Loughborough University. He is also an overseas professor in Computer Science Department at Central China Normal University, People's Republic of China. His research interests include Internet-based control, safety critical systems and artificial intelligence. He received his PhD in Control Engineering from Zhejiang University in 1991. He is a member of Institute of Measurement and Control (MinstMC) and a Charted Engineer (CEng) in UK.

Lian-Sheng TAN was born in 1965. He is a full professor now and the Head of the Department of Computer Science, Central China Normal University. Dr. Tan obtained his Ph.D. degree from Loughborough University in UK in 1999. He was a postdoctoral research fellow doing researching in computer networks with School of Information Technology and Engineering, University of Ottawa, Canada in 2001. His research interests are in modeling, congestion control analysis, and performance evaluation of computer communication networks.

Hong-Yan Li is a master at the Department of Computer Science, Central China Normal University. Her research area is Internet-based process control. She is also a teacher at Department of Computer and Electronic Science, Hubei University of Economics.

Research on Grid Security for OGSA*

Zuguang Fan, Fangmin Li, Quan Liu
School of Information Engineering, Wuhan University of Technology
Wuhan 430070, China
Email: lifangmin@mail.whut.edu.cn, fzg007294@sohu.com, qliu@pubilc.wh.hb.cn
Tel: 13973255023, 13545068921, 13908659571

ABSTRACT

Grid computing is concerned with the sharing and coordinated use of diverse resource in distributed "virtual organization." The dynamic and multi-institutional nature of VOs (virtual organizations) presents challenging security issues that need new security-enabled schemes. In this paper ,we make a brief discussion about Security Architecture within OGSA(Open Grid Service Architecture) , we then describe in detail approaches developed to support the GT3 implementation of the Open Grid Services Architecture, a new initiative aimed at recasting key Grid concepts within a service-oriented framework., and mainly analyze the GT3 grid security infrastructure both from the point of the resource and the user to illustrate the concrete security implementation based on Web services security mechanisms for credential exchange and other purposes.

Keywords: Grid, OGSA, Web services, GT3 (globus toolkit 3), GSI, VOs (virtual organizations)

1. INTRODUCTION

The grid regarded as the next-generation internet has attracted more and more attention from all over the world, the term "grid" refers to systems and applications that integrate and manage resources and services distributed across multiple control domains [1] [2].

Grid computing has emerged as an important new field, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications and high-performance orientation. A common scenario within Grid computing involves the formation of dynamic "virtual organizations" comprising groups of individuals and associated resources and services united by a common purpose but not located within a single administrative domain. What distinguishes a VO from a classical organization is that it may gather individuals and/or institutions that have agreed to share resources and otherwise collaborate on an ad-hoc, dynamic basis, while they continue to belong to different real organizations, each governed by their own set of internal rules and policies.

The need to support the integration and management of resources within VOs introduces challenging security issues. Grid computing research has produced security technologies based on the use of VO as a bridge among the entities participating in a particular community or function. Resorting to the results of the research, the widely-used software system called the Globus Toolkit (GT) builds its primary security infrastructure –GSI(Grid Security Infrastructure) which is name given to the part of the Globus Toolkit that implement security functionality.

Web services technology have had a set of policies, protocols, and methods to implement its security,, integration of GSI with OGSA make it possible for grid computing to take advantage of the emerging Wed services security technologies.

2. SECURITY CHALLENGES IN GRID ENVIRONMENT

The security challenge faced in a Grid environment can be summarized into three points: [6]

(1) integration solutions where existing services needs to be used, and interfaces should be abstracted to provide an extensible architecture;
(2) interoperability solutions so that services hosted in different virtual organizations that have different security mechanisms and policies will be able to invoke each other; and
(3) solutions to define, manage and enforce trust policies within a dynamic Grid environment.

A solution within a given category will often depend on a solution in another category.. For example, any solution for federating credentials to achieve interoperability will be dependent on the trust models defined within the participating domains and the level of integration of the services within a domain. Defining a trust model is the basis for interoperability but trust model is independent of interoperability characteristics. Similarly level of integration implies a level of trust as well has a bearing on interoperability.

2.1 The Integration Challenge

For both technical and pragmatic reasons, it is unreasonable to expect that a single security technology can be defined that will both address all Grid security challenges and be adopted in every hosting environment. Existing security infrastructures cannot be replaced overight , the feasible way is that we must implement the security functionalities based existing security infrastructure.

Thus, to be successful, a Grid security architecture needs to step up to the challenge of integrating with existing security architectures and models across platforms and hosting environments. This means that the architecture must be implementation agnostic, so that it can be instantiated in terms of any existing security mechanisms (e.g., Kerberos, PKI); extensible, so that it can incorporate new security services as they become available; and integratable with existing security services.

2.2 The Interoperability Challenge

Services that traverse multiple domains and hosting

* Supported by China Hunan NSF #03JJY1012.

environments need to be able to interact with each other, thus introducing the need for interoperability at multiple levels [5]:

(1) At the protocol level, we require mechanisms that allow domains to exchange messages. This can be achieved via SOAP/HTTP, for example.

(2) At the policy level, secure interoperability requires that each party be able to specify any policy it may wish in order to engage in a secure conversation—and that policies expressed by different parties can be made mutually comprehensible. Only then can the parties attempt to establish a secure communication channel and security context upon mutual authentication, trust relationship, and adherence to each other’s policy.

(3) At the identity level, we require mechanisms for identifying a user from one domain in another domain. This requirement goes beyond the need to define trust relationships and achieve federation between security mechanisms (e.g., from Kerberos tickets to X.509 certificates). Irrespective of the authentication and authorization model, which can be group-based, role-based or other attribute-based, many models rely on the notion of an identity for reasons including authorization and accountability. It would be nice if a given identity could be (pre)defined across all participating domains, but that is not realistic in practice.

2.3 The Trust Relationship Challenge

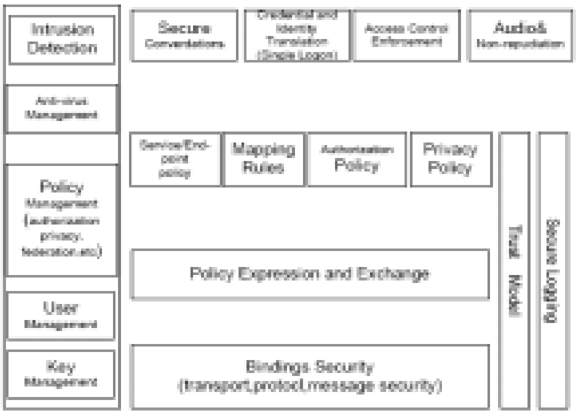
Generally, grid service request must span multiple security domains, so trust relationship among these domains play an important role in the outcome of such end-to-end traversals. A service needs to make its access requirements available to interested client entities, so that they understand how to secure request access to it. The trust relationship problem is made more difficult in a Grid environment by the need to support dynamic, user-controlled deployment and management of transient services.

3. GRID SECURITY MODEL

Industry efforts have rallied around Web services (WS) as an emerging architecture that has the ability to deliver integrated, interoperable solutions. Ensuring the integrity, confidentiality and security of Web services through the application of a comprehensive security model is critical, both for organizations and their customers – which is the fundamental starting point for constructing virtual organizations.

The security of a Grid environment must take into account the security of various aspects involved in a Grid service invocation. This is depicted in the Figure 1[6]:

The Grid security model groups all security management functions applicable to various aspects of binding, policy and federation. Addressing the management of various aspects of the security infrastructure will satisfy the manageability requirement on the Grid environment.The security model must provide a mechanism by which authentication credentials from the service requestor’s domain can bu transkated into the service provider’s and vice versa. This translation is required in order for both ends to envaluste their mutual access policies based on the establisjed credentials and the quality of the established channel.



Intrusion Detection
Anti-virus Management
Policy Management (authorization privacy, federation,etc)
User Management
Key Management
Secure Conversations
Credential and Identity Translation (Single Logon)
Access Control Enforcement
Audio& Non-repudiation
Service/End-point policy
Mapping Rules
Authorization Policy
Privacy Policy
Policy Expression and Exchange
Bindings Security (transport,protocl,message security)
Trust Model
Secure Logging

Figure 1. Componets of Grid Security Model

GT3 (the Globus Toolkit) and its accompany Grid Security Infrastructure (GSI3) provide the first implementation of OGSA mechanisms. The Grid Security Infrastructure (GSI) is a set of tools, libraries and protocols used in Globus to allow users and applications to securely access resources, based on public key encryption, X.509 certificates, and SSL protocol. Extensions to these standards have been added for single sign-on and delegation. The GT’s implementation of the GSI adheres to the Generic Security Service API (GSS-API), promoted by the IETF.

4. GT3 SECURITY ARCHITECTURE

4.1 Overview

Figure 2[4] shows a user invoking a job on a resource GT3 installed. The steps in this process, and their security components, are:

(1) The user generates a job instantiation request with a description of the job to be started. The user then signs this request with their GSI proxy credentials and sends the signed request to the Master Managed Job Factory Service (MMJFS) on the resource.

(2) The MMJFS runs in a non-privileged account. It verifies the signature on the request and establishes the identity of the user who sent it. It then determines the local account in which the job should be run. Currently this is done by using the grid-map file and user's grid identity.

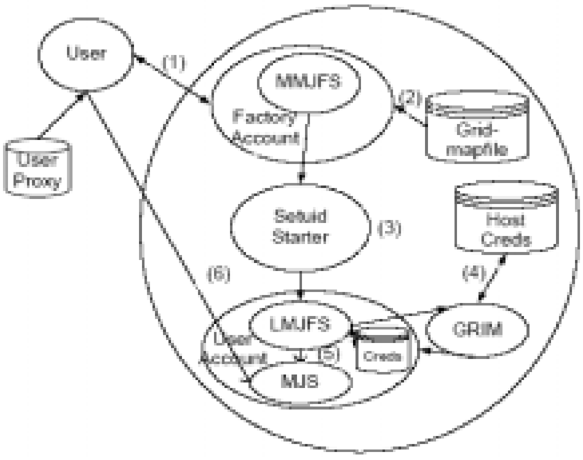


Figure 2. GT3 Job Invocation

(3) Assuming the user does not already have a Local Managed Job Factory Service (LMJFS) running in their account, the MMJFS invokes the setuid starter process to start one. The setuid starter is a small setuid program running with root privileges that has the sole function of starting LMJFS in user's account.

(4) Once the LMJFS starts up, it uses the Grid Resource Identity Mapper (GRIM) to acquire a set of credentials. GRIM is a setuid program that accesses the local host credentials and from them generates a proxy for the LMJFS. This proxy credential has embedded in it the user's Grid identity, local account name and local policy about the user. The latter policy is obtained from the Grid map file entries that apply to that local account.

(5) The MMJFS then forwards the original user-signed job instantiation request from the user to the LMJFS. The LMJFS verifies the signature on the request to make sure it has not been tampered with and to make sure it was created by a user that is authorized to run in the local user account. Once these checks are successfully completed, the LMJFS instantiates a Managed Job Service (MJS), presents it with the user’s request, and returns a reference to the MJS to the user.

(6) The user then connects to the MJS. The user and MJS then perform mutual authentication, the user using their 3proxy and the MJS using the credentials acquired from GRIM. The MJS authorizes the user as a valid user to access the local account it is running in. The user authorizes the MJS as having a credential issued from a appropriate host credential and containing a Grid identity matching it's own, thus verifying the MJS it is talking to is running not only on the right host, but in an appropriate account. The user would then delegate GSI credentials to the MJS for the job to use and start the job running.

4.2 Server Side Security

Figure 3 shows the JAX-RPC handlers that are involved in security related message processing on a server. A message arrives from the client the soap engine invoke handlers as follows [3]:

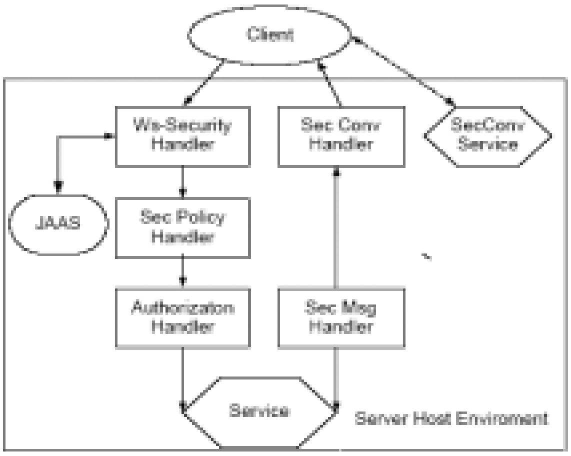


Figure 3. Server side Message Flow

(1) The first of these handlers, the WS-Security handler, searches the message for any WS-Security headers. From those headers it extracts any keying material, which can be either in the form of an X509 certificate and associated certificate chain or a reference to a previously established secure conversation session.

(2) The next handler, the security policy handler, checks that incoming message fulfill any security requirements the service may have .These requirements are specified, on a per-operation basis, as part of a security descriptor during deployment.

(3) The security policy handler is followed by authorization handler. This handler verified that the principle established by the WS-Security handler is authorized to invoke the service.

(4) The message is finally handed off to the actual service for processing after the message has passed the authorization handler.

(5) Replies from the service back to the client are processed by two outbound handlers: the GSI Secure Conversation message handler and the GSI Secure Message handler. The GSI Secure Conversation message handler deals with encrypting and signing messages using a previously established security context, whereas the GSI Secure Message handler deals with messages without an established contact by signing the messages using X509 certificates.

4.3 Client Side Security

In contrast to the server side, where security is specified via deployment descriptors, client side security configuration is handled by the application .The client side application can specify to use either the Secure session or GSI Secure Message security approaches .It does this by a per message property that is processed by the client side security handlers. The figure 4 shows the Client Side Message Flow. [3]:

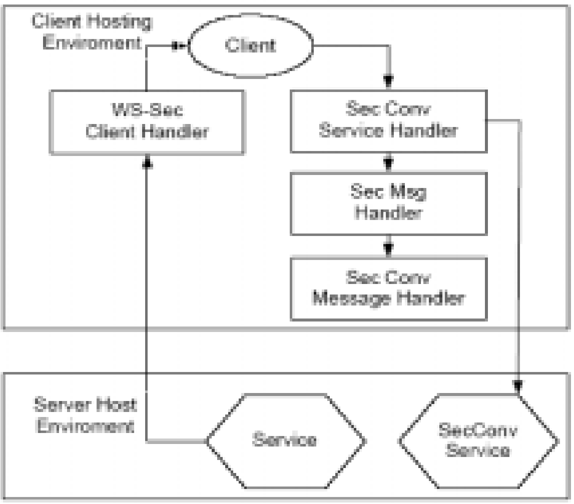


Figure 4. Client Side Message Flow

There are three outbound client side security handlers:

(1) The secure conversation service handler is only operational if GSI Secure Session mode is in use. It establishes a security session with a secure conversation service collocated with the service to which the client aims to communicate. It will also authorize the service by comparing the service’s principal/subject obtained during session establishment with a value provided by the client application. Once the session has been established the handler passes on the original message.

(2) The next handler in the chain, the secure message handler,

is only operational if GSI Secure Session mode is in use. It signs and/or encrypts messages using a security session established by the first handler.

(3) The third outbound handler is operational only if GSI Secure Message mode is in use. It handles signing of messages in GSI Secure Message operation.

The client side inbound handler (the WS-Security client handler) deals with verifying and decrypting any signed and/or encrypted incoming messages. In the case of GSI Secure Message operation it will also authorize the remote side in a similar fashion to the outbound secure conversation service handler.

5. CONCLUSION

Security is of fundamental importance to the grid service. There is a great deal of security specifications, such as WS-privacy, WS-Secure, WS-policy, WS-trust, Convention etc., produced by the Web Services community that are adopted to implement the security in a grid environment, whereas the security for web services is not the security for OGSA that also can help web services smooth difficulty in security. GT3, which is based in the emerging Wed Services and the OGSA, leverages Web Services for its security functionality, its development provides a basis for a variety of future work. The Globus Toolkit version 3 (GT3) represents the latest evolution of the Grid Security Infrastructure. Nowadays, there are still lots of Grid security Challenges in GSI, the combination of GSI and existing security mechanisms in Web service is an inevitable trend.

6. REFERENCES

[1] Foster, C. Kesselman, S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International J. Supercomputer Applications, 15(3), 2001.

[2] Mark Baker, Rajkumar Buyya and Domenico Laforenza The Grid: A Survey on Global Efforts in Grid Computing http://www.grid.com.cn/gct/GridSurvey.pdf

[3] Jarek Gawor, Sam Meder, Frank Siebenlist, Von Welch ,GT3 Grid Security Infrastructure Overview

[4] Nataraj Nagaratnam , Philippe Janson Security Architecture for Open Grid Services http://www.gridforum.org

[5] USC Information Sciences Institute, GT3 Developers and Administrators Tutorial , http://www.globus.org/

[6] Ian Foster, Carl Kesselman, Jeffrey M. Nick, Steven Tuecke, The Physiology of the Grid——An Open Grid Services Architecture for Distributed Systems Integration

[7] Von Welch 1 Frank Siebenlist 2 Ian Foster 1,2 John Bresnahan 2, Security for Grid Services

[8] Mary Thompson,Security Implications of Typical Grid Computing Usage Scenarios, University of Virginia

[9] Mary R. Thompson, Robert Cowles,CA-based Trust Issues for Grid Authentication and Identity Delegation, Grid Certificate Policy Working Group

[10] Frank Siebenlist2, Ian Foster12, GSI3: Security for Grid Services, University of Chicago, Department of Computer Science

[11] Von Welch, Frank Siebenlist, Ian Foster, Security for Grid Services, Argonne National Laboratory, Mathematics and Computer Science Division

[12] Jarek Gawor, Sam Meder, Frank Siebenlist, Von Welch ,GT3 Grid Security Infrastructure Overview

Fangmin Li, male, born in June, 1968. Doctor, full professor, his research interests are in QoS (Quality of Service), network security, and multimedia communication technology.

ZuGuang Fan, male, was born in Feb. 1979, is a postgraduate in the School of Information Engineering, Wuhan University of Technology. His research direction is Grid computing.

Security Prototype Framework Design for Open Grid Services Architecture (OGSA)*

Yang Qiulin, Zhou Zude, Li Fangmin
School of Informational Engineering, Wuhan University of Technology
Wuhan, Hubei, China
Email: sword@mail.whut.edu.cn zudezhou@mail.whut.edu.cn lifangmin@mail.whut.edu.cn
Tel: 02787299697 02787651445 02762356877

ABSTRACT

Building on concepts and technologies from the Grid and Web services communities, Open Grid Services Architecture defines an uniform exposed service semantics (the Grid service). This paper gives out a brief introduction to the definition, structure and characteristics of grid. It also gives an outline of the security requirements and security infrastructure of OGSA. Based on Globus Tookit 3.0 security infrastructure, we have designed in this paper a convenient and dynamic security prototype frameworks by the aid of pluggable module, builds the security group communication policy. It fully ensures security for OGSA, while it reduces the step of security negotiating, enhances the efficiency of interoperability between entities, improves the universal applicability, expansibility and dynamic of GSI. Using the security frameworks, a secure model for a shared virtual enterprise was constructed which provides a security platform for digital manufacture.

Keywords: Grid, OGSA, Web services, Security, Group Communication, OA.

1. INTRODUCTION

Grid computing is a way of organizing computing resources so that they can be flexibly and dynamically allocated and accessed, often to solve problems requiring many organizations' resources. The resources can include central processors, storage, network bandwidth, databases, applications, sensors and so on. The objective of grid computing is to make resources available (that is, processing capacity, information, input sensors, output and storage devices and networking facilities) so they can be more efficiently exploited [1].

Grid concepts and technologies were first developed to enable resource sharing within far-flung scientific collaborations[2][3]. Applications include collaborative visualization of large scientific datasets (pooling of expertise), distributed computing for computationally demanding data analyses (pooling of compute power and storage), and coupling of scientific instruments with remote computers and archives (increasing functionality as well as availability) [6].

We argue that Grid concepts are critically important for commercial computing, because of a solution to new challenges relating to the construction of reliable, scalable, and secure distributed systems. These challenges derive from the current rush, driven by technology trends and commercial pressures, to decompose and distribute through the network previously monolithic host-centric services.

2. OPEN GRID SERVICES ARCHITECTURE (OGSA)

Grid technologies can be aligned with Web services technologies [10] [11] to capitalize on desirable Web services properties. We call this alignment—and augmentation—of Grid and Web services technologies an Open Grid Services Architecture (OGSA) [2][4], with the term architecture denoting here a well-defined set of basic interfaces from which can be constructed interesting systems, and open being used to communicate extensibility, vendor neutrality, and commitment to a community standardization process. This architecture uses the Web Services Description Language (WSDL) to achieve self-describing, discoverable services and interoperable protocols, with extensions to support multiple coordinated interfaces and change management [4]. OGSA leverages experience gained with the Globus Toolkit to define conventions and WSDL interfaces for a Grid service, a (potentially transient) stateful service instance supporting reliable and secure invocation (when required), lifetime management, notification, policy management, credential management, and virtualization. OGSA also defines interfaces for the discovery of Grid service instances and for the creation of transient Grid service instances. The result is a standards-based distributed service system that supports the creation of the sophisticated distributed services required in modern enterprise and interorganizational computing environments. We view a Grid as an extensible set of Grid services that may be aggregated in various ways to meet the needs of VOs, which themselves can be defined in part by the services that they operate and share.

3. SECURITY INFRASTRUCTURE FOR OPEN GRID SERVICES ARCHITECTURE

The security challenges faced in a Grid environment can be grouped into three categories, A solution within a given category will often depend on a solution in another category:

Integration solutions where existing services needs to be used, and interfaces should be abstracted to provide an extensible architecture; Interoperability solutions so that services hosted in different virtual organizations that have different security mechanisms and policies will be able to invoke each other; Solutions to define, manage and enforce trust policies within a dynamic Grid environment.

To be successful, a Grid security architecture needs to step up to these challenges with existing security architectures and models across platforms and hosting environments. The Grid security model groups all security management functions applicable to various aspects of binding, policy and federation[5][9]. These include key management for

* Supported by: national key scientific and technological project (No:2003BA103C)

cryptographic functions, user registry management, authorization, privacy and trust policy management and management of mapping rules which enables federation. It may also include the management of intrusion detection, anti-virus services and assurance information enabling service requestors to discover what security mechanisms and assurances a hosting environment can offer. Addressing the management of various aspects of the security infrastructure will satisfy the manageability requirement on the Grid environment. The Grid security model is a framework that is extensible, flexible, and maximizes existing investments in security infrastructure. It allows use of existing technologies such as X.509 public-key certificates, Kerberos shared-secret tickets and even password digests. Therefore, it is important for the security architecture to adopt, embrace and support existing standards where relevant. Given Grid services [8] are based on web services, Grid security model will embrace and extend the Web services security standards that evolve in the industry. Specifically, given that OGSA is a service oriented architecture based on Web services (i.e. WSDL based service definitions), the OGSA security model needs to be consistent with web services security model. The web services security roadmap [WSR] provides a layered approach to address web services, and also defines SOAP security bindings [7]. Like any other service, security services should be exposed as web services (i.e., with a WSDL definition) and should expose functionality while hiding implementation details.

4. THE DESIGN FOR DYNAMIC PLUGGABLE SECURITY MODULE

The OGSA security Architecture allows use of existing and evolving technologies. Based on Globus Tookit 3.0 security infrastructure, this paper will design a convenient and dynamic security prototype frameworks by the aid of pluggable module, build the security group communication policy for deploying security architecture at liberty.

4.1 Build Security Group Communication, Maintain the Relation of Group Members.

This paper establishes a security communication-layer tree model, which consisted of Root Group Proxy by task leader and multilevel trust proxy and group members (As Figure 1 indicates).This model use protocol groups of security group communication including multilevel proxy authentication, secret key assign, security communication, secret management and so on.

Security communication-layer tree model consists of root group proxy, multilevel trust proxy and bottom group members. Root group proxy is the root nod of Grid, it produces the group secret key and child secret keys of all trust proxy from different level, it also will make a decision to let new member join a suited child group. “a” nod and it’s child nods together make up of child group, and they share child group secret key. The secret key (bi) of the member is assigned by its father nod (a). The member and its father nod share the secret key together. Thus every member possesses multilevel secret key including the individual secret key, the child group key and the group’s root key, meanwhile a certain secret key belongs to multi-members.

When a new member (u) takes part in this group, we must assign the new member’s individual secret key and update some secret keys including the group’s root key and the secret keys of child groups in which the new member will join for insuring that the new member can’t know those old messages appearing before it join. Using the encrypt technology of RSA and the method of multi-broadcasting, assigning each level secret key is realized through sending updated messages to each level members. In order to preventing deceivability and afresh sending, each level proxies use owns private key to sign and append postmark. When a member leave from a group, we must delete the member’s nod and it’s secret key, then update other’s secret keys for insuring that the member can’t know the group’s messages until it join again in the later.

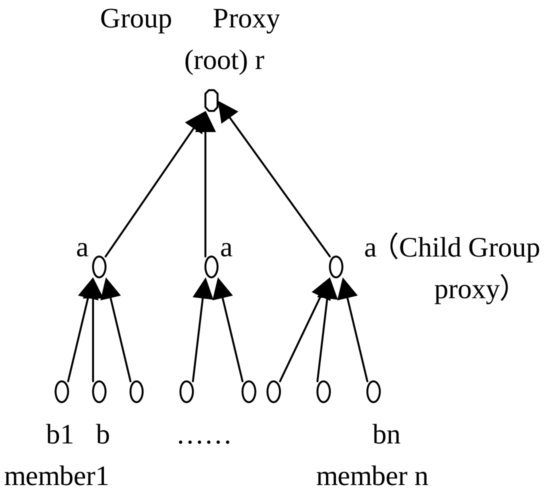


Figure 1 Security communication-layer tree model

After building communication group in security, it is very important to maintain the relation of group members through making use of unitive security polices and access control schemes in way of regarding group as units. The group sever maintain, index and authentication the relation between members. The group severs save authentic credentials from users and Grid entities. In these credentials, the group members’ information proves their private status including the unit to which they are belong, entity name and their credential and so on. When a user begin to request services, he use his UseProxy (UP) to index the group to which resource entity accessed belongs, then request the authorization credentials of needful Grid services from authorization sever based on the authorization credentials presented by own group. Therefore, maintaining the relation of group members and interoperating in way of building group will reduce complexity of interoperation.

4.2 Design and Build Dynamic Pluggable Security Service Component

In Open Grid Services Architecture and the software tool –Globus Tookit 3.0, the Grid security model allows use of existing technologies such as X.509 public-key certificates, Kerberos shared-secret tickets and even password digests. It is important for the security architecture to adopt, embrace and support existing standards and evolving standards. In the Grid security model, there are some characteristics including: entities which interoperate each other may locate different security group, the security strategies of each group may be different, a service request may span several different security groups, extensible security strategy should be provided when updating the security polices in group domain. All above lead to that it is very complex when entities interoperate in security.

In order to solve the universal applicability, expansibility, dynamic of GSI, we design a pluggable security service component including a strategy monitoring component, use proxy and a security strategy database. (As Figure 2 indicates)

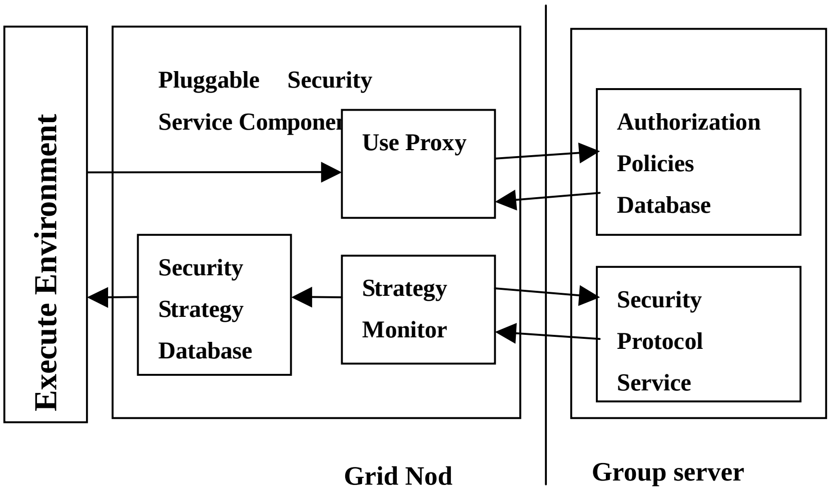


Figure 2 Dynamic Pluggable Security Service Component

Use Proxy delegates an entity's rights to access resource services and Single Logon is held up to realize interoperate, as be defined in the draft-GGF[a]. We access Grid services addressing the Dynamic Pluggable Security Service Component. Strategy Monitor model will monitor those security protocols used by remove communication group. An entity can login remove group serve after it's Use Proxy successfully complete the act of authentication. If the remove group server demand interoperability using own security protocols, entity should download these security policies from the serve in dynamic and build these policies on native nod combining native security strategy database, then achieve security policies up which both interoperating sides all hold, as a result, entity can complete it's task in security. Based on these security policies harmonized, application program can produce correlative code to achieve interoperate in security with remove services. The security strategy database embrace existing technologies such as X.509 public-key certificates, Kerberos shared-secret tickets, PGP public-key technologies and even password digests. Based on the demand coming from remove serve, entity may in dynamic upload the right security protocol for unifying the protocol used by both sides, then it can login remove sever. After completing resource map on the remove serve, entity can use the remove resource. The group server should unify all native members to adopt coincident security polices together, and decide the security police which external accessing entities will use. In the native group, all nods acquiescently use the native security polices, needn't monitor and download security strategies from native group serve. Therefore, the complexity of interoperate will be reduced. For example, the entity can access resource services without authorization in a trust domain, all entities in a group domain use a security protocol commonly.

The dynamic pluggable security service component is fit extensive, dynamic environment of Grid services. It not only improves the universal applicability, expansibility, dynamic of OGSA security model, but also reduces the complexity of interoperation.

5. TYPICAL APPLICATION

Using the Infrastructure for Open Grid Services Architecture（OGSA），a digital manufacture platform was constructed. There are virtual enterprise A and virtual enterprise B. If user group located on B want to complete their manufacture task, they must access A and use resource services of A. Each entity is connected with high speed Ethernet in physic layer, is linked through the protocols of Internet in link layer and assisted using Grid middleware and Grid software. We use above security model designed in the paper to construct security Grid platform for digital manufacture. The model can offer security services well for the Grid of digital manufacture.

6. CONCLUTION

Based on Globus Tookit 3.0, we have built a dynamic extensible security mechanism providing a group communication approach. We realized the security model as Grid services. It fully ensures security for OGSA, while it reduces the step of security negotiating, enhances the efficiency of interoperate between entities, improves the universal applicability, expansibility and dynamic of GSI. Now, enormous fund was devoted to the research of Grid and the research of Grid security model becomes more important. Grid is developing rapidly combining Web services technology. Therefore, the Grid security architecture will extend and leverage the security of Web services, which is also the aspect that we will research in the future.

7. REFERENCES

[1] Foster and C. Kesselman, editors, The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, 1998.

[2] Mark Baker, Rajkumar Buyya and Domenico Laforenza The Grid: A Survey on Global Efforts in Grid Computing http://www.grid.com.cn/gct/GridSurvey.pdf

[3] I. Foster. The Grid: A New Infrastructure for 21st

Century Science. Physics Today, 55(2):42-47, 2002.

[4] I. Foster, C. Kesselman, S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International J. Supercomputer Applications, 15(3), 2001.

[5] GWD (draft-ggf-sec-arch) GGF OGSA Security Workgroup http://www.ggf.org/ogsa-sec-wg

[6] I. Foster, C. Kesselman, J. Nick, S. Tuecke. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. January, 2002.

[7] Security in a Web Services World: A Proposed Architecture and Roadmap, http://www-106.ibm.com/developerworks/library/ws-secmap/

[8] S. Tuecke, K. Czajkowski, I. Foster, J. Frey, S. Graham, C. Kesselman. Grid Service Specification. Draft 2, 6/13/2002, http://www.globus.org

[9] Shanli Li, Zhihui Dou, Yu chen, Peng Liu. Grid compute. 2002.10 (in Chinese)

[10] Graham, S., Simeonov, S., Boubez, T., Daniels, G., Davis, D., Nakamura, Y. and Neyama, R. Building Web Services with Java: Making Sense of XML, SOAP, WSDL, and UDDI. Sams, 2001.

[11] Kreger, H. Web Services Conceptual Architecture. IBMTechnical Report WCSA 1.0, 2001.

Yang Qiulin, male, was born in July 1980, is a postgraduate in Information Engineering Institute, Wuhan University of Technology. His research interests are in grid computing, network security.

Zhou Zude, male, professor, tutor of doctor, is the President of Wuhan University of Technology, his research interests are CNC Theory and technology, Intelligent Control, Digital Manufacturing, Reliability and Fault Diagnosis of the Modern Manufacturing Systems and etc.

The Design and Implement of a Distance Education System Based on Improved MVC Pattern

Ran Chunyu, Hu Hengying, Chen Caixian
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430070, China
Email: hhy_carot@sohu.com **Tel:** 027-87297484

ABSTRACT

With the development and popularization of Internet, distance education has already entered the stage based on Web. Due to the shortcomings of Web application and development, MVC (Model-View-Controller) pattern is introduced. The separation of logic programming from interface design is effectively archived by the three kinds of technology correlated with Java: JSP, Servlet and JavaBeans. This paper analyses MVC pattern. On the purpose of the unanimous solution of distance education and commercialization, an improved MVC pattern based on the conventional one is introduced. It puts the emphases on using Java to realize each component in the improved MVC pattern. It offers a feasible solution for distance education based on Web.

Keywords: distance education, MVC, Controller, Model.

1. INTRODUCTION

The roots of pure distance education go back at least 170 years to the beginning of correspondence study. As radio developed during the First World War and television in the 1950s, distance education found new way to expand its influence. [1] With the development of computer, multimedia, communication, especially the swift development of Internet, the means of distance education has qualitative leaps and distance education become modern distance education under the new technology condition. It efficiently provides real-time or non-real-time interact between teachers and students in case of distance instead traditional education model restraint by time and space. Moreover, it also expands the scale of teaching, improves teaching quality and reduces the educational cost. Under distance education, students can arrange the time and place for study more freely, choose the content of study and study plan independently. They also can put forward the question among study and can be answered in time at any time. It benefits individual study even more. [2, 3]

The modern distance education based on Internet structures the opening environment system. But in the present application develop based on Web, there still exist many places that less than satisfactory, such as low reusable degree of procedure, tedious and difficult maintaining, weak adaptability to change of the procedure, sharing responsibility between Webpage designers and software developers, restricting of Webpage design and software development and resulting low developing efficiency. Therefore, a kind of rational, effective programming pattern——MVC (Model-View-Controller) pattern emerges as the time requires. Designing Web application by using MVC pattern can basically solve the problems mentioned above. In this paper, the first section provides how to divide the system into modules; the second section summarizes the MVC pattern and an improved pattern based on the conventional one; in the third section, designs of Model, Controller and View are given in details; the fourth section summary follows.

2. PARTITION OF FUNCTION MODULE

Distance education is an all-around teaching management application system. It includes many facets of education system. The facets are not isolated but related. From the administrators of the whole system perspective, there are three related roles: student, teacher and administrator. From the view of functions, there are three subsystems: a subsystem for students to learn, a subsystem for teachers to teach and a subsystem for system administrators to manage the whole system. Different user names and passwords are required when they enter their spaces.

Student Space: The subsystem for students to learn makes up the student space. The student space serve the user with student’s identity the functions below: course introduction, teacher introduction, course arrangement, teaching announcement, on-line study, on-line test, course homework, discussion, resource and score inquiry online.

Teacher Space: The subsystem for teachers forms the teacher space. When users with teacher identification log into this space, they can maintain teaching announcement, manage the course job, manage teaching resources, manage discuss, administrate the online exam pool, maintain the basic information and issue the students’ achievement.

Administrator workroom: At present, this module is in charge of verifying the identity of the registered user. Only the user who went through the checking can log in to and enter the relevant system.

3. THE SOLUTION OF MVC PATTERN

MVC is designed initially by Smalltalk. The pattern partitioned off function to three interrelated components —Model, View and Controller. As illustrated in Figure 1.

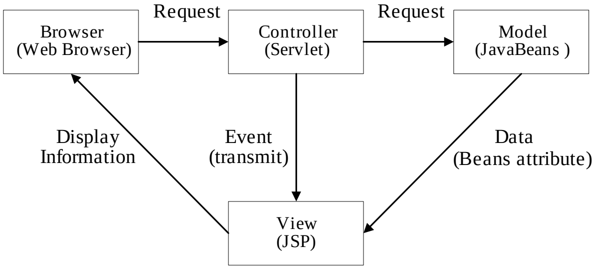


Figure 1 MVC architecture

Model realizes the business logic of the application. It's the core of application, encapsulates data and behavior in essence, including the rule to control and revise data, offers a set of methods to inquire and change Model state.

View realizes the appearance of module. It is the external manifestation of the application. Its main function is to show users, by a certain form, the results which business logic module deal.

Controller receives user's requests, and passes user's data to business logic module. And then it invokes corresponding business logic modules to deal with the requests. At last, result page according to the user's needs are created by corresponding view modules and returned to browser. It links Model and View together. Generally, Servlet is applied to play the role of Controller because the Controller controls the procedure flow and doesn't need output to the client. [4]

The advantages of this pattern are: 1) loosely-coupled: many kinds of components can interact in a flexible way; 2) parallel development: the duty is clear and it is possible to partition the whole system into components so that different person could develop at the same time. The structure is clear so as to easily integrate and maintain; 3) expansibility: Controller can expand with the modules; 4) reusable quality: it can improve the reusable quality by encapsulated the business logic in the component. [5]

In this system, the conventional MVC is improved. As illustrated in Figure 2. The design will contribute a lot to the general and uniform solutions to modern distance education system. And it may also help to lead the education supporting system to commerce.

The improvement abstract Controller output, regard a Servlet in common use as the controller of all requests. The controller draws out the parameters from the requests objects, and then makes an instantiation or obtains a certain corresponding Servlet in memory according to the parameters to respond the request. The certain corresponding Servlet is called Action in the structure. It is a Servlet essentially, but it inherits from abstract Action class. Abstract Action class will be introduced in 4.2.2.

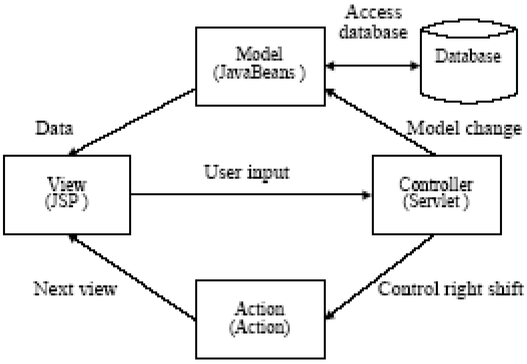


Figure 2 Architecture of improved MVC

4. THE IMPLEMENTATION OF IMPROVED MVC DESIGN PATTERN

4.1 The Design for the Model Classes

The model classes are applied to carry out Model components in MVC pattern, most are presented as JavaBean components. There are three kinds of classes: business object classes, application object classes and testing classes.

4.1.1 The Business Object Classes

The business object classes are comparatively easy to design. Every class supports a table in backstage database. Certainly, they can be associated with other kind of data source in database, for instance, view in database, the stored procedure of returned data set, etc. It worth nothing here that twelve business object classes exist in the system, associated with the tables in the database of the system:

- com.whut.masm.model.Users
  //user's basic data
- com.whut.masm.model.UserDate
  //users individualized information
- com.whut.masm.model.AdminUser
  //manage users and corresponding authority
- com.whut.masm.model.Borad
  //big board in answers and questions subsystem
- com.whut.masm.model.Borads
  //concrete zone in answers and questions //subsystem
- com.whut.masm.model.Forum
  //user speech in answers and questions subsystem
- com.whut.masm.model.Elite
  //quintessence management in answers and //questions subsystem
- com.whut.masm.model.Homework
  //homework data in homework management //subsystem
- com.whut.masm.model.News
  //teaching announcement data
- com.whut.masm.model.Exam
  //exam pool
- com.whut.masm.model.Resource
  //teaching resource data
- com.whut.masm.model.General
  //course data in common use

4.1.2 The Application Object Classes

The business object classes are used to delegate all entities the system known, however the application object classes act as the controllers of the business object classes. Three kinds of classes are designed in the system.

Model Classes: During the system operation, the business object resides in the application container class named Model. It is called container because "aggregation" in OOD (Object Oriented Design) is adopted. The business object is regarded as an attribute of the Model class. The table below describes the design of Model classes.

The WebModel class needs to implement three methods:

- void init(ServletContext context) uses to take out parameters from the web.xml describing descriptor, including JDBC drivers name, the URL of the database connection, and so on.
- The other two are to implement the HttpSessionBindingListener interface. One is void valueBound(HttpSessionEvent event), the other is void valueUnbound(HttpSessionEvent event). Realizing the interface is to notify the WebModel class if

HttpSessionBindingEvent event happened, such as a session is set invalid or out of time. We can release or shut down the database connection in the latter method when session time out or invalidate.

Table 1 Model classes

Attribute(private)	Method	Description
String jdbcDriver **String databaseURL** **Connection con**	public void connect()	connect or disconnect the database
	public void disconnect()	
	public boolean isConnected()	
	public String getJdbcDriver()	
	public void setJdbcDriver(String)	
	public String getDatabaseURL()	
	public void setDatabaseURL(String)	
String examID **List exams**	public void newExam(Exam exam)	add an examination question
	public void updateExam(Exam exam)	update an examination question
	public void deleteExam(Exam exam)	delete an examination question
	public void examSearch(String strSearch)	exams search for the specified examination question, and update the exams by the result set
	public Exam getExam()	get the Exam object associated with the examID
	public List getExams()	get the examination question list of the exams property
	public String getExamID()	get the ExamID property from the current Model
	public void setExamID(String)	set the ExamID's value in the current Model
……	……	define and implement properties and method in the Method for every business objects

Util Class: This is a class in common use. Its mainly function is conversion of the form of data.

4.1.3 Testing Class

This class is a tool class to test the Model class. Using it, you can simulate the controller updating the Model's states or making method invocation.

4.2 The Design for Controller Classes

The controller component is to operate the Model and change its states in line with user's input. In this system, it is designed to change the states of the aggregate object at present, and then chooses the next view to respond the custom according to the request.

The system uses ControllerServlet, a single Servlet, to act as the controller and shift the control right. How to change Model's states and which view to choose decided by special Action.

4.2.1 The Design for ControllerServlet

ControllerServlet is the driver of changing Model's states. It buffers the Action class that had been called in Session. When a request comes, ControllerServlet looks over Session whether the instances of the corresponding Action classes are in the Session. If there were not, ControllerServlet would take the action key from the request path information in order to load the special Action classes; If the instances had already loaded, it would use the instances directly.

4.2.2 The Design for Action Class

The Action class is the real component to handle business logic. Action object gets the necessary parameters initialized from the ControllerServlet, and also obtains the right to change the Model states and to control power. Although the designs of Action class and functions are required to be closely related, some function in common use can be drawn out.

The abstract class can be simply designed. It includes Servlet request and respond instance variables, ServletContext and Model. Except the getter() and setter() methods of those variables, there is an abstract method run(). Method run() is the only one must be implemented by each Action class.

The implement of concrete Action class depends on the function module, but they all inherit from the abstract Action class. In the method run() of their own, they invoke the methods in Model to finish the states changing, creating request sender and sending the request to the next view.

4.3 The Design of View

In view of the description above, the View implements the appearance of module, and works as external manifestation of the application. View ought to create the dynamic pages that should return to the users based on the results of business processing. JSP(Java Server Page) suits to program dynamic pages, so it's the best choice of developing the pages.

4.4 Flow Chart

We have boiled down how to implement every component in the adopted MVC pattern in detail. The flow chart below is to show how the Controller, Model and View work together in testing on line module, for example, in this distance education system responding and processing user's requests. As illustrated in Figure 3.

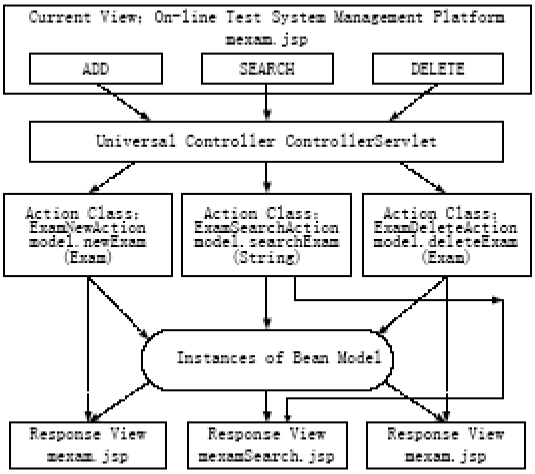


Figure 3 Flow chart

5. CONCLUSION

We introduce MVC pattern in distance education based on Web and take the full advantages of its loosely coupled, expansibility, and reusable quality. We also improved the traditional pattern, in order to make the design clearer, the integration easier and the maintenance more convenient. In sum, MVC pattern has great value and meaning in distance education system development.

6. REFERENCES

[1] McIsaac, M.S.,Gunawardena, C.N. ,"Distance education", http://seamonkey.ed.asu.edu/~mcisaac/dechapter/index.html

[2] Hu Yuerong, "The Development Nature of Distance Education and Appraisal", Science and Technology progress and policy, June 2002, pp.166-167.

[3] Shen Ruiming, Tang Zhiliang, Ding Dayu, Zheng Ke, "Distance Education System and its Application", Microcomputer Applications, No.7, 1999, pp.40-42

[4] Sun Ying, Xu Junhua, Zhang Yi, He Qingfeng, "Web Application of MVC Pattern and its Implementation with Java", Computer Engineering and Applications, No.17, 2001, pp.160-163.

[5] Zhang Yanqiu, Chen Chuan, He Mingde, "Designing JSP/Servlet+EJB Web Applications Based on MVC Design Mode", Computer Engineering, Vol.27, No.11, November 2001, pp.71-73

Ran Chunyu is a Full Professor and a head of computer system institute in computer science and technology department, Wuhan University of Technology. He graduated from Beijing architectural material college in 1969 with specialty of automatic control in engineering department. He once studied in Nanjing University and Huazhong University for computer courses, and then went to the Rumanian International Training Center for further study. He have undertaken some research items of "ninth five-year national key science and technology project". Now, he is carrying on some major researches of "key scientific research projects at provincial bureau level". More than 30 articles of him have been published in native or foreign journals and 5 teaching materials are compiled with his attendance. His research areas are computer applications like network database, communication security, graphics and images, and so on.

Hu Hengying is a graduate student of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. Her research areas are network database, information processing and so on.

Application of Data Mining Technology to Intrusion Detection System

Xia Hongxia, Shen Qi, Hao Rui
Computer Science, The Wuhan University of Technology
Wuhan, Hubei (Province) 430070, China
Email: shenqi1127@163.net Tel.: (027)87290674

ABSTRACT

Based the analysis of current Intrusion Detection technologies, the paper introduces the Data Mining Technology to the Intrusion Detection System, proposes a system architecture as well as a pattern strategy of automatic update. By adopting the Data Mining Technology, the frequency patterns could be dug out from a mass of network events. So, effective examination rules could be discovered, which would be then used to instruct the analysis of IDS network intrusion. Meanwhile, the usage of the pattern strategy of automatic update that adopts the ways of network real-time analysis has improved the efficiency and the veracity of the mining greatly. The integration of them would be effective in solving the problem of high misreport and false alerts rate in the traditional Intrusion Detection Systems.

Keywords: Intrusion Detection, Data Mining, Association Rules, and Frequent Episodes

1. INTRODUCTION

With the development of information industry, the computer has become an absolutely necessary tool in which amounts of important information has been stored. Meanwhile, more and more computers are getting rid of the isolation and connecting to the Internet, which greatly protrudes the importance of information security. In the near future, a considerable number of attacks in Internet would bring great intimidation to the security of computer network. So, it becomes an important subject of research that how to detect the intrusion effectively and then take some proper measures for protection in order to ensure the safe and steady operation of the network and the computers. The author does some researches on the problem of using Data Mining Technology in the IDS (Intrusion Detection System), and designs a system structure of the IDS on which a introduction of new strategy about automatic renew pattern is based.

2. THE CURRENT SITUATION OF IDS

Intrusion Detection, just as its name implies, is the detection, or in other words, identification of a behavior of intrusion. It collects information on some key points of the computer network or the computer system and then analyzes it so as to decide whether a behavior breaking the safety strategy or an evidence of being attacked lies there. The integration of both the hardware and the software used to detect the intrusions is IDS that is Intrusion Detection System.

The IDS could be divided into three kinds according to the difference in the derivation of data: (1) IDS based on the host; (2) IDS based on the network; (3) distributed IDS based on the integration of both.

The Intrusion Detection analysis technology could be separated into the Intrusion Detection based on the behavior and Intrusion Detection based on the knowledge. The detection based on the behavior, which is called Anomaly Detection too, is a judgment of whether there happens an intrusion according to the usage situation of resource or the user’s behavior; and the detection based on the knowledge, which also has the name of Misuse Detection, means detecting the intrusion by the appearances of some pre-formulated patterns of intrusion. It is based on some attack measures that we have already known. The intrusion is described mainly through the analysis of the characters, the conditions, the arrangement and the relationship between the events in the process. Most of the current Intrusion Detection Systems adopt the Misuse Detection method whose patterns are always pre-formulated by the security experts. At the same time, the patterns need continuous renewal in order to catch up with the development of the Intrusion technology, which in fact exists many limitations. The data on the network changes with the variety of the network application, while the effect of Misuse Detection method works only under the premise of definitely pre-defined patterns. It means that the Misuse Detection method cannot modify its patterns automatically according to the change of data in the network and is therefore incapable for the new technologies of attacks. Meanwhile, the renewal of the detection patterns depends on expert׳s eyes and hands only. It will inevitably lead to the failure in the timely renewal of patterns and increase the rate of misinformation. So there needs an automatic tool to detect the intrusion patterns or a tool to offer the assistance for the security experts. Applying the Data Mining technology to the IDS, we can half-automatically pick-up the characters of intrusions according to the historical flow data of network and then use them to classify the serial records on network and detect the intrusion.

3. CONSTRUCTING THE MODEL OF INTRUSION DETECTION BY THE DATA MINING TECHNOLOGY

The advantage of applying Data Mining technology to the Intrusion Detection System lies in its ability of mining the succinct and precise characters of intrusions in the system from large amounts of information automatically. It can solve the problem of difficulty in picking-up rules and in coding of the traditional Intrusion Detection system. At the same time, by concluding the forecast rules through the cataloging tree, the efficiency of search on instruction pattern would be improved a lot. Expecting to detect the abnormal patterns of network caused by the intrusions, this article shows the analysis in the amount of network data flow and a view of considering the data as the center while taking the intrusion detection as a mission of data analysis would be adopted here.

3.1 The System Structure of IDS

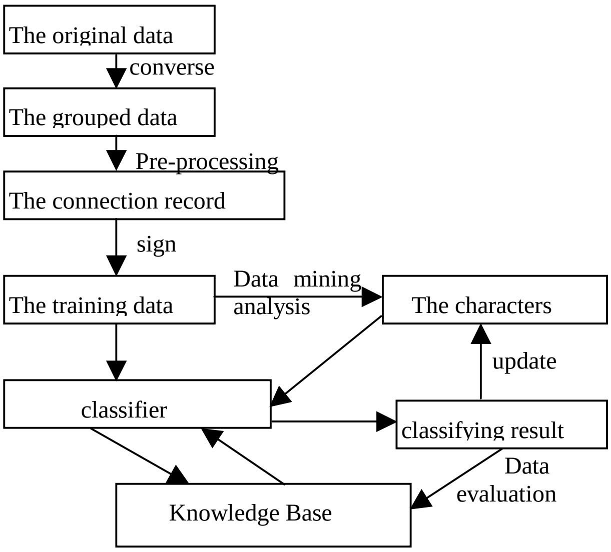


Figure 1 The System Structure of IDS

In Figure 1, the original data is a binary data obtained from the network. First, the binary data would be conversed into grouped data packages in formation of ASCII. Then, these packages would be disposed into a group of serial connection records composed of characters including the start time of connection, the last period of the connection, the original address, the object address and so on. Meanwhile, these records would be signed in order to separate them into the normal connection records and the intrusive connection records so that a training data collection with the sign would be formed. Then, with the help of association analysis algorithm and the episode analysis algorithm belong to the Data Mining algorithms, we could work out the frequent patterns such as the association rules and the frequent episode. Based on those findings, a collection of characters would be achieved through the selection of particularities that construct some accessional character of the connection records such as the time statistical character and the connection statistical character. Eventually, a model of Intrusion Detection System would be generated in the light of a classifier constructed by the classification algorithms. The whole course needs continuous repeated evaluation. If the effect is not good, we have to renew the collection of characters based on the result of classification and choose a better one. Simultaneously, we also need continuous repeated training and test of data so that the knowledge base would be renewed through the evaluation until then it maintains in a comparative steady situation [4]. Then, the integration of the collection of characters which we have obtained, the knowledge base, the classifier and the program of data gathering and data pre-processing could form a effective IDS based on the network.

3.2 The Working Process of IDS using the Data Mining Technology

By the method above, we could work out a framework of mining the audit data that is designed to obtain an intrusion detection model.

3.2.1 Data Pre-processing

The data dealt in the experiment are data packages captured from the network by the program of tcpdump. However, these captured data could not be analyzed directly by the classification algorithms. So, we need to make some pre-processing about it and pick up some meaningful characters from it.

Since the captured network data packages line in the sequence of their appearing time in the network and there might establish lots of connections in the same period in the data file of tcpdump, it would lead to packages of the same connection unregulated. So, in order to gather some relative information, we at first need to generalize all of the packages that are involved to one link into one connection record.

Then, we use a script to scan each row of the data in the files of ASCII formation in tcpdump and generalize all the data packages that belong to one link into a connection record. To each TCP connection, the job for the script program is following:

(1) In the step of establishment, examine whether 3-way handshake of TCP connection has been built up normally. If they are not, then check which situation they end in and what reason they failed. For instance, the connection is denied, the receiving of link-response data packages fails or receive some response packages of the connection that are not deserved, etc.

(2) In the step of transmission, inspect all of the data packages and the control packages, set certain records of counter and certain statistical values that are relative to the connection, such as the repeat rate, the repeated ACK, the number of bytes transmitted in two directions and so on.

(3) In the step of the end for the connection, check the situation of the end, such as the normal end (both of the communicators send and receive the package of fin), the interrupt, half-connection (only one host send the package of fin) and the break of connection, etc.

Because there is no course of establishment and end in data transmission based on the UDP protocol, each UDP data package could be considered as a connection, and so is each data package of ICMP [1].

The original data of tcpdump is conversed into a collection of data with the formation of each connection for each record. Then, we could go further in establishing the Intrusion Detection System based on the classification of network behaviors.

3.2.2 Mining Patterns from Audit Data

In the IDS, the basic idea of Mining patterns from Audit Data is mining the association rules and the frequent episodes from the audit data in order to obtain the patterns about the relationship between the records and between the attribute characters of the audit records. These patterns could be considered to a kind of statistical generalize of behavior of system recorded by the audit data since they described the relationship between the character of system and the events episode. Therefore, we could use these patterns to guide the gathering of audit data and the process of selecting character.

(1) Association Rules

The association rules mean the rules about certain relationship in a group of objects in the collection of data. Since there always be some relationship between the execution of the program and the users' behavior that always be expressed in the collection of connection data, the aim of mining association rules is to obtain the relativity of many characters from the table of database so as to find the relationship among the property of each record and establish the outline of normal or abnormal situation of use. In order to obtain association rules [2][3], the association mining algorithms making use of the minimum support and the minimum confidence would be helpful for the output of the "meaningful in the statistic" pattern. But, if there are just so few constraints, lots of unpractical rules would be generated, and what's worse, there even might generate a misguiding effect. Therefore, the interest based on the character is in need, or in other word, we have to constrain the generation of collection of frequent items with the help of key character. Moreover, according to the needs of different clients, the key character is also different. For example, if some analytical network task needs evaluating certain situation of the network service, then the service would be the key character; similarly, the dst-port would become the key character when some analytical task is interested in the port of object. The association rules is usually expressed like this : X→Y[c,s], X,Y are the records of the database, c is the confidence, and the s means the support.

(2) Frequent Episodes

In the Intrusion Detection, the data we analyze is always some kind of episode structures. For example, large amount of audit data gathered from the inspection of network could be seen as a group of events episodes based on the time. Moreover, an attack behavior is usually expressed as an episode character in certain time, such as the Syn Flood. One of the basic problems in analyzing this kind of event episodes is to find which ones are frequent episodes, or in other word, a serial events that happen frequently and have close quarters. The association analysis is mining the transverse relationship between different items of data record, while the episodes analysis is discovering the longitudinal relationship between the different data records and expressing the pattern of serial audit records through the frequent episodes. Here, we could work out the frequent episodes by a extend algorithm. First, we use the key character to find the collection of frequent association items, and then generate the pattern of frequent episodes according to the collection of association items. The frequent episodes is usually expressed as followings: X,Y→Z[c,s,w]. In that formulation, X, Y, Z means the records of the database, c means the confidence, s is the support and w is the time window.

3.2.3 The Selection of Character

According to the narration on above, proper selection of key character and recited character, which are very essential to the computer and the detection of the mode of intrusion, could differentiate the normal pattern and the attack pattern effectively. However, we could not always guess these selections because it would burden heavily to the users. So, an iterative process of pattern mining and compare, constructing character from modes, mode establishment and evaluation is needed [5]**.** Each of the iterative steps should choose a different combination of the key characters and the recite characters. The chosen items are confined in the basic characters, such as the service, the dst-host, the src-host, the src-port and so on, while the time stamp is neglected because of infrequence. Since the intrusions always aim at the victim host in the network, we could take the service as the key character and the dst-host as the recited character in the beginning. For each iterative process, the collection of the characters and the classification as the result are recorded in TP (true predication, namely the detection) and FP (fault predication, namely misinformation). In the last, the process chooses the collection of characters in the best-classified model.

In step of compare, we could use the program of frequent episodes to deal with the normal connection data and the intrusion data, then compare the pattern which we have gained so as to discover the "only intrusion pattern" before constructing the statistical character (average, count number, etc) and link them as the additional character to the connection records eventually.

3.2.4 Classification

The ultimate aim of intrusion detection is to group every audit record into a discrete collection of classification so as to judge whether it is an intrusion behavior or a normal one. Through Data Mining in the IDS, enough "normal" and "abnormal" audit data was collected. Meanwhile, each character has a sign of class. Then the classification algorithm could calculate a pattern, and make use of the discernable character value to describe every sign. In order to improve the effect of classification, there is a need for choosing the value of max information profit (the max reduce of the entropy) as the test value of current node so that the information needed for classifying the surplus collection of training data could be the minimum. For example, inspecting the record of TCP connection in the table 1(this record could be obtained through the tcpdump program). In this table, the "hot" is the number of visiting sensitive directories and documents in the system, such as the visit to "/var/log", the number of establishing and executing the sid program and so on; the "compromised" is the evidence for the attacks to the system, such as a large number of "file/path not found" errors and the emergence of lots NOP dictate in the data packages. We can obtain the rules in table 2 by the classified learning program, RIPPER.

Table 1 TCP connection record

label	**service**	**flag**	**hot**	**failed_logins**	**compromised**	**root_shell**	**su**
normal	ftp	SF	0	0	0	0	0
buffer_overflow	telnet	SF	3	0	2	1	0
normal	http	SF	0	0	0	0	0
normal	http	SF	0	0	0	0	0
guess_passwd	telnet	SF	0	6	0	0	0
buffer_overflow	telnet	SF	3	0	2	1	0
normal	telnet	SF	0	0	0	3	1
normal	ftp	SF	0	0	0	0	0
…	…	…	…	…	…	…	…

Table 2 the RIPPER rules obtained from the TCP connection record

RIPPER rules	**The instruction of the rules**
guess_passwd: -failed_logins>=6	If the number of user's failure in logging in is

	more than 6, this telnet connection is an attack of code guessing.
buffer_overflow: -hot=3, compromised=2, root_shell=1	If the behaviors of the hot are 3, while the number of the appearance of attacks to the system is 2 and the user obtains the root shell, then the telnet connection is an attack to buffer.
……	……
normal: -true	If none of above is satisfied, it is a normal session.

4. THE STRATEGY OF AUTOMATIC RENEWAL IN THE IDS

In the system of intrusion detection based on the Data Mining technology, it will cost a lot to gain the data for training. Moreover, the behavior of intrusions has been changing continuously. In order to cut down the cost for training data obviously and improve the cute degree for the intrusion that keep on changing all the time, an introduction of new pattern of automatic renewal is needed.

For an object deserved to be protected, if the pattern found in the audit data is correspond to the behavior of object, not only the IDS should make some response to it, but also the data would be gathered to the database simultaneously. After gathering some audit data about the object, we could keep on mining some relative patterns by using a new audit data and adding new founded rules to those we have already known until the whole collection of rules tends to be steady.

5. CONCLUSION

Aiming at the shortcomings of the current ways of intrusion detection, the article offers a system framework of constructing the IDS through the Data Mining technology based on the integration of the Anomaly Detection and Misuse Detection, shows the feasibility and the effect of the usage of Data Mining technology in the intrusion detection. However, for the purpose of improving the precision, feasibility and the expansibility of the IDS, lots of problems are still left to be resolved, such as the researches on the widely attack behaviors, picking-up minor of character from the original data, obtaining relative condition attributes and decision attributes, develop better picking-up pattern algorithm, learning algorithm and compare algorithm, etc.

6. REFERENCES

[1] W. Lee and S. J. Stolfo. Data Mining Approaches for Intrusion Detection [A]. In: Proceedings of the 7th USENIX Security Symposium[C], San Antonio, TX, 1998.

[2] Zhong Luo, Xia Hongxia, Yuan jingling, Study and Improvement on Hierarchical Algorithm of Association Rule, Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV, Proceedings of SPIE Vol.4730 (2002)

[3] Wu Ji, Huang Chuanhe, Wang Lina, Wu Xiaobing. Intrusion Detection System Based on Data Mining. Computer Engineering and Applications Vol.40, No.4, 2003.

[4] Xue Jingfeng, Cao Yuanda. Intrusion Detection Based on Data Mining. Computer Engineering Vol.29, No.9, 2003.

[5] Wen Zhiyu, Tang Hong, Wu Yu. Application of Data Mining Technology to Intrusion Detection System. Computer Engineering and Applications Vol.40, No.17, 2003.

Xia Hongxia is an associate professor. Her research interests are in database application technology, distance education and intelligent method.

Shen Qi is a graduate student of Computer Science and Technology Department, Wuhan University of Technology. She graduated from Wuhan University of Technology in 2002. Her research interests are in intelligent method and network security.

Design and Implementation of CSP Module

Rao Wenbi[1], Xiong Huiyue[2], Tang Chunming [3], Li Wei [4]
[12]School of Computer Science and Technology, Wuhan University of Technology
Wuhan,Hubei,430070,China
Email: raowenbi@yahoo.com.cn **Tel:** +86(0) 2752236288
[3]Information and Communication Engineering College, Harbin Engineering University
[4]Department of Computer and Systems Sciences, Royal Institute of Technology
Stockholm,16440, Sweden
Email: liwei@dsv.su.se **Tel:** +46(0)8161662

ABSTRACT

Along with fast development of Internet, the communicating degree of information is more improved. Cryptographic technology is used in the course of information communicate, since the problem of information security is outstanding increasingly. CSP, which is used to implement cryptographic operations, is the base of Microsoft's security applications frame and serves, but it can only be used in Windows Operating System and the speed of cryptographic operations is limited due to its software implementation. In order to resolve the above problems and really meet Chinese cryptographic standards, a new CSP module named as XyCSP which make the CSPs of Microsoft as reference, is designed in this paper. The design and implementation of XyCSP with three forms and the procedure of writing XyCSP module are all presented. Therefore the application developer can encrypt and decrypt the message with CryptoAPI and keep the message securely transmission.

Keywords: XyCSP, design, CryptoAPI, Cryptographic technology

1. INTRODUCTION

To protect the user's sensitive private data, CryptoAPI can provide some services that enable application developers to add encryption / decryption of data to their Microsoft Window-based application. Actually all the cryptographic operations are performed by independent modules known as Cryptographic Service Providers (CSPs), which contain implementations of cryptographic standards and algorithms. Currently we can use several software CSPs by Microsoft, such as: Microsoft Strong Cryptographic Provider, Microsoft AES Cryptographic Provider, Microsoft DSS Cryptographic Provider and so on[1]. Although these CSPs can implement many different cryptographic algorithms and even when implementing the same algorithm there are different key sizes and padding, they really have some limitations: the software CSPs have the less tamper-resistant and would be inconvenient in the fields of the interactive logon, e-mail signing, e-mail decryption and remote access authentication; additionally it would be difficult to apply to other platforms for the available CSPs shipped with the Window Operating Systems.

To resolve the above problems and really meet our needs, here we have the CSPs of Microsoft as reference, and design our own CSP module named XyCSP which can make up some shortages of software CSP. XyCSP utilizes the storage hardware: EKEY[2] and IC card, and make remote access authentication, interactive logon, e-mail signing and e-mail decryption more possible. With some adjustments, XyCSP can be run in the Linux System and provides the cryptographic services. Furthermore, an extra algorithm named as SSF33 is added in the XyCSP module according to Chinese cryptographic standards.

In this paper, we will present the architecture of XyCSP, and describe the implementing process of XyCSP module.

2. DESIGN OF XYCSP

Applications do not communicate directly with a CSP. Instead, applications call CryptoAPI functions exposed by the Operating System's Advapi32.dll and Crypt32.dll, then the Operating System filters these functions calls and passes them through CryptoSPI on to the appropriate CSP functions. Figure 1 shows the relationship clearly.

In the following, we will focus on the design of XyCSP in service provider layer and implement the XyCSP module.

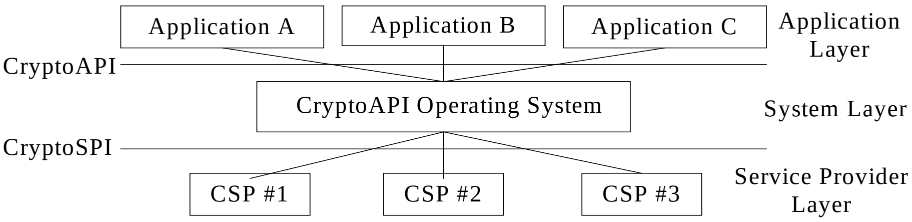


Fig 1 the Architectural of CSP

2.1 The Architecture Of Xycsp Module

The Cryptographic Services Manager doesn't assume any particular form factor for a CSP. Indeed, CSPs can be instantiated in hardware, software or both. There are two obvious distinctions between hardware and software implementations of CSPs, one is the degree of trust that the application receives by using a given CSP, another is the cost of developing that CSP. A hardware implementation should be more tamper-resistant than a software implementation. Hence a higher level degree of trust is achieved by the application. Software CSPs are the default in the Microsoft Operating System and are portable in that they can be carried as an executable file. Additionally, the modules that implement a CSP must be digitally signed (to authenticate their origin and integrity), and they should be made as tamper-resistant as possible.

In XyCSP module, the requirement extends to both software and hardware implementations. In order to have the good performance and modifiability, XyCSP module is divided into several parts. The structure would look like as figure 2. In the following, the module of hardware key database, PIN-based data protection and the CSP class are described.

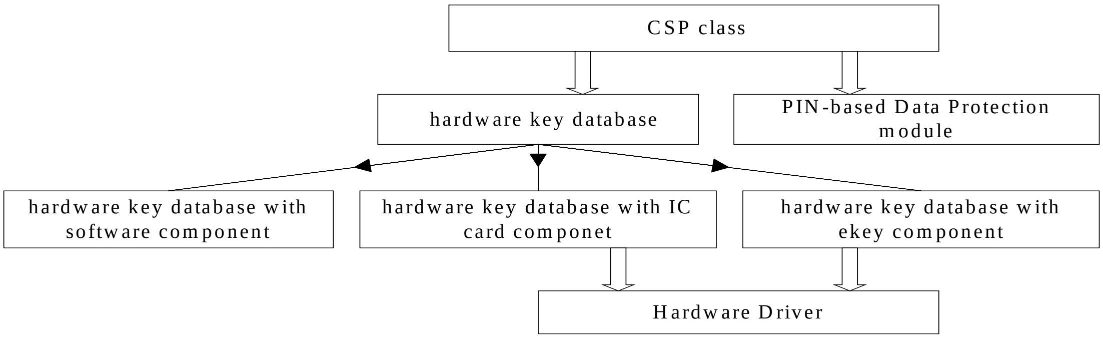

CSP class
hardware key database
PIN-based Data Protection module
hardware key database with software component
hardware key database with IC card componet
hardware key database with ekey component
Hardware Driver

Fig 2 the structure of my CSP

2.2 The Module Of Hardware Key Database

The most important component of XyCSP is hardware key database, which have different hardware forms as the Figure 2 shown: software component, IC card component and EKEY component. In general, the software component in this layer has no ability to keep and manage the key well. But EKEY and IC card belong to the storage hardware, they can be used to keep the cert and private key or personal information, and be able to authenticate a user for physical or electronic access. But there is a difference between the EKEY and IC card. EKEY belongs to smart card with an embedded encryption algorithm microchip and memory. The key and operate encryption can be generated inside the EKEY, but the export key can not be exported. EKEY has a higher level degree of security. IC card belongs to memory card, which works with the card reader and only builds in memory to store the key and personal information. It has no ability to generate the key or encrypt the data. Applications must generate the key in the memory and then import the key to IC card. When the data needs to be encrypted, the key must be exported outside the IC card and operate the encryption in the application. But the IC card is cheaper than EKEY.

Although the set of hardware key database has three forms, it provides the layer of CSP class developers with a single interface to support, which includes function CreateContainer, GenRandom, FindContainerAndKeypair, GenSymmeticKey, GenAsymmeticKey, ImportSymmeticKey, ExportSymmeticKey and so on.

2.3 The Module Of PIN-Based Data Protection

An additional module for the hardware key database module is PIN-based data protection module, which is mainly used in the IC card and EKEY components. To authenticate the user, this module would pop a dialog box and ask user to input the personal identification number, and verify the input PIN. It could make the remote access and interactive logon to have double secure protections.

2.4 The Layer of CSP Class

The top layer in figure 2 is the set of CSP base classes, which mainly defines the data objects[3].

(1) Class CCSP

```
class CCSP {
public:
  CONTAINERLIST m_containerlist[CSP_MAX];
public:
      ……  // other operations
}
```

Class CCSP is used to manage CContainer objects. The m_containerlist[] element shows how many container objects have been created in my CSP and every pointer address of container objects.
jk
(2) Class Ccontainer

```
class CContainer {
public:
   HARD_API m_hardAPI;    //hardware interface LPTSTR
   m_ContainerName;
private:
   HMODULE m_hHardDLL;
   /* the handle of hardware*/
    KEYLIST       m_keylist[CSP_MAX]; HASHLIST
    m_hashlist[CSP_MAX];
    ……  // other elements
    public:
    ……  // other operations
}
```

Class CContainer manages CKey objects and CHash objects. The m_keylist[] element and the m_hashlist[] element are used to store every pointer address of the key object, the hash object and the handle the objects refer to.

Furthermore, we have defined two base classes: class CKey and class CHash. These classes are mainly used to generate and manage the key objects and the hash objects. According to the supported algorithms, we can derive other classes to generate and manage the actual objects. So here we inherit the class CRSA, class CRC4, class C3DES, C33key from class CKey, and inherit class CMD5、 class CSHA1 from class CHash.

3. IMPLEMENTATION OF XYCSP

The creation procedure of XyCSP is as following: firstly creating a CSP DLL just like any other DLL, and exposing all of the CryptoSPI functions, then implementing these functions; Secondly writing the CSP setup program; Thirdly running test code that calls CryptoAPI functions to test CSP’s implementations of those functions; Finally signing the CSP by Microsoft and testing the CryptoSPI functions with CryptoAPI function.

3.1 The cryptographic algorithms

Before writing a CSP, a writer must select the cryptographic algorithms and obtain implementations for them. In XyCSP module, we support the symmetric key algorithms: the Single DES and Triple DES, RC4, SSF33; and also support the asymmetric key algorithm: RSA, the hash algorithm: MD5 and SHA-1.

In XyCSP module, Triple DES algorithm takes three different 64-bit keys with 192 bits in total, and implements three DES operations in the sequence encrypt-decrypt-encrypt shown in figure3.

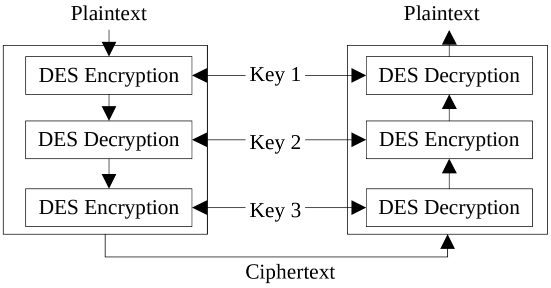


Fig3 the Procedure of Triple-DES Encryption and Decryption

RC4 key is used with 128-bit key. SSF33 is applied to EKEY component and IC card component. It provides the cardholder with sufficient security. In the RSA algorithm, we implement encryption and digital signature. The process is shown as figure 4 [4].

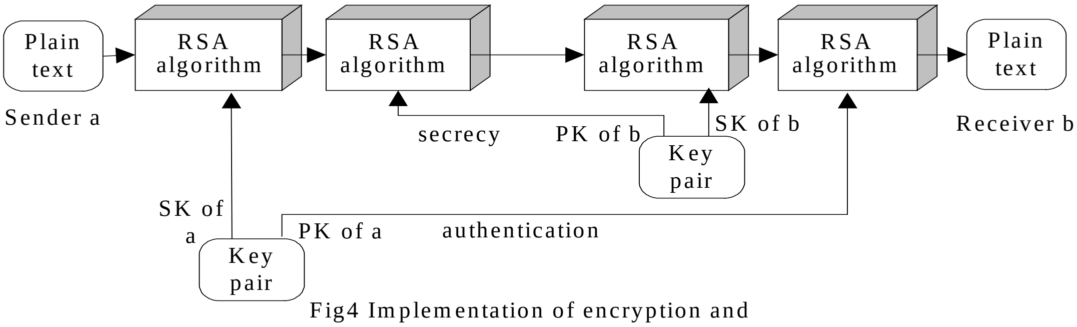


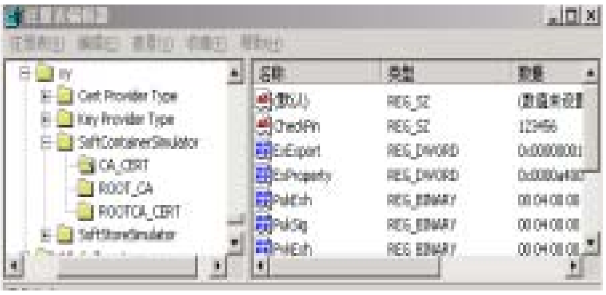
Fig4 Implementation of encryption and signature with RSA

The MD5 algorithm produces a 128-bits message digested as output. The SHA-1 takes a message of less than 264 bits in length and produces a message digest with 160-bit in length. After all these have been accomplished, a custom CSP can be created.

3.2 Implementing the CryptoSPI function

We would create a CSP DLL just like any other DLL, and expose all of the CryptoSPI functions, then implement these functions. These functions include CSP connection functions, the key generation and exchange functions, the data Encryption and Decryption functions, the hash and signature functions.

The key problem of implement these functions is how to store the key in the different hardware forms. In the software component, we would create CContainer object and CKey objects in the memory and register them in the regedit. From the following figure, it shows that three container objects are kept under softContainerSimulator directory. Every container keeps the values of the asymmetric key pairs and the attributes of the pairs, of course, the key values have been dealt [5].



- xy
  - Cert Provider Type
  - Key Provider Type
  - SoftContainerSimulator
    - CA_CERT
    - ROOT_CA
    - ROOTCA_CERT
  - SoftStoreSimulator

名称	类型	数据
(默认)	REG_SZ	(数值未设置)
CheckPin	REG_SZ	123456
ExExport	REG_DWORD	0x00000001
ExProperty	REG_DWORD	0x0000a40
PubExch	REG_BINARY	00 04 00 00
PubSig	REG_BINARY	00 04 00 00
PriExch	REG_BINARY	00 04 00 00

For the EKEY and IC card components, we only have to reset the content of the hardware and create the root file folder as the container. Under the folder we create the files for public key, private key and other symmetric keys.

3.3 Testing The XYCSP Module

In general, we would run test code that calls CryptoSPI functions directly to check whether the CSP functions work well and operate encryption\decryption\signature correctly. A more effective way that has interoperability with the different CSPs is used in this paper. For example, in order to test the encrypts correctness of RSA algorithm, Microsoft Base Cryptographic Provider v1.0, named as MSCSP, would be used to generate the RSA key pair and export the public key. Then this public key would be imported into XyCSP and encrypt the inputting message. Finally the MSCSP is used to decrypt the cipher and ensure whether the result is same with the inputting message [6]. The implementation code is shown as following:

```
char msprovider[]="Microsoft Base Cryptographic Provider
                        v1.0";
char text[]="this a text to be encrypted!";
CryptAcquireContext(&hProvMS,"interoperability",msprovide
                  r, 0x01, CRYPT_NEWKEYSET);
CPAcquireContext(&hProvWH, NULL,
                  CRYPT_VERIFYCONTEXT|CSP_SOFTCARD,
                  0);
CryptGenKey(hProvMS, AT_KEYEXCHANGE, 1024<<16 |
                  CRYPT_EXPORTABLE, &hMSPvkkey);
CryptExportKey(hMSPvkkey,0,PUBLICKEYBLOB, 0, buffer,
                  &nbufferlen);
CPImportKey(hProvWH, buffer, nbufferlen, 0, 0,
                  &hWHPubkey);
memset(buffer, 0x00, 2000);
nbufferlen=2000;
memcpy(buffer, (BYTE*)text, ntextlen);
CPEncrypt(hProvWH, hWHPubkey, 0, true, 0, buffer,
                  &ntextlen, nbufferlen );
CryptDecrypt(hMSPvkkey, 0, true, 0, buffer, &ntextlen);
CPReleaseContext(hProvWH,0);
CryptReleaseContext(hProvMS,0);
```

By analogy, we can use XyCSP to generate RSA key pair or other symmetric keys and import the public key (the private key or the symmetric key) to MSCSP, operate the encryption to verify the key generation and encryption correctly; for SSF33 key, MSCSP can’t support it, we also can interoperability with software XYCSP to check its validity.

Finally XyCSP is signed by Microsoft and the CryptoSPI functions are tested by CryptoAPI functions.

4. CONCLUSION

CSP serves as the base architecture of security application. To satisfy clients’ requirement, providers have to program CSP according the cryptographic standards and algorithms, so the application programmer could load CSP module to encrypt and decrypt the message.

5. REFERENCES

[1] Chen Yan-xue. The Theory and the Utility Of the Information Security. CHINA Railway Publishing House, Apr. 2001

[2] Xuhua Ding, Gene Tsudik. Simple Identity-Based Cryptography with Mediated RSA. Lecture Notes in Computer Science. Vol. 2612, 2003.

[3] Duan Gang. Encryption and Decryption . Publishing House of Electronics Industry, Jun. 2003

[4] Karl Brincat. On the Use of RSA as a Secret Key Cryptosystem. Designs, Codes and Cryptography 2001,22(3): 317-329.

[5] RSA Laboratories: PKCS #1 RSA Cryptography Specification Version2.0, 1998

[6] A. Palacious and H. Juarez. Cryptography with cycling chaos. Physics Letter. A 2002,303(5-6) 345-351.

Rao Wenbi is a Full Associate professor in School of Computer Science and Technology, Wuhan University of Technology. She received a Ph.D. in 2000 at the same University. She is now a visiting researcher in Department of Computer and Systems Sciences, Stockholm University and Swedish Royal Institue of Technology (2003-2004).

A Robust Approach to Authentication of Binary Image for Multimedia Communication

Jin WU[1, 2], Bei-bei XIA[1], Jian LIU[2], Jin-wen TIAN[2]
[1] College of Information Science and Engineering, Wuhan University of Science and Technology
Wuhan, Hubei, 430081, China
[2] Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology
Wuhan, Hubei, 430074, China
Email: hust_wu@163.com, lliw@163.com **Tel:** +86 (0)27 62000502

ABSTRACT

In recent years, digital watermarking techniques have been proposed to protect the copyright of multimedia data. Different watermarking schemes have been suggested for images, while there is little discussion about binary image. Basing on the analysis of the principle of digital watermarking and a spatial domain algorithm, we propose a watermarking algorithm---the discrete cosine transforms (DCT) to realize the authentication of binary image. The experiment results show that the DCT algorithm is feasible to binary image. Compared with the spatial domain algorithm, the watermarked image by the DCT almost has no distortions and is robust to common signal distortions, including geometric manipulations.

Keywords: Digital watermarking, Binary image, DCT, Authentication, Robust, Multimedia Communication.

Ⅰ INTRODUCTION

Digital watermarking is an important technology that has many applications. Many watermarking schemes have been suggested for images, audio, and video streams; however, the authentication problem is difficult for binary images because of their simple binary nature. Embedding of authentication signals into binary images will cause destruction of image contents, arouse possible suspect from invaders. Therefore, a good solution should take into consideration not only the security issue of reducing the possibility of being tampered with imperceptions but also the effectiveness of reducing image distortion resulting from authentication signal embedding. In this paper, an authentication method for binary images based on discrete cosine transform (DCT) is proposed, it is robust to different types of attacks comparing to a spatial domain authentication algorithm.

Ⅱ PRINCIPLES OF DIGITAL WATERMARK

There are three functional components required to embed a watermark in an image, including a watermark generator, a watermark carrier, and a carrier modifier. A watermark carrier is a list of data elements from the original image used for encoding the watermark. The watermark is a sequence of noise-like signals, based on a secret decryption key and generated pseudorandomly[1]. The carrier modifier adds the generated noise signals to the selected carrier. Embedding the watermark and detecting the watermark are the operations in the watermarking of digital media, which enable the owner to be identified. The watermarking scheme can be represented symbolically by

$$I_w = E(I_0, W) \tag{1}$$

Where I_0, W and I_w denote the original multimedia signal (audio, image or video), the watermark containing the information that the owner wishes to embed and the watermarked signal respectively. The embedding function E modifies I_0 according to W [1][2]. Fig.1 shows a general scheme of embedding digital watermarking.

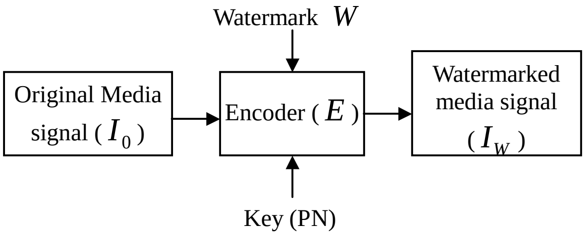

Watermark W
Original Media signal (I_0)
Encoder (E)
Watermarked media signal (I_W)
Key (PN)

Fig. 1 Embedding system of digital watermarking.

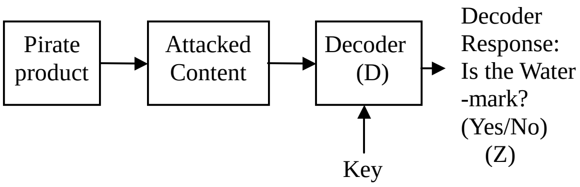


Fig. 2 Detecting system of digital watermarking.

For watermark detection, a detecting function D is used. This operation is represented by

$$W' = D(R, I_0) \tag{2}$$

where R is the signal to be tested, whether it is watermarked or not, and R could be a distorted version of I_w. The extracted watermark sequence W^t is compared with W, and a Yes/No decision is made. The decision is based on a correlation measure Z [3], as follows:

$$Z(W', W) = \begin{cases} 1 & c \geq \gamma_0 \\ 0 & otherwise \end{cases} \tag{3}$$

where c is the value of the correlation between W and W', γ_0 is a positive threshold. Fig. 2 shows a general scheme of detecting digital watermarking.

Ⅲ A SPATIAL DOMAIN ALGORITHM FOR BINARY IMAGE

A. Process of Authentication Code Embedding

The input to the proposed authentication code embedding process includes a cover image I with L blocks, two keys K_1 and K_2, and two random number generators f_1 and f_2. The output is a stego-image I' in which authentication codes are embedded. The steps of the embedding algorithm are as follows [4].

1) Use f_1 with K_1 as the seed to generate a sequence of L random numbers $c_1, c_2 ......c_l$, each with m bits, as the authentication codes.

2) Embed each c_i into a corresponding block B_i in I in the following way to yield $I^{'}$:

2.1) Use f_2 and K_2 as the seed to select randomly in B_i a certain number, say n, of code holders, each including a certain number, say m, of ordered pixels.

2.2) Compare c_i with each code holder g_k by matching respectively the m corresponding bits in c_i and g_k, and find the optimal code holder g_{opt} with the minimum number of different bit values.

2.3) Replace, if necessary, the value of each bit in g_{opt} with the corresponding one in c_i to complete the code embedding work.

B. Process of Authentication Code Verification

On the other hand, the proposed authentication code verification process is as follows. The input includes a stego-image I' with L blocks, as well as the two keys K_1 and K_2 and the two random number generators f_1 and f_2 used in the authentication code embedding process, the output is an authentication report for I'.

1) Re-generate the L m -bit authentication codes $c_1, c_2 ......c_l$, using f_1 and K_1.

2) Verify each c_i in a corresponding block B_i in I' in the following way:

2.1) Use f_2 and K_2 to reselect in B_i the n m -pixel code holders.

2.2) Compare c_i with each code holder g_k by matching the m corresponding bits in c_i and g_k, and label B_i as tampered if there exists no code holder with content identical to c_i ; or as authentic, otherwise.

3) If there exists any block being labeled tampered, report negative fidelity and output all tampered blocks; otherwise, regard the entire image of I' untampered.

Ⅳ PROPOSED DCT ALGORITHM FOR BINARY IMAGE

Rather than embedding the watermark in the spatial domain as the former algorithm, our proposed scheme is based on embedding a pseudorandom sequence of real numbers in the DCT coefficients of each segment of the host image.
The proposed algorithm can be carried out though three steps as following.

A. Image Split and Segment Split.

First the host image $f(x, y)$ is divided into 8*8 image segments which are not covered each other, noted that $B_k, k = 0,1....,k-1$. Namely

$$f(x, y) = \mathrm{Y}_{k=0}^{k-1} B_k = \mathrm{Y}_{k=0}^{k-1} f_k(x', y'), 0 \le x', y' < 8 \qquad (4)$$

The greater the image texture, the higher the watermark perceptibility region value. That is to say it can be embedded much greater watermark signal. According to the complexity of the image partial texture, it is an effective method to improve solidity of the watermark by increasing the intensity of the embedded watermark as could as possible. Therefore, the image segments are classified into two species: first is those whose texture are weak texture, noted S_1; second is the ones whose texture are strong, noted S_2. Because the edge points represent the break points of the image pixels, so the more edge points in the image segment, the greater the image texture will be. Therefore, edge points density can be used to classify the segments.

Segment classification is demonstrated as the following:

If

$$number\{e(x, y) \ne 0, (x, y) \in B_k\} \le T_1 \qquad (5)$$

then

$$B_k \in S_1 \qquad (6)$$

otherwise

$$B_k \in S_2 \qquad (7)$$

In which, $e(x, y)$ represents the binaration edge image extracting form the host image $f(x, y)$, T_1 is the region value set in advance.

B. DCT and Watermark Embedding.

After classification, every segment should be carried through DCT transform:

$$F_k(u', v') = DCT\{f_{k(}x', y'), 0 \le x', y' < 8\}, 0 \le u', v' < 8 \qquad (8)$$

The watermark w is made up of random sequence which is followed Caussian distribution $N(0,1)$, the length is K, namely $W = \{x_i, 0 \le i < K-1\}$. Watermark coding is organized by changing DC coefficient.

$$F_k'(u',v') = \begin{cases} F_k(u',v')(1+\alpha x_i) & if \quad u'=v'=0 \\ F_k(u',v') & otherwise \end{cases} \tag{9}$$

In which α is draw gene. The above formula shows that the iterative watermark single intensity is proportional to the average brightness of the image segment. According to Weber law, αx_i should less than 0.02 in theory. According to the organization of the common image which has different texture feature, in this scheme

$$\alpha = \begin{cases} 0.015, & B_k \in S_2 \\ 0.006, & B_k \in S_1 \end{cases} \tag{10}$$

C. Inverse DCT

After DCT inverse transform to the image which has adjusted in DCT domain, the embedded watermark image is got as:

$$f'(x,y) = \Upsilon_{k=0}^{k-1} IDCT\{F_k'(u',v')\} \tag{11}$$

After that, watermark detection can be carried on. The hypothesis inspection is used in the watermark detection method as:

$$H_0 : E = F^* - F = N \quad (nowatermark)$$
$$H_1 : E = F^* - F = W^* + N(having \quad watermark)$$

In which, F^* and F represent embedded image and host image respectively, W^* is the detecting watermark sequence, N is noise. The detecting watermark could be extracted through this formula [5]:

$$W_k^* = F_k^*(0,0) - F_k(0,0); \tag{12}$$

$$W^* = \{x_i^*, 0 \le i < K-1\} = \Upsilon_{k=0}^{k-1} W_k^* \tag{13}$$

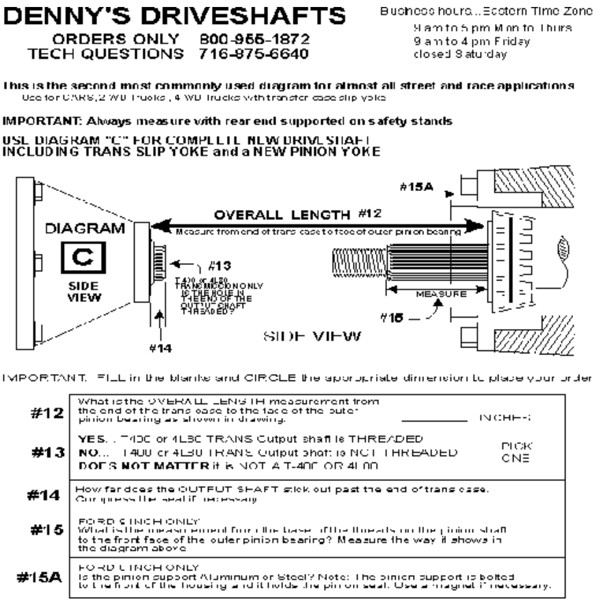
The similarity between W^* and W is

$$\rho(W^*, W) = \sum_{i=0}^{k-1} (x_i^* \bullet x_i) \Big/ \sqrt{\sum_{i=0}^{k-1} (x_i^*)^2} \tag{14}$$

The standard of whether the watermark exists is that: if $\rho(W^*, W) > T_2$, the watermark W exists in the detecting image, otherwise, the watermark doesn't exist. The false positive probability and the false negative probability should be accounted for when choosing T_2. If T_2 decreased, the false negative probability will reduced but the false positive probability will be enhanced. If T_2 increased, the false positive probability will reduced but the false negative probability will be enhanced. If W^* and W are not relative, the probability of $\rho(W^*, W) > T_2$ is equal the one of Gaussian distributing random variable which is T_2 as bigger as the average value.

Ⅴ EXPERIMENTS AND RESULTS

In order to test the proposed watermark scheme, many binary images were used in this experiment. Several common signal processing techniques and geometric distortions were applied to these binary images to evaluate if the detector can reveal the presence of the image owner's watermark. In this way, we can measure the algorithm robustness to various kinds of attacks. In this section, experimental result obtained on a piece of binary image as in Fig. 3 is displayed.



DENNY'S DRIVESHAFTS

ORDERS ONLY 800-955-1872
TECH QUESTIONS 716-875-6640

Business hours...Eastern Time Zone
9 am to 5 pm Mon to Thurs
9 am to 4 pm Friday
closed Saturday

This is the second most commonly used diagram for almost all street and race applications
Use for CARS, 2 WD Trucks, 4 WD Trucks with transfer case slip yoke

IMPORTANT: Always measure with rear end supported on safety stands

USE DIAGRAM "C" FOR COMPLETE NEW DRIVE SHAFT
INCLUDING TRANS SLIP YOKE and a NEW PINION YOKE

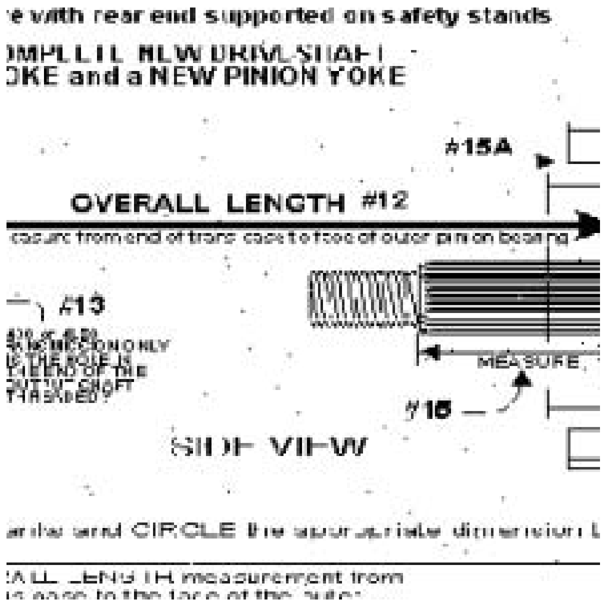
DIAGRAM
C
SIDE VIEW
#15A
OVERALL LENGTH #12
Measure from end of trans case to face of outer pinion bearing
#13
T400 or 4L80 TRANSMISSION ONLY IS THE HOLE IN THE END OF THE OUTPUT SHAFT THREADED?
MEASURE
#16
#14
SIDE VIEW

IMPORTANT: FILL in the blanks and CIRCLE the appropriate dimension to place your order

#12	What is the OVERALL LENGTH measurement from the end of the trans case to the face of the outer pinion bearing as shown in drawing. ________ INCHES
#13	**YES**... T400 or 4L80 TRANS Output shaft is THREADED **NO**... T400 or 4L80 TRANS Output shaft is NOT THREADED **DOES NOT MATTER** it is NOT A T-400 OR 4L80 PICK ONE
#14	How far does the OUTPUT SHAFT stick out past the end of trans case. Compress the seal if necessary
#15	FORD 9 INCH ONLY What is the measurement from the base of the threads on the pinion shaft to the front face of the outer pinion bearing? Measure the way it shows in the diagram above
#15A	FORD 9 INCH ONLY Is the pinion support Aluminum or Steel? Note: The pinion support is bolted to the front of the housing and it holds the pinion seal. Use a magnet if necessary.

Fig. 3 The host binary image

A comparison of the watermark embedded image basing on the former spatial domain algorithm and DCT algorithm are shown in Fig. 4 and Fig. 5 respectively. From these two figures, we can see that there are many dots on the embedded image with the spatial algorithm; however, the result image basing on the proposed DCT algorithm almost has no distortions.



e with rear end supported on safety stands
OMPLETE NEW DRIVE SHAFT
OKE and a NEW PINION YOKE

#15A
OVERALL LENGTH #12
easure from end of trans case to face of outer pinion bearing
#13
T400 or 4L80 TRANSMISSION ONLY IS THE HOLE IN THE END OF THE OUTPUT SHAFT THREADED?
MEASURE
#16
SIDE VIEW

anks and CIRCLE the appropriate dimension t

ALL LENGTH measurement from
is case to the face of the outer

Fig. 4 The partial enlarge image of the embedded image with the spatial domain algorithm

The response of a given mark is compared to T2 (T2=3, experimentally found to be the maximum detector response of extracting the virtual watermarks) to decide whether the mark is present or not. The watermarked images are exposed to common image processing operations including the geometric manipulations to test the algorithm robustness. The results of Wiener filtering, JPEG compression and cropping are discussed.

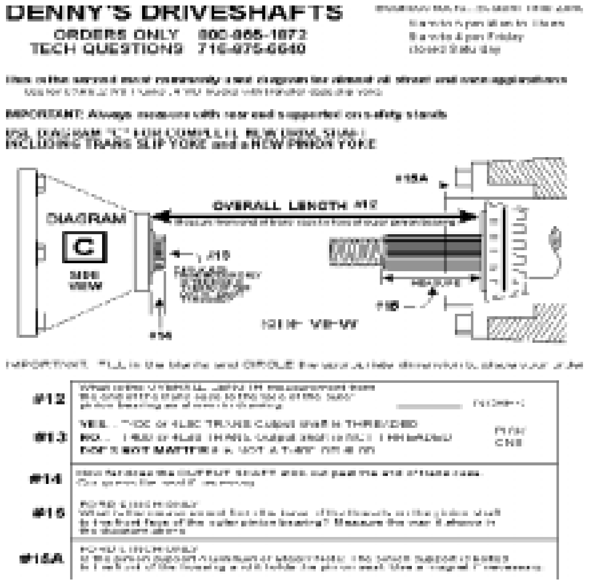

DENNY'S DRIVESHAFTS
ORDERS ONLY 800-865-1872
TECH QUESTIONS 716-875-6640
DIAGRAM
C
SIDE VIEW
OVERALL LENGTH #12
#15A
#16
#14
TOP VIEW
#12
#13
#14
#15
#15A

Fig. 5 The embedded image with the proposed DCT algorithm

Wiener filtering is able to remove the outliers in the image without reducing the sharpness of the image. This is shown in Fig. 6, where the effect of image filtering on watermark detection as the output of correlation coefficient versus the filter size is displayed. Also, this figure shows that when the binary image is watermarked based on the proposed scheme then filtered using Wiener filtering, it will not have a severe impact on the watermarking robustness.

The watermarked images are exposed to JPEG for different compression ratios (CRs). The correlation coefficient of the watermarking detector after compression using JPEG is presented in Fig. 7. The JPEG compression attack results in compression ratios ranging from 1:1 to 45:1.

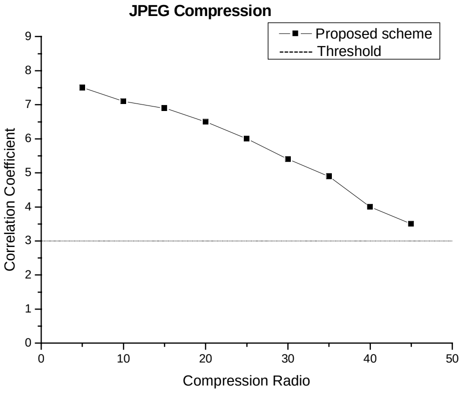


Fig.7 Correlation output of watermark detector response of successive compression ratios from 5 to 45 using JPEG

To achieve compression, high frequency discrete cosine transform (DCT) terms can be eliminated without affecting the image quality, but it might affect the integrity of the watermark information placed in the high frequency terms. This is clear that the robustness declines as the watermarked image is compressed more. However, the proposed system robustness is better than the former spatial domain algorithm, because after JPEG compression basing on the former algorithm, the watermark can't be detected entirely.

Following the watermark compression, the images were cropped, and the detection procedure was applied. A percentage of the image is cropped from 10% to 90%. The correlation coefficient output at each cropping level is shown in Fig. 8.

It can be concluded that the proposed system method is robust enough to the cropping attack, even when only a small percentage of the image area remains. The similarity for the proposed technique is high and above the detection threshold. This is because the watermark is inserted in each watermark segment as explained. Fig. 7 also demonstrates the superiority of our system compared to the former scheme. This is because through cropping, using the former algorithm, the authentication code embedded into the image cannot match the original one.

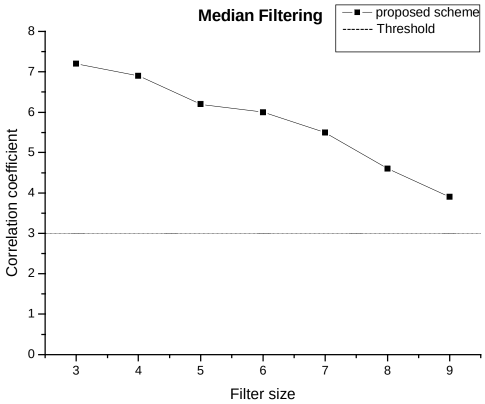


Fig. 6 Correlation output of watermark detector response of different Wiener filter size

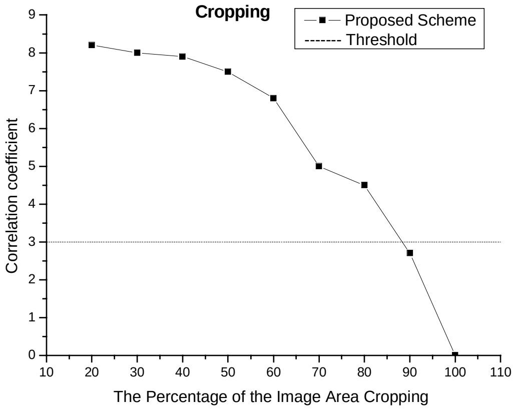


Fig. 8 Correlation coefficient as a function of the percentage of the image area cropping

VI CONCLUSION

To conclude, we have presented a new robust watermarking scheme and demonstrated its superior performance using different experiments. The results of experiments show that this approach is very promising, because it is robust to common image processing distortions. Our proposed system outperforms the spatial algorithm. For the compression attacks, it is found that the robustness against JPEG compression is achieved for a compression ratio (CR) of up to 45. Moreover, robustness against Wiener filters is shown for the 3*3 up to 9*9 pixel neighborhood. Also, the proposed system is inherently resistant to geometric manipulation. This is because we are utilizing the robust extracted features, which are covariant to these geometric transformations. Finally, we have shown that cropping attack is not effective in destroying the watermark.

VII REFERENCES

[1]. A. Bors and I. Pitas, “Image watermarking using DCT domain constraints,” in Proc. IEEE. Int. Conf. Image Processing, Lausanne, Switzerland, Sept. 1996, pp. 231~234.

[2]. M. D. Swanson, M. Kobayshi, and A. H. Tewfik, “Multimedia dataembedding and watermarking technologies,” Proc. IEEE, Vol.86, June 1998, pp. 1064~1087.

[3]. F. Hartung and M.Kutter, “Multimedia watermarking techniques,” Proc. IEEE, Vol.87, July 1999, pp. 1079~1106.

[4]. Chih-Hsuan Tzeng and Wen-Hsiang Tsai, “A new approach to authentixation of binary images for multimedia communication with distortion reduction and security enhancement,” IEEE Communications Letters, Sep. 2003, Vol.7, No.9.

[5]. Mohamed A. Suhail, and Mohammad S. Obaidat, “Digital Watermarking-Based DCT and JPEG Model,” IEEE Transactions on Instrumentation and Measurement, Oct, 2003, Vol.52, No.5.

Wu Jin received the B.Eng. degree in Electronics and Information Engineering in 1988 from Huazhong University of Science and Technology, and the M.Eng. Degree from Beijing University of Science and Technology in 1997 now is a doctor of Huazhong University of Science and Technology. She is a Vice Professor, also a head of Department of Electronics and Information Engineering, Wuhan University of Science and Technology. She has published over 10 Journal papers. Her main areas of interest are image processing and analysis, pattern recognition, information safety, multimedia telecommunication.

Xia Beibei received the B.Eng. degree in Computer Science and Technology in 2001 from China University of Mining and Technology, now is a master of Wuhan University of Science and Technology. Her main areas of interest are digital watermarking, network security.

Researches on Parallel Intrusion Detection Methods Based-on Network Processor

Li Fangmin [1, 2], **Zhang Huifu** [1,2], **Yang Ka** [2]
[1]**School of Computer Science and Engineering , Hunan University of Science and Technology, Xiangtan, 411201**
[2]**School of Information engineering, Wuhan university of technology, Wuhan, 430070**
Email: lfm68@sina.com **Tel:**008613973255023

ABSTRACT

According to the request of network development and the characteristic of Network Processor, a new network parallel intrusion detection model based on IXP1200 is presented. The parallel process efficiency is improved greatly by using a new method for load distribution based on Micro-Engine. Also, in order to generate more efficient Micro-C code, we suggest a new software architecture that reduces the time of operating memory and checking string.

Keywords: Network Processor, Intrusion Detection, parallel, char string checking

1. INTRODUCTION

1.1 Introduction of Network Processor

Network Processors，a new type of programmable processor which is used to process and forward data, mainly have two functional modules which are network processor unit and special intelligent coprocessor unit. The intelligent coprocessor which is supported by embedded operating system is the core of network system, and is used to control network processor unit and other hardware unit. Network processor unit generally uses a multi-threaded structure that can accomplish high-speed, strong-capacity intelligent data processing functions such as receiving and forwarding packets, packet header processing and route query etc. The hardware architecture of the network processor greatly increases the processing speed. The parallel working frame of multi-processor can improve the processing ability of network processors. Its inner hardware multi-threaded structure can reduce the average execution time when accessing registers and improve the chips efficiency. Real time requires different tasks accomplished by different processor, which can reduce packets delay in some degree. Seen from system flexibility, intelligent coprocessor unit is programmable, and network processor unit generally consists of more than one programmable RSIC, which provides adequate assurance for the network processors' flexibility.

The microengines of IXP1200 can fetch data directly from SRAM, SDRAM, and the FBI, but cannot exchange data directly between them. Instead, microengines can exchange messages through shared memory accesses that are expensive: for example, reading 4 bytes from the SRAM takes 22 cycles.

1.2 The brief introduction of Intrusion Detection System

We depend more and more on the network with the development of network techniques and scales. Meanwhile, network attack is becoming increasingly popular and complicated, which means it is far beyond the need of network security when only depending on the static security defense techniques such as the traditional operation system reinforcing and firewalls isolating. Intrusion detection systems have the functions such as, monitoring and analyzing the behavior of the user and system; auditing system configuration and leak; estimating sensitive system and integration of data; recognizing attack action; making a statistic of abnormal behavior; collecting packets related to system automatically; auditing and tracing the behavior against secure rules; recording the hacker behavior by trap servers, etc. Thus, administrators can monitor, audit and estimate the system more effectively.

1.3 Shortcomings of the present intrusion detection system

Intrusion detection system can't detect all packets effectively, and the network based intrusion detection system is difficult to match the network development speed. Capturing each packet from network then analyzing and matching to see whether there is a certain character of attack will consume much time and system resources. Currently, IDS has worked well in 10M~100M network when detecting all the packets with many attack characters. With the development of fast network techniques such as ATM, GIBT Ethernet etc, the speed of network development is far beyond that of packets pattern analysis.

The methods of detecting and analyzing are single while attack methods are becoming more and more complex. A kind of single analysis method based on the pattern matching and statistic is hard to find some attacks. Furthermore, both of them have their own virtues, so in one system it tends to use different methods at the same time. Nowadays, nearly all IDSs use a single analysis method.

The update of attack character libraries is not immediate enough. The pattern matching based analysis methods are fit for most intrusion detection systems appropriately, which needs the newest updated libraries. But now many intrusion detection systems haven't provided any methods such as "push technology" to update attack characters at any moment. Currently, it's evident that the need of network security can't be satisfied when considering many new leaks and attacks occur every day.

2. RESEARCHES ON PARALLEL INTRUSION DETECTION METHODS BASED ON NETWORK PROCESSOR

2.1 The advantage of network processor

A traditional intrusion detection system is based on PC infrastructure, and it can detect the intrusion events mainly by pattern matching or anomaly detection. IDSs based on PC infrastructure have bottleneck both in sniffing and processing

This work is supported by China Hunan Education Science Foundation #02A049 and Hunan NSF #03JJY1012.

ability. At present, the best way is to analyze the TCP/IP protocol stacks in NP, which actually releases not only their own work but also the burden of CPU and the efficiency are improved also. We can make the most of the NP microengine to realize parallel IDS analysis to reach the thread processing speed. Both pattern matching and anomaly detection have their inherent drawbacks, but we employ NP as a special CPU through infusing the virtue of pattern matching and anomaly detection together with protocol analysis techniques. In order to ensure the important information system running safely, this technology based on the fast network combines the virtue of NP and traditional intrusion detection with protocol analysis techniques, improves the network security by researching on the intrusion detection techniques used in large-scale network.

2.2 Researches on the execution efficiency of microengine

2.2.1 Parallel and mutex

We should thoroughly consider the condition when six microengines perform simultaneously, so that we can acquire great execution efficiency when writing microcode programs. Theoretically, this will occur if six microengines can perform simultaneously at any given time. But, it is inevitable to see mutex in any given multi-process or multi-processor system. For example, when a process executes writing operation on a date, others must wait till it’s over (e.g. read and write data). The numerous mutex operations decrease the system efficiency undoubtedly. The same problem exists in the IXP 1200 also. If a microengine takes lots of cycles to obtain and release a lock, while these cycles cannot used by other threads which are waiting for the same resources, the threads can only stagnate and the parallel execution efficiency cannot display fully. So we should try to avoid numerous mutex operations when writing a microcode program, and consider whether one thread alone can process a packet to avoid mutex.

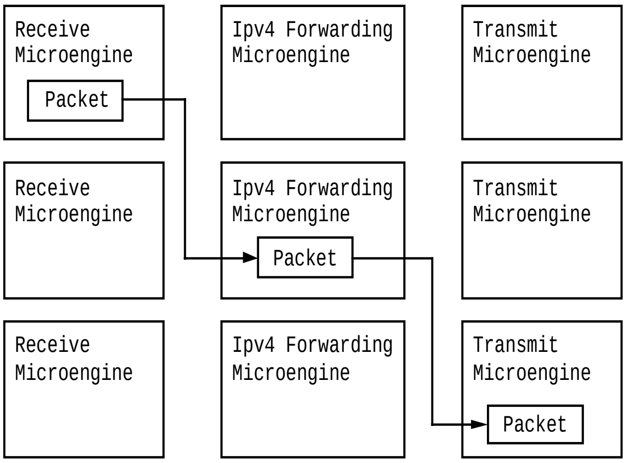


Fig.1. Using three microengines to receive, forward and Send respectively

It is scarce to use several threads to process one packet in most designs, but it is common to execute different functional threads to do it one by one. Considering that each microengine has 2kb of instruction memory, we just think of using several engines to realize the function if one cannot do. Below are two designs for one application [1]:

In figure 1, packages are received by a receiving microengine thread; then it is processed by an Ipv4’s forwarding microengine thread and finally sent by a sending microengine thread.

In figure 2, receiving microengine and Ipv4’s forwarding microengine are integrated to one microengine, and the sending microengine remains.

The design above varies from the practical applications. Each packet is processed in full by a single microengine, which is adopted in this design. The basic approach is therefore to assign each packet to a single worker for its entire processing: receiving, analyzing, sending etc. This results in a reasonably balanced system, as a busy worker cannot issue requests for more work and therefore new packets will be assigned to the least busy microengine. In this way, the chance to access shared resources dramatically decreases and the parallel efficiency increases.

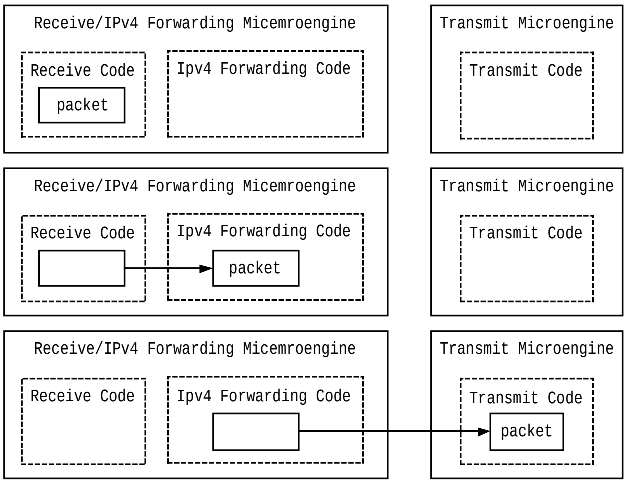


Fig.2. Combine the receiving and forwarding functions in one microengine

2.2.2 Researches on mutex scheme

Because of the multi-threaded structure of the IXP1200, a worker for a packet can be either an individual thread or an entire microengine. We therefore consider two different methods presented below.

2.2.2.1 Thread-based scheme

Thread-based scheme allocate the whole packet to one thread in each microengine. (Showed in figure 3)

This design has the advantage of simplicity, and yields the same codes for all threads. One drawback of the thread-based scheme is that the registers of each uEngine must be equally divided among the four threads. Each thread has to fetch the headers of its packet in its local registers. Each uEngine contains 128 32-bit general registers, and each thread can obtain 32*4=128-bytes general purpose registers only. 54 bytes are needed for Ethernet, IP and TCP headers. When the thread fetches a packet contained more than 128-bytes, parts of the data are stored in SDRAM or SRAM temporarily, and they will be moved to the registers only when they need processed. The whole process is expensive. Another disadvantage is that processing of the four packets inside the uEngine is done in an interleaved manner, meaning that only

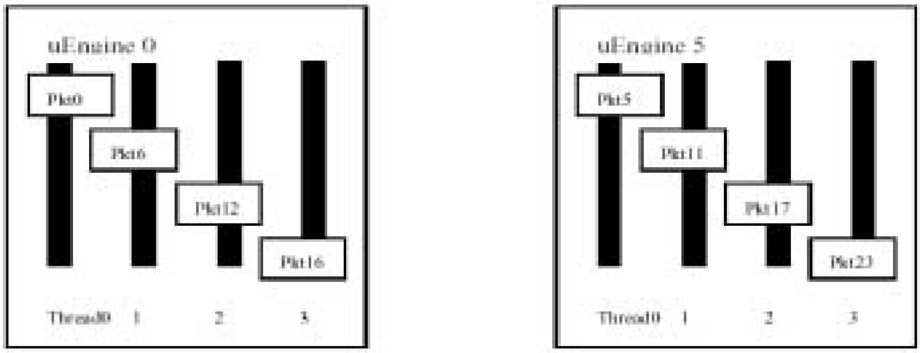


Fig.3. Thread-based scheme

one thread is active and only one fourth of these registers are actually used at any given time. The waste brought by the interplay between threads and SDRAM or SRAM can be alleviated via thread communication, but whether can we improve the performance of the system by reducing the interplay? As mentioned above, more threads load to more mutex manipulation. In an intrusion detection system, there is usually only one port to receive all packets, and numerous mutex manipulations appear when using multi-thread. First, we allow one thread in each uEngine to receive a packet. On the second step, up to six threads (one for each uEngine) race to lock the input port (i.e. access). The winning thread receives a packet, releases the acquired lock, and commences data checking. Meanwhile, the other threads can perform the same synchronization method in order to serve the next packet. Many cycles are wasted in vain during the competing course which can be overcome by the uEngine based scheme.

2.2.2.2 Microengine-based scheme

In the microengine-based scheme, an entire uEngine is allocated for serving a packet, and only one thread is in executing. (Showed in Figure 4). This thread holds all the resources such as registers. Meanwhile, it is responsible for the entire packet processing including moving the packet between uEngine and SDRAM the actual header processing.

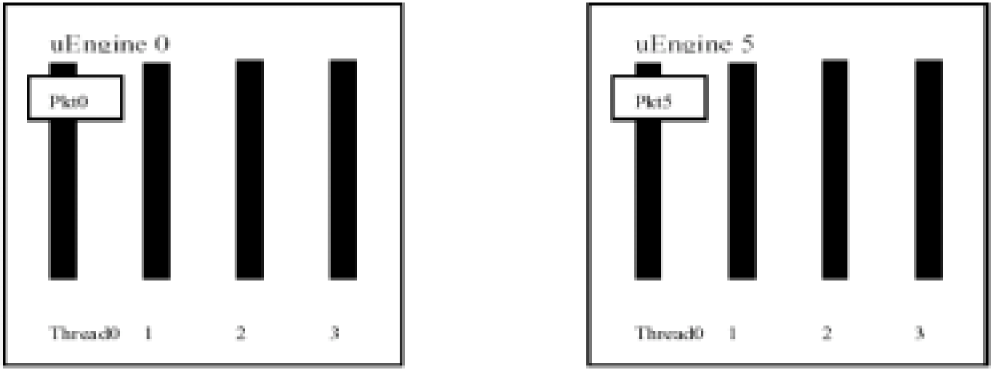


Fig.4. Microengine-based scheme

In contrast with the thread-based scheme, the uEngine based scheme can store the entire header (54bytes) and all (or parts) data of each packet to registers. The volume of the registers in each uEngine is 512 bytes, a small part of which is used to store the temporary results. In other words, we try to put the data to be analyzed in the register copy operations from SDRAM or SRAM and greatly enhance the performance of the system.

This scheme was supported further with a simpler synchronization method. Each uEngine signals the next one to start polling for arrived packets. Therefore accesses to memory are avoided, and the system behaves much more smoothly.

The two schemes demonstrate similar performance. However, since the uEngine-based scheme achieves much more, it was chosen to work with.

2.3 The network engine design involving microcode

The network engines obtain the possible intrusion information by analyzing the packets in the network. It's a tool accomplished not only data generating and forwarding, but also some data analyzing. All its functions equal to an integrated NIDS.

2.3.1 Innovation of detection matching method

Pattern matching used both in first and second generation intrusion detection system is a packet analysis technology based on attack characters. Its virtues such as fast analyzing speed and low error report ratio are far beyond. There are great disadvantages purely using pattern matching method, so we use protocol analysis and pattern matching method to analyze network packets in network engine. The process of pure pattern matching method is below:

Analyze each packet to find whether it has some attack characters, and analyzing shows below.

Compare the packets with attack characters from the header.
If the compared results are the same, a possible attack is detected.

If different, compare from the next position.

An attack character matching is over until detecting all bytes of attacks or packets accomplished.

For each attack character, repeat from step 2.

The matching toward a packet finished until each attack character matching accomplished.

The problems of traditional pattern matching:

Great calculation: For a special network, the maximum comparison times in one second are attack character bytes×packets×packets per second×numbers of attack characters. If the length of all attack characters is 20 bytes, average length of packets is 30 bytes, 3000 packets per second, 4000 characters character libraries. Then comparative times per second are 20×300×30,000×4,000=720.000.000.000.

Accuracy detection: Traditional pattern matching detection methods can only detect special attack which makes the detection fail even meets a minute transfiguration.

The essential problem is that it treats packets as orderless and random byte stream. Even though it knows nothing about inherent architecture of the packets, it matches the images or voices transmitted in network. But network communication protocol is a high formatted, specific meaning and valued data stream, and we can get an improved efficiency and accurate result if combine the protocol analysis and pattern matching method.

Protocol analysis effectively utilizes the hierarchy of network protocol and related knowledge to judge whether some attack characters exist quickly, which greatly reduces the matching calculation.

2.3.2 Protocol analysis and data analysis

The protocol analysis function is differentiating protocol styles of packets so that we can use mapped data analysis program to detect packets. We can make all the protocols into a protocol tree; a specific protocol is a node of the tree structure, which can be described by a binary tree. A network packet analysis is a path from root to a certain leaf. We can realize the flexible protocol analysis function by dynamically maintaining and configuring this tree in program.

We can add user-defined protocol node into this tree structure, for example, in HTTP protocol ask for URL as a node into this tree then make different ways in URL as child nodes. In this way, we can subdivide data and improve detection efficiency.

General intrusion detection systems realize this binary tree in memory. Analyzing packets is to search a node in a binary tree, find all function chains which detect this protocol, and utilize all its functions to do the pattern matching analysis quickly. In IXP 1200, we realize binary tree not in memory but directly in microinstruction when considering each microengine has 2k of instruction memory, thereby the execution time towards memory are reduced. (The microcode 2 in the example showed below)

```
…………………………………………………………
if (ETHPROTOCOL==0x800&&PROTOCOL=0X6){
  if (PROT2==0x50){
               if (IP2==0xa000001){
       if (IP1=0xa000000){
                              if (DSIZE>0X200){
/*Action for “tcp 10.0.0.0 any ->10.0.0.1…”*/
}}}
if (IP2==0xa000002){
       if (IP1=0xa000001||IP1==0xa000002){
              if (ACK>0x200){
/*Action for “tcp [10.0.0.1 10.0.0.2] any …”*/
}}}
…………………………………………………………
}//<<<<PORT2
   if(PORT2==0x14){
       if(IP2==0xa00012c) {
              if(DSIZE>0x200) {
/*Action for “tcp any any ->10.0.0.300 20 … */
}}}}
```

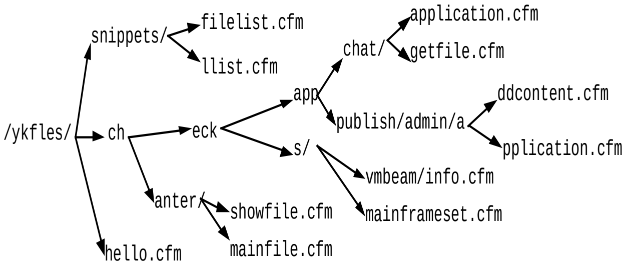
We can see that the tree-based microcode performs effectively and avoids redundant matching analysis. It’s analyzed that the efficiency is increased by 21.22% on average when using this tree.

2.3.3 Aho-Corasick_Boyer-Moore Hybrid Algorithm

The string-matching algorithm adopted by former IDSs is generally based on Aho-Corasick or Boyer_Moore algorithm, both of which have low efficiency as the analyzed string has to be compared with each character thoroughly. For example, certain data should be matched with all the character string bellowed and analyzed.

/ykfles/hello.cfm
/ykfles/checkapp/chat/application.cfm
/ykfles/checkapp/chat/getfile.cfm
/ykfles/checkapp/publish/admin/addcontent.cfm
/ykfles/checkapp/publish/admin/application.cfm
/ykfles/checks/vmbeam/info.cfm
/ykfles/checks/mainframeset.cfm
/ykfles/chanter/mainfile.cfm
/ykfles/chanter/showfile.cfm
/ykfles/snippets/list.cfm
/ykfles/snippets/filelist.cfm

Aho-Corasick or Boyer_Moore algorithm must start from the first character string to the last one in turn, so there are 11 repetitions in this example. We introduce an improved algorithms——Aho-Corasick or Boyer_Moore algorithm, for the two algorithms do the blind match one by one ignoring the characters of all the strings. This algorithm analyzes all the characters and displays all the information with a tree, which is shown below:



We start from the root node Later when we analyze character string matching and go on to compare the next children node when meeting the same parent node. If the child node doesn’t have a single successful matching or the children node is vacant, we consider the matching process fail and repeat the above procession toward the next character string till the last character string. Only one matching is needed to finish all the work from begin to end in the algorithm and its efficiency can increase 10 times than general algorithm especially when numerous character strings have similar prefix.

3. CONCLUSIONS

This paper introduces the basic conception of network processor, gives a simple description of the basic architecture of IXP 1200 produced by Intel Co., then introduces the IDS and the present IDS' deficiency, and finally gives a research method which based on network processor to realize the IDS. The emphases on how to improve the parallel processing ability and the microcode efficiency of IXP 1200 have been discussed. We provide two kinds of schemes which are based on thread and microengine, and analyze both the advantages and the disadvantages. Considering the character of IDS protocol analysis and system structure of IXP 1200, the paper makes full use of the 2k of instruction memory realized the binary tree directly in microcode. This bin tree is used in optimized protocol analysis and improves the performance of the system furthest. This paper only gives a primary discussion on how network processor realizes IDSs, and we'll gradually perfect the functions in future to get a better joint with the network processor in order to meet with the need of the fast development in present networks.

4. REFERENCES

[1]Erik J.Johnson and Aaron R.Kunze .IXP1200 Programming. The Microengine Coding Guide for the Intel IXP1200 Network Processor Family. INTEL PRESS.2002.

[2]Charitakis, D. Pnevmatikatos, E.Markatos, and Anagnostakis. Code Generation for Packet Header Intrusion Analysis on the IXP1200 Network Processor. September 2003.

[3]Neil Desai. Increasing Performance in High Speed NIDS. 2003.

[4]J.Frank.Artificial Intelligence and intrusion detection: Current and future directions. In Proceedings of the 17th National Computer Security Conference, October 1994.

A General Dynamic Secret Sharing Algorithm in Distributed System

[1] Li Xiaoxin, [2] Guo Qingping, Zhang Feng
School of Computer Science and Technology, Wuhan University of Technology, Yujiatou Campus,
Wuhan 430063, China.
Email: liwonder@sina.com[1] qpguo@mail.whut.edu.cn[2]

ABSTRACT

In order to produce, manage and protect distributed key more safely, especially for the private key of the CA center, Country's military secret and so on, a General Dynamic Secret Sharing Algorithm has been proposed, which based on the study of a Proactive Secret Sharing Algorithm [3] and a General Verifiable Secret Sharing protocol [6]. The algorithm makes Secret Sharing more practical and more applicable. And some new conception has been presented, such as Virtual Shares, Increased Shares and Working Shares. A scheme of dynamic renewing each minimal authorized subset has also been presented and realized, which makes adversary don't know how to attack and makes the Secret Sharing system more safely.

Key Words: secret sharing; general dynamic secret sharing; access structure; minimal authorized subset; virtual shares.

1. PRESENT THE PROBLEMS

For the field of communication security, the "tenth five" 863 plan of our country treated distributed key algorithm and parallel technology as an important study task and pointed out," Stress the study of symbol computing technology in the application of cryptography, the design and realization of distributed and parallel algorithm in network environment." The thought of Secret Sharing is very important for the study of distributed key and has great influence especially on Distributed Key Generation[7] and Management, and Private Key Conservation and Management[3] in distributed environment. However, almost all secret sharing algorithms available still have shortcomings. Due to these shortcomings, an information-theoretic secure general dynamic secret sharing algorithm, which is applicable to arbitrary monotone access structures and could be used to generate and manage distributed key, is proposed.

1.1 Analyze the proactive secret sharing algorithm in Ref.[3]

In order to protect long-lived and highly sensitive private keys, a Proactive Secret Sharing Algorithm for Dlog-based secret has been presented in Ref.[3] , which is based on Ref.[2] and using Pedersen-VSS algorithm. In this algorithm, shares are renewed periodically (without changing the shared secret) so that information gained by adversaries in one period will be useless after the shares are renewed in next period. This algorithm has improved greatly the security of highly sensitive private keys and can detect the dubious participants. It is fairly successful in security.

However, analyze the algorithm carefully and we can discover some shortcomings as following:
(1) Pedersen-VSS, on which the algorithm is based to share secret, is based on (t,n) threshold scheme. Threshold schemes are all based on an assumption that all participants are equal in status, right and dependability. However, the assumption is very hard to satisfy in fact.
(2)At the beginning of the algorithm, there is only one dealer distributes shares to participants in *ShareGen* Algorithm. But it is very insecure to distribute shares only by one person. If the dealer were controlled by a adversary, share-holders would not be able to verify the validity of the shares they have received.

In Ref.[6], ZHANG Fu-Tai has presented an information-theoretic secure general VSS (for short, ZFT_GVSS), which can be used to resolve the first shortcoming. Based on ZFT_GVSS, a secure and applicable Distributed Key Generation Protocol has been designed in Ref.[7]. This protocol can be used to resolve the second shortcoming.

1.2 Analyze ZFT_GVSS

Let's see ZFT_GVSS. Analyze Ref.[6] carefully and we can know that ZFT_GVSS is more general than Pedersen-VSS just because the former has imported monotone access structures Γ and its radix Γ_0. The participants of each minimal authorized subset in Γ_0 could cooperate to reconstruct the shared secret. But the number of participants in each minimal authorized subset doesn't need to be equal. Thereby, the confine of threshold scheme is broken. That is to say that the number to reconstruct the shared secret is decided by the number of participants in each minimal authorized subset, not by the threshold value appointed at the beginning. Obviously, this scheme is more progressive than threshold schemes.

However, threshold schemes and ZFT_GVSS scheme both tell everyone, including the adversaries, which participants can be combined to reconstruct the shared secret: they are any t participants in the (t,n) threshold scheme and are the participants of any minimal authorized subset in ZFT_GVSS scheme. Assuredly, these schemes have offered specific targets to the adversary. He could reconstruct the shared secret as long as the adversary could bribe any one of the combinations. And it will be useless to renew the shares of the participants who have been bribed.

And, the threshold schemes and ZFT_GVSS scheme suppose that there are some strong admissible participants who would never be bribed. However, which participant could promise that he would never be bribed?

How to make the adversary have no way to know the combination, in which the participants could be combined to reconstruct the shared secret? Or even if he knew such a combination and could bribe all the participants in it, how to make him not be able to use them to reconstruct the shared secret, for such a combination has been overdue? These are the kernel problem in this paper.

According to the analysis above, this paper presents a General Dynamic Secret Sharing Algorithm to resolve the above problems. Call it general, because the algorithm is based on the thought of general secret sharing, not on threshold scheme; Call it dynamic, because the shared secret will be dynamic generated, not be appointed by one dealer; the share of each participant is renewed periodically, not fixed; the participants who could be combined to reconstruct the shared secret are changed periodically, not appointed at the beginning.

2. DESCRIPTION OF THE ALGORITHM

2.1 The definitions of some parameters before Algorithm

(1) The parameters of the system are in accordance with Ref.[6]. But the participants of the system are slightly different. List them as following (according to our habits, we changed some identifiers of the parameters):

System parameters: the public values are (p,q,g,h), where p,q are large primes, $q\,|\,(p-1)$, Z_q is the only subgroup with order q, g and h are two generators in Z_q with order q. None knows the discrete logarithm $\log_g h$.

Each participant in the system belongs to a set $P_{\mathrm{a}} = \{p_1, p_2, \Lambda\ , p_n\}$, in which each participant act as dealer as well as participant.

Some explanations: Define $E(a,b) = g^a h^b$. The formula $a^{(m)}$ expresses the value of a in period m. There are many places where multinomial is constructed by some points in the paper. We all use the method of Lagrange interpolation and don’t point it out clearly.

(2) Redefine monotone access structure

Usually, the participants have different statuses in actual life. According to the statuses, the shared secret S might need to be shared and recovered by different members come from different statuses.

For example, a key is asked to be inputted when to launch an A-bomb. Many administrative staffers have shared the key and the key should have some participants come from different statuses recovered. Now there are administrative staffers and solution schemes listed in Table 1.

Table 1

Statuses Schemes	President (1)	Vice president (3)	Councilmen in senate (20)	Councilmen in commons (50)
Scheme 1	1	2	10	0
Scheme 2	0	2	10	20

As table 1, we know that there is one president, 3 vice presidents, 20 councilmen in senate, and 50 councilmen in commons. There are two schemes to launch an A-bomb. E.g., in scheme 1, one president, any two of the three vice presidents and any half of councilmen in senate can recover the key.

Based on this instance, the following scheme was proposed: divide the n participants into r levels, such as $R_1, R_2, \Lambda\ , R_r$. R_i has Δ_i participants, $n = \sum_{i=1}^{r} \Delta_i$. Each minimal authorized subset is defined as Table 2. This table shows the minimal authorized subset A_i needs Δ_{ij} participants in level R_j, $\Delta_{ij} \le \Delta_j$.

Table 2. The relation between the level and the minimal authorized subset

Level authorized subset	$R_1(\Delta_1)$	$\cdots$	$R_r(\Delta_r)$
A_1	Δ_{11}	$\cdots$	Δ_{1r}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_t	Δ_{t1}	$\cdots$	Δ_{tr}

Base on this relation table, we define Γ_0, the radix of the monotone access structure Γ: $\Gamma_0 = \{A_1, A_2, \Lambda\ , A_t\}$.

System broadcasts the table before the algorithm begins.

Because the Share Distribution Algorithm, Sharing Renewing Algorithm, the Minimal Authorized Subset Renewing Algorithm will all use the random multinomial of degree t to share some value, such as each participant’s x_i and y_i in the Sharing Distribution Algorithm. Hence, based on Algorithm *Joint-RS(t)*[3], a Random Multinomial Sharing Algorithm will be presented at first, which will be suitable for the General Secret Sharing Algorithm. The Random Multinomial Sharing Algorithm is the soul of the General Dynamic Secret Sharing Algorithm.

2.2 Random Multinomial Sharing Algorithm $RS(d,P,func,t,r,a,b,c)$

Means of the parameters:

d —a dealer；

P —a set of participants, which can include some participants and subsets of some participants , e.g. $P = \{p_1, p_3, \{p_2, p_5, p_8\}, \Lambda\ \}$；

$func$ — point out the purpose of running this algorithm；

t —the degree of the multinomial $u(z), v(z)$, generated by d；

r,a,b —prescribe the property the two multinomial should meet：

$u(r) = a, v(r) = b$； To do the same job, each participant running $RS(d,P,func,t,r,a,b,c)$ should have the same value of r .

c —if $func = 0$， c is the number of new shares period; if $func = 1$， C is the number of current shares period.

The function of this algorithm: the dealer d generates two random multinomials, by which d distributes an **Increase Share**, (which will be used to increase the old shares of participants in P) to each participant. After each participant in P has received and verified his increase share, he uses his old share to add his increase share to get a new share; d constructs a multinomial and sends a increase share to the virtual participant of each minimal authorized subset in Γ_0 (when $func = 0$) or in P (when $func = 1$).

Algorithm $RS(d,P,func,t,r,a,b,c)$ will be used in many cases. However, in sum, there are two kinds of cases as the following:

① $func = 0$, the share of each participant in P and the share of the virtual participant in each minimal authorized subset need to be renewed forever(i.e., all the old shares will not be used again or will be out of date), this case is relative to periodicity(here, P must be equal P_a);

② $func = 1$, the participants in P need to use new shares to do a specific job(so, we call the new share as Job Share). The purpose is not to renew the shares but to complete some specific job. This case is irrespective to periodicity. Here, the new shares are just temporary, and the old shares should be reserved.

Some explanations before the algorithm:

Because the case of $func = 1$ is irrespective to periodicity, when $func = 1$, there will be no (m) at the top right corner of all identifiers; most processes of the 2 kinds of cases of are consistent with each other, the default case is $func = 0$; and for the case of $func = 1$, all the identifiers' (m) should be gotten rid of; if the processes of the 2 kinds of cases are different from each other, notations will be given.

Suppose the introduced parameters be $d = p_o \in P_a$, $r = r^{(m)}$, $a = a_o^{(m)}, b = b_o^{(m)}$, $c = m$; If $m = 0$, make $m - 1 = 0$.

The steps of the algorithm:

(1) Release all the subsets in P

Suppose $P = \{P_1, P_2, \Lambda\}$. Because the element P_i in P can be a single participant or a subset of some participants, in order to assure the algorithm can run accurately, the elements of all subsets in P should be released if there are. Operate as following:

Rebuild a new set P' and copy all elements in P into P'. Release all the elements of all the subsets in P' into P' and delete all the subsets in P'. Now each element in P' is a single participant. And then, check if there are repeated elements in P'. For the repeated elements, only leave one in P' and delete the others.

After these operations, there are w participants in P', and $P' = \{\gamma_j \mid \gamma_j \in P_a, \gamma_j = p_i, j = 1, \Lambda, w\}$. Use $\#(\gamma_j)$ to express the subscript of the corresponding of γ_j one in P_a.

(2) d distributes increase share (α_i, β_i) to each participant γ_j in P'.

① d select two random multinomial from Z_q:

And the two multinomial meet $u_o^{(m)}(r^{(m)}) = a_o^{(m)}, v_o^{(m)}(r^{(m)}) = b_o^{(m)}$. Calculate and broadcast the proof, $(a_{ok}^{(m)}, b_{ok}^{(m)})$, of each modulus in the two multinomial:

$E_{ok}^{(m)} = E(a_{ok}^{(m)}, b_{ok}^{(m)}) = g^{a_{ok}^{(m)}} h^{a_{ok}^{(m)}}$, $k = 0,1,\Lambda, t$. Send (α_i, β_i) to each participant γ_j in p'. Here, $i = \#(\gamma_j)$, $\alpha_i = u_o^{(m)}(i), \beta_i = u_o^{(m)}(i)$.

② Every participant γ_j in p' verify if the increase share (α_i, β_i) he receives meets:

$$E(\alpha_i, \beta_i) = \prod_{k=0}^{t} (E_{ok}^{(m)})^{i^k} \bmod p \tag{1}$$

If the left was not equal the right, γ_j broadcasted a complaint "complaint p_o".

③ If d received a complaint from γ_j, d should broadcast α_i and β_i.

④ If all the participants in P' have complained d or as the answer of one complaint, the values α_i and β_i broadcasted by d can not meet equation (1), d will be thought as **a controlled server**.

⑤ If d is not a controlled server, execute step (3); if d is a controlled server, operate as following:

If $func = 0$, γ_j makes α_i and β_i into 0 and goes on to step (3); else the algorithm fails and ends.

(3) Each participant γ_j in P' gets a new share or a job share

After verifying the share, γ_j gets a new share $(s_i^{(m)}, e_i^{(m)})$ ($i = \#(\gamma_j)$) and $s_i^{(m)} = s_i^{(m-1)} + \alpha_i$, $e_i^{(m)} = e_i^{(m-1)} + \beta_i$. $(s_i^{(m-1)}, e_i^{(m-1)})$ is the share of last period. γ_j deletes $(s_i^{(m-1)}, e_i^{(m-1)})$ and reserve the new share $(s_i^{(m)}, e_i^{(m)})$.

If $func = 1$, γ_j gets a temporary share $(s_i', e_i')(i = \#(\gamma_j))$. We call it as Job Share. Here $s_i' = s_i^{(m)} + \alpha_i$, $e_i' = e_i^{(m)} + \beta_i$. Attention, the Job Share is just used to complete a specific job and the old share $(s_i^{(m)}, e_i^{(m)})$ is still reserved. After the job is finished, γ_j deletes his Job Share.

(4) Renew the virtual shares of the minimal authorized subsets

If $func = 0$, renew the virtual shares of all the minimal authorized subsets in Γ_0; if $func = 1$, just renew the virtual shares of all the minimal authorized subsets in P. Operate as the following:

For the minimal authorized subset $A_i = \{p_{i_1}, p_{i_2}, \Lambda, p_{i_k}\}$, d uses $(0, u_o^{(m)}(0))$ and $(i_1, u_o^{(m)}(i_1)), (i_2, u_o^{(m)}(i_2)), \Lambda, (i_k, u_o^{(m)}(i_k))$ to construct a multinomial with degree k, $I_{ko}^{(m)}(z)$, and gets $s_{i(n+1)}^{(m)} = s_{i(n+1)}^{(m-1)} + I_{ko}^{(m)}(n+1)$; if $func = 1$, $s_{i(n+1)}' = s_{i(n+1)}^{(m)} + I_{ko}(n+1)$.

And then, d uses $(0, v_o^{(m)}(0))$ and $(i_1, v_o^{(m)}(i_1)), (i_2, v_o^{(m)}(i_2)), \Lambda, (i_k, v_o^{(m)}(i_k))$ to construct a multinomial with degree k, $IV_{ko}^{(m)}(z)$, and gets $e_{i(n+1)}^{(m)} = e_{i(n+1)}^{(m-1)} + IV_{ko}^{(m)}(n+1)$; if $func = 1$, $e_{i(n+1)}' = e_{i(n+1)}^{(m)} + IV_{ko}(n+1)$.

Now, A_i gets its virtual share $(s_{i(n+1)}^{(m)}, e_{i(n+1)}^{(m)})$ or its virtual Job Share $(s_{i(n+1)}', e_{i(n+1)}')$ in period m. And p_o

broadcasts the virtual share or the virtual Job Share.

For each minimal authorized subset $A_i = \{p_{i_1}, p_{i_2}, \Lambda, p_{i_k}\}$, all the participants in it check the validity of the virtual share $(s^{(m)}_{i(n+1)}, e^{(m)}_{i(n+1)})$ or the virtual Job Share $(s'_{i(n+1)}, e'_{i(n+1)})$ by equation $E(s^{(m)}_{i(n+1)}, e^{(m)}_{i(n+1)}) = \prod_{j=0}^{k} (\prod_{k=0}^{t} (E^{(m)}_{ok})^{(i_j)^k})^{b_j}$. Here, $i_0 = 0$, $b_j = \prod_{t=0}^{k} \frac{n+1-i_t}{i_j - i_t}$.

(5) At last, p_o deletes all the random multinomial and all the increase shares he has given out.

2.3 Share distribution algorithm

(1) Dynamic generate all the minimal authorized subsets in Γ_0

① The system selects at random a participant p_i as a dealer D. In accord with table 2, D selects at random participants for each minimal authorized subset A_i . Get $A'_i = \{p_{i_1}, p_{i_2}, \Lambda, p_{i_k}\}$. D sends the subscript of A_i to each participant in A'_i. We call this step as D arranges for A_i.

② After p_{i_t} receives the information i, p_{i_t} check the number *Subsets* of the minimal authorized subsets that he has joined. In order to avoid him to be too tired p_{i_t} has the right to refuse to join A_i. If *Subsets* is equal to some value *Tire*, p_{i_t} will refuse to join any minimal authorized subsets(if *Subsets* =*Tire*, we think p_{i_t} is in a **Saturated State**); if *Subsets* < *Tire*, p_{i_t} accepts the arrangement and adds 1 to *Subsets*.

③ If D received a refusing information coming from p_{i_t}, D select another participant p_m in the level that p_{i_t} lies in and replace p_{i_t} with p_m and repeat step ③; If D hasn't received any refusing information, arrangement for the minimal authorized subset A_i is ok and D will arrange for the next minimal authorized subset A_j.

④ After D has arranged all the minimal authorized subsets, D broadcasts the members of all the minimal authorized subsets to all the participants and arranges p_{n+1} as a virtual member for all the minimal authorized subsets. We calls the share of the virtual member in each minimal authorized subset as **Virtual Share**. The Virtual Share of A_i in period m is $(s^{(m)}_{i(n+1)}, e^{(m)}_{i(n+1)})$ and $(s^{(0)}_{i(n+1)}, e^{(0)}_{i(n+1)})$ is initialized to zero.

(2) All the participants of the system (not including the virtual member) execute the following steps at the same time, handing out shares to all the participants and virtual shares to all the minimal authorized subsets (Virtual Shares are public to all the participants):

①Each participant p_o initialize his first share $(s^{(0)}_i, e^{(0)}_i)$ to zero and selects two elements x_o, y_o in Z_q. x_o acts as the secret that p_o will share to all and y_o acts as the correlative random value.

② p_o makes $d = p_o, P = \{p_1, p_2, \Lambda, p_n\}, func = 0, t = n, r = 0, a = x_o$, $b = y_o, c = 0$ and execute $RS(d, P, func, t, r, a, b, c)$.

(3) After the parallel executions over, the shares with regard to the shared secret S of the members and the virtual member in each minimal authorized subset $A^{(\omega)}_i = \{p_{i_1}, \Lambda, p_{i_k}\}$ are:

$$s^{(0)}_{i_\alpha} = \sum_{o=1}^{n} u^{(0)}_o(i_\alpha) \quad (\alpha = 1, \Lambda, k), \tag{2}$$

$$s^{(0)}_{i(n+1)} = \sum_{o=1}^{n} I^{(0)}_{ko}(n+1) \tag{3}$$

And the shares with regard to e is similar to S and don't show them again.

If the shred secret is a private key, the public key y will be $y = \prod_{o=1}^{n} E^{(0)}_{o0}$.

Now we will execute the Shares Renewing Algorithm to renew the shares of all the participants every period (we call the period as Share Renewing Period, for short, Period) but the shared secret will not be changed.

2.4 Shares Renewing Algorithm

Our Shares Renewing Algorithm is in accord with the *Refresh(m)* algorithm in Ref.[3] on the whole. However, our algorithm will renew the virtual shares.

(1) All the participants $p_o \ (o = 1, \Lambda, n)$ of the system execute the following operations every other Period (suppose the current period be $m - 1$):

Make $d = p_o, P = \{p_1, p_2, \Lambda, p_n\}, func = 0, t = n,$ $r = 0, a = b = 0$, $c = m$ and execute $RS(d, P, func, t, r, a, b, c)$.

(2) After the parallel executions over, the shares with regard to the shared secret S of the members and the virtual member in each minimal authorized subset $A^{(\omega)}_i = \{p_{i_1}, \Lambda, p_{i_k}\}$ are:

$$s^{(m)}_{i_\alpha} = s^{(m-1)}_{i_\alpha} + \sum_{o=1}^{n} u^{(m)}_o(i_\alpha) \quad (\alpha = 1, \Lambda, k) \tag{4}$$

$$s^{(m)}_{i(n+1)} = s^{(m-1)}_{i(n+1)} + \sum_{o=1}^{n} I^{(m)}_{ko}(n+1) \tag{5}$$

And the shares with regard to e is similar to S and don't show them again.

2.5 Minimal Authorized Subsets Renewing Algorithm

The members and the Virtual Shares in each minimal authorized subset A_i should be renewed every other period (we call this period as **Minimal Authorized Subsets Renewing Period**). Suppose the current Minimal Authorized Subsets Renewing Period be ω and the Period be m . Now we will renew $A^{(\omega)}_i = \{p_{i_1}, p_{i_2}, \Lambda, p_{i_k}\}$.

(1) Renew the members

Select one member from each level having members in $A_i^{(\omega)}$ and use this member select a substitution from the members haven't joined $A_i^{(\omega)}$ in the level the member lies in and replace the member with the substitution. However, the substitution must meet two conditions: First, haven't joined A_i recently; second, not be in the Saturated State. Now, the members of A_i has been renewed and $A_i^{(\omega+1)} = \{p_{i_1}^{'}, p_{i_2}^{'}, \Lambda, p_{i_k}^{'}\}$. A_i broadcasts his members and makes $(s_{i(n+1)}^{(m)}, e_{i(n+1)}^{(m)})$ =0.

(2) Renew the virtual shares

Step1. Select another two minimal authorized subsets from Γ_0：

$A_j^{(\omega)} = \{p_{j_1}, p_{j_2}, \Lambda, p_{j_\theta}\}$, $A_h^{(\omega)} = \{p_{h_1}, p_{h_2}, \Lambda, p_{h_\rho}\}$. $A_j^{(\omega)}$ and $A_h^{(\omega)}$ execute the following steps at the same time:(e.g. $A_j^{(\omega)}$)

Step1.1 Each member $p_{j_\delta}(\delta = 1, \Lambda, \theta)$ in $A_j^{(\omega)}$ executes the following steps at the same time:

Make $d = p_{j_\delta}$, $P = \{A_j^{(\omega)}, A_i^{(\omega+1)}\}$, $func = 1$, $t = n, r = n+1$, $a = b = 0$, $c = m$ and execute $RS(d, P, func, t, r, a, b, c)$. If return any p_{j_δ} is a controlled server, select other minimal authorized subsets from Γ_0 and execute Step1.1 again. If $RS(d, P, func, t, r, a, b, c)$ was completed successfully, the members in $A_i^{(\omega+1)}$ would get Job Shares:

$$s_{i_\alpha}^{'} = s_{i_\alpha}^{(m)} + \sum_{\delta=1}^{\theta} u_{j_\delta}(i_\alpha), (\alpha = 1, \Lambda, k) \tag{6}$$

$$s_{i(n+1)}^{'} = s_{i(n+1)}^{(m)} + \sum_{\delta=1}^{\theta} \mathrm{I}_{kj_\delta}(n+1) = \sum_{\delta=1}^{\theta} \mathrm{I}_{kj_\delta}(n+1) \tag{7}$$

The members in $A_j^{(\omega)}$ would get Job Shares:

$$s_{j_\alpha}^{'} = s_{j_\alpha}^{(m)} + \sum_{\delta=1}^{\theta} u_{j_\delta}(j_\alpha), (\alpha = 1, \Lambda, \theta) \tag{8}$$

$$s_{j(n+1)}^{'} = s_{j(n+1)}^{(m)} + \sum_{\delta=1}^{\theta} J_{\theta j_\delta}(n+1) \tag{9}$$

Step1.2 $A_j^{(\omega)}$ compute Job Secret for $A_i^{(\omega+1)}$

$A_j^{(\omega)}$ uses $(j_1, s_{j_1}^{'}), \Lambda, (j_\theta, s_{j_\theta}^{'})$ and $(n+1, s_{j(n+1)}^{'})$ $\theta + 1$ points and gets:

$$s^{'} = \sum_{\alpha=1}^{\theta} s_{j_\alpha}^{'} \left(\prod_{\substack{\beta=1 \\ \beta \neq \alpha}}^{\theta} \frac{-j_\beta}{j_\alpha - j_\beta} \cdot \frac{-(n+1)}{j_\alpha - (n+1)}\right) + s_{j(n+1)}^{'} \prod_{\beta=1}^{\theta} \frac{-j_\beta}{n+1-j_\beta} \tag{10}$$

In the same way, $A_j^{(\omega)}$ gets $e^{'}$.

$(s^{'}, e^{'})$ the secret shared by the current Job Shares of $A_j^{(\omega)}$ and $A_i^{(\omega+1)}$. So we call $(s^{'}, e^{'})$ as **Job Secret**. $A_j^{(\omega)}$ sends $(s^{'}, e^{'})$ to $A_i^{(\omega+1)}$.

Step2. $A_i^{(\omega+1)}$ receives $(s^{'}, e^{'})$ and uses the Job Shares gotten by the Increase Shares sent by the members in $A_j^{(\omega)}$ to do the following steps：

$A_i^{(\omega+1)}$ constructs multinomial by $(i_1, s_{i_1}^{'}), \Lambda, (i_k, s_{i_k}^{'})$ and $(0, s^{'})$points:

$$I_k(z) = \sum_{\alpha=1}^{k} s_{i_\alpha}^{'} \left(\prod_{\substack{\beta=1 \\ \beta \neq \alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{n+1}{i_\alpha}\right) + s^{'} \prod_{\beta=1}^{k} \frac{z - i_\beta}{-i_\beta}$$

Make $z = n+1$ and get:

$$s_{i(n+1)}^{(m)'} = I_k(n+1) \tag{11}$$

In the same way, $A_j^{(\omega)}$ gets $e_{i(n+1)}^{(m)'}$.

Then compute：$s_{i(n+1)}^{(m)} = s_{i(n+1)}^{(m)'} - s_{i(n+1)}^{'}$, (12)

$e_{i(n+1)}^{(m)} = e_{i(n+1)}^{(m)'} - e_{i(n+1)}^{'}$.

Now $A_i^{(\omega+1)}$ gets its virtual share $(s_{i(n+1)}^{(m)}, e_{i(n+1)}^{(m)})$, marked as $([s_{i(n+1)}^{(m)}]_j, [e_{i(n+1)}^{(m)}]_j)$.

In the same way, $A_i^{(\omega+1)}$ would get another virtual share $([s_{i(n+1)}^{(m)}]_h, [e_{i(n+1)}^{(m)}]_h)$.

If $([s_{i(n+1)}^{(m)}]_j, [e_{i(n+1)}^{(m)}]_j)$ = $([s_{i(n+1)}^{(m)}]_h, [e_{i(n+1)}^{(m)}]_h)$, $A_i^{(\omega+1)}$ could confirm its virtual share $(s_{i(n+1)}^{(m)}, e_{i(n+1)}^{(m)})$ = $([s_{i(n+1)}^{(m)}]_j, [e_{i(n+1)}^{(m)}]_j)$ = $([s_{i(n+1)}^{(m)}]_h, [e_{i(n+1)}^{(m)}]_h)$ and also could confirm $A_j^{(\omega)}$ and $A_h^{(\omega)}$ are **Creditable Minimal Authorized Subsets**.

If $([s_{i(n+1)}^{(m)}]_j, [e_{i(n+1)}^{(m)}]_j) \neq ([s_{i(n+1)}^{(m)}]_h, [e_{i(n+1)}^{(m)}]_h)$, there will be at least one dishonest subset among $A_j^{(m)}$ and $A_h^{(\omega)}$. So $A_i^{(\omega+1)}$ can't confirm its virtual share. Another minimal authorized subset has to be selected from Γ_0. Going on to compute and compare the virtual shares in the above way. To operate in this way until $A_i^{(\omega+1)}$ gets two virtual shares that are equal in value when the virtual share of $A_i^{(\omega+1)}$ and the credible minimal authorized subsets could be confirmed.

In this procession, the system can find which minimal authorized subsets are dishonest and can use the Creditable Minimal Authorized Subsets to rebuild all the shares of dishonest minimal authorized subsets by Shares Rebuilding Algorithm [3].

(3) For all the minimal authorized subsets except $A_i^{(\omega+1)}$, execute step (1) and step (2). In step (2), just use the selected Creditable Minimal Authorized Subsets.

(4) At last, the two Creditable Minimal Authorized Subsets announce their virtual shares are out of date. And then all the participants cooperate to execute the Shares Renewing Algorithm.

2.6 Shared Secret Reconstructing Algorithm

For a minimal authorized subset $A_i^{(\omega)} \in \Gamma_0$ and $A_i^{(\omega)} = \{p_{i_1}, \Lambda, p_{i_k}\}$ and the shares of $A_i^{(\omega)}$ are: $(s_{i_1}^{(m)}, e_{i_1}^{(m)}), \Lambda, (s_{i_k}^{(m)}, e_{i_k}^{(m)})$ and $(s_{i(n+1)}^{(m)}, e_{i(n+1)}^{(m)})$. If these shares have been verified correct, $A_i^{(\omega)}$ could use $(i_1, s_{i_1}^{(m)}), \Lambda, (i_k, s_{i_k}^{(m)})$ and $(n+1, s_{i(n+1)}^{(m)})$ to reconstruct the shared secret s.

To construct a multinomial with degree of k:

$$I_k^{(m)}(z) = \sum_{\alpha=1}^{k} s_{i_\alpha}^{(m)} \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)} \qquad (13)$$

$$+ s_{i(n+1)}^{(m)} \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

Get $s = I_k^{(m)}(0)$.

For short, we call this process as "$A_i^{(\omega)}$ reconstruct the shared secret s".

3. CORRECTNESS PROOF

Suppose $A_i^{(\omega)}$ in the following lemmas as $A_i^{(\omega)} \in \Gamma_0$ and $A_i^{(\omega)} = \{p_{i_1}, p_{i_2}, \Lambda, p_{i_k}\}$; and "the normal conditions" in the following lemmas means that all the members in $A_i^{(\omega)}$ have received safely the Increase Shares sent by all dealers and they can show their correct shares when they want to cooperate to complete one job.

Lemma 1. Under the normal conditions, any $A_i^{(\omega)}$ can reconstruct the shared secret s by the shares ($s_{i_\alpha}^{(0)}$, $\alpha = 1, \Lambda, k$) and virtual shares ($s_{i(n+1)}^{(0)}$) distributed by each participant executing Shares Distributing Algorithm.

Proof.

Use $A_i^{(\omega)}$ to recover the shared secret. By equation (13),

$$I_k^{(0)}(z) = \sum_{\alpha=1}^{k} s_{i_\alpha}^{(0)} \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z-i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha-(n+1)} + s_{i(n+1)}^{(0)} \prod_{\beta=1}^{k} \frac{z-i_\beta}{n+1-i_\beta}$$

By equation(2)、equation(3),

$$I_k^{(0)}(z) = \sum_{\alpha=1}^{k} \left(\sum_{o=1}^{n} u_o^{(0)}(i_\alpha)\right) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z-i_\beta}{i_\alpha-i_\beta} \cdot \frac{z-(n+1)}{i_\alpha-(n+1)}$$

$$+\left(\sum_{o=1}^{n} I_{ko}^{(0)}(n+1)\right) \prod_{\beta=1}^{k} \frac{z-i_\beta}{n+1-i_\beta}$$

$$= \sum_{o=1}^{n} \left(\sum_{\alpha=1}^{k} u_o^{(0)}(i_\alpha)\right) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z-i_\beta}{i_\alpha-i_\beta} \cdot \frac{z-(n+1)}{i_\alpha-(n+1)}$$

$$+\left(\sum_{o=1}^{n} I_{ko}^{(0)}(n+1)\right) \prod_{\beta=1}^{k} \frac{z-i_\beta}{n+1-i_\beta}$$

$$= \sum_{o=1}^{n} \{ \left(\sum_{\alpha=1}^{k} u_o^{(0)}(i_\alpha)\right) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z-i_\beta}{i_\alpha-i_\beta} \cdot \frac{z-(n+1)}{i_\alpha-(n+1)}$$

$$+ I_{ko}^{(0)}(n+1) \prod_{\beta=1}^{k} \frac{z-i_\beta}{n+1-i_\beta} \}$$

By Shares Distributing Algorithm and $RS(d, P, func, t, r, a, b, c)$, $I_{ko}^{(0)}(n+1)$ in the above equation is got from the multinomial $I_{ko}^{(0)}(z)$ with k degrees constructed by p_o using $(0, x_o)$ and $(i_\alpha, u_o^{(0)}(i_\alpha))$ ($\alpha = 1, \Lambda, k$). Whereas, using $(n+1, I_{ko}^{(0)}(n+1))$ and $(i_\alpha, u_o^{(0)}(i_\alpha))$ ($\alpha = 1, \Lambda, k$) to construct multinomial with k degrees:

$$I_{ko}^{(0)'}(z) = \sum_{\alpha=1}^{k} u_o^{(0)}(i_\alpha) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z-i_\beta}{i_\alpha-i_\beta} \cdot \frac{z-(n+1)}{i_\alpha-(n+1)} \quad , \quad \text{make}$$

$$+ I_{ko}^{(0)}(n+1) \prod_{\beta=1}^{k} \frac{z-i_\beta}{n+1-i_\beta}$$

$z = 0$, and $I_{ko}^{(0)'}(0) = x_o$ can be gotten.

So, $I_k^{(0)}(z) = \sum_{o=1}^{n} I_{ko}^{(0)'}(z)$, $I_k^{(0)}(0) = \sum_{o=1}^{n} I_{ko}^{(0)'}(0) = \sum_{o=1}^{n} x_o = s$.

Proof ends.

Lemma 2. If $I_k^{(m)}(z)$ (in equation （13）) means the multinomial, by which $A_i^{(\omega)}$ reconstruct the shared secret s in Period m, and $I_{ko}^{(m)}(z)$ (in step (4) of $RS(d, P, func, t, r, a, b, c)$ algorithm) means the multinomial, by which p_o renews the virtual share of $A_i^{(\omega)}$, the following equation will be correct when $z = 0, i_1, \Lambda, i_k, n+1$, $m \geq 1$:

$$I_k^{(m)}(z) = I_k^{(m-1)}(z) + \sum_{o=1}^{n} I_{ko}^{(m)}(z) \quad , \quad \text{and} \quad I_{ko}^{(m)}(0) = 0$$

$(o = 1, \Lambda, n)$.

Proof.

By equation (13),

$$I_k^{(m)}(z) = \sum_{\alpha=1}^{k} s_{i_\alpha}^{(m)} \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z-i_\beta}{i_\alpha-i_\beta} \cdot \frac{z-(n+1)}{i_\alpha-(n+1)} + s_{i(n+1)}^{(m)} \prod_{\beta=1}^{k} \frac{z-i_\beta}{n+1-i_\beta} .$$

By equation(4)、equation (5),

$$I_k^{(m)}(z) = \sum_{\alpha=1}^{k} \left(s_{i_\alpha}^{(m-1)} + \sum_{o=1}^{n} u_o^{(m)}(i_\alpha)\right) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z-i_\beta}{i_\alpha-i_\beta} \cdot \frac{z-(n+1)}{i_\alpha-(n+1)}$$

$$+ \left(s_{i(n+1)}^{(m-1)} + \sum_{o=1}^{n} I_{ko}^{(m)}(n+1)\right) \prod_{\beta=1}^{k} \frac{z-i_\beta}{n+1-i_\beta}$$

$$= \{ \sum_{\alpha=1}^{k} s_{i_\alpha}^{(m-1)} \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)} + s_{i(n+1)}^{(m-1)} \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta} \}$$

$$+ \sum_{\alpha=1}^{k} (\sum_{o=1}^{n} u_o^{(m)}(i_\alpha)) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}$$

$$+ (\sum_{o=1}^{n} I_{ko}^{(m)}(n+1)) \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

$$= I_k^{(m-1)}(z) +$$

$$\sum_{o=1}^{n} (\sum_{\alpha=1}^{k} u_o^{(m)}(i_\alpha)) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}$$

$$+ (\sum_{o=1}^{n} I_{ko}^{(m)}(n+1)) \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

$$= I_k^{(m-1)}(z) +$$

$$\sum_{o=1}^{n} \{(\sum_{\alpha=1}^{k} u_o^{(m)}(i_\alpha)) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}$$

$$+ I_{ko}^{(m)}(n+1) \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta} \}$$

$I_{ko}^{(m)}(n+1)$ is equal to $I_{ko}^{(m)}(z)$ when $z = n+1$. By step (4) of $RS(d,P,func,t,r,a,b,c)$ and the Minimal Authorized Subsets Renewing Algorithm, $I_{ko}^{(m)}(z)$ is constructed by points $(0,0)$ and $(i_1,u_o^{(m)}(i_1)),(i_2,u_o^{(m)}(i_2)),\Lambda ,(i_k,u_o^{(m)}(i_k))$. So, 0 is equal to the multinomial

$$I_{ko}^{(m)'}(z) = \sum_{\alpha=1}^{k} u_o^{(m)}(i_\alpha) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)} + ,$$

$$I_{ko}^{(m)}(n+1) \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

Constructed by points $(n+1, I_{ko}^{(m)}(n+1))$ and $(i_1,u_o^{(m)}(i_1)),(i_2,u_o^{(m)}(i_2)),\Lambda ,(i_k,u_o^{(m)}(i_k))$ when $z=0$. That is to say, $I_{ko}^{(m)}(z)$ and $I_{ko}^{(m)'}(z)$ are equal to each other in the points $(0,0)$, $(i_1,u_o^{(m)}(i_1))$, $\cdots$, $(i_k,u_o^{(m)}(i_k))$, $(n+1,I_{ko}^{(m)}(n+1))$. That is to say, when $z = r^{(m)}, i_1, \Lambda , i_k, n+1$,

$I_k^{(m)}(z) = I_k^{(m-1)}(z) + \sum_{o=1}^{n} I_{ko}^{(m)}(z)$. *Proof ends.*

Lemma 3. After all the participants of the system has renewed all the shares of the system for $t\,(t \geq 1)$ times, any minimal authorized subset $A_i^{(\omega)}$ can reconstruct the shared secret S.

Proof.

Just to prove: when the Period is t, the value of the multinomial $I_k^{(t)}(z)$ with k degree constructed by $A_i^{(\omega)}$ is S when $z=0$. By Lemma 2, $I_k^{(t)}(0) = I_k^{(t-1)}(0) + \sum_{o=1}^{n} I_{ko}^{(t)}(0) = I_k^{(t-1)}(0)$, we can get: $I_k^{(t)}(0) = I_k^{(t-1)}(0) = \Lambda = I_k^{(0)}(0) = s$.

Proof ends.

Lemma 4. Use the Minimal Authorized Subset Renewing Algorithm and renew any minimal authorized subset $A_i^{(\omega)} \in \Gamma_0$ to $A_i^{(\omega+1)}$. There will be : $A_i^{(\omega+1)} \in \Gamma_0$, and $A_i^{(\omega+1)}$ can reconstruct the shared secret S.

Proof.

(1) By the method renewing the members of $A_i^{(\omega)}$ in the Minimal Authorized Subset Renewing Algorithm, the members of $A_i^{(\omega+1)}$ are still in accord with the definition in table 2. So, $A_i^{(\omega+1)} \in \Gamma_0$.

(2) Now, we prove $A_i^{(\omega+1)}$ can reconstruct the shared secret S.

Still suppose $A_i^{(\omega+1)} = \{ p_{i_1}, \Lambda , p_{i_k})$, the creditable minimal authorized subset is $A_j^{(\omega)} = \{ p_{j_1}, \Lambda , p_{j_\theta} \}$ and the current Period is m.

To prove $A_i^{(\omega+1)}$ could reconstruct the shared secret S is to prove the value of

$$I_k^{(m)}(z) = \sum_{\alpha=1}^{k} s_{i_\alpha}^{(m)} \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}$$

$$+ s_{i(n+1)}^{(m)} \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

is equal to S when $z=0$.

At first, because $A_j^{(\omega)}$ can reconstruct the shared secret S, we can get

$$s = \sum_{\alpha=1}^{\theta} s_{j_\alpha}^{(m)} \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{\theta} \frac{-j_\beta}{j_\alpha - j_\beta} \cdot \frac{-(n+1)}{j_\alpha - (n+1)} + \tag{14}$$

$$s_{j(n+1)}^{(m)} \prod_{\beta=1}^{\theta} \frac{-j_\beta}{n+1-j_\beta}$$

By equation (6), $s_{i_\alpha}^{(m)} = s_{i_\alpha}^{'} - \sum_{\delta=1}^{\theta} u_{j_\delta}(i_\alpha)$; by equation (7) and equation (10),

$s_{i(n+1)}^{(m)} = s_{i(n+1)}^{(m)'} - \sum_{\delta=1}^{\theta} \mathrm{I}_{kj_\delta}(n+1)$. So we can get:

$$I_k^{(m)}(z) = \sum_{\alpha=1}^{k} (s_{i_\alpha}^{'} - \sum_{\delta=1}^{\theta} u_{j_\delta}(i_\alpha)) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}$$

$$+ (s_{i(n+1)}^{(m)'} - \sum_{\delta=1}^{\theta} \mathrm{I}_{kj_\delta}(n+1)) \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

$$= \{ \sum_{\alpha=1}^{k} s_{i_\alpha}^{'} \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)} + s_{i(n+1)}^{(m)'} \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta} \} -$$

$$\sum_{\delta=1}^{\theta} \{ (\sum_{\alpha=1}^{k} u_{j_\delta}(i_\alpha)) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}$$

$$+ \mathrm{I}_{kj_\delta}(n+1) \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta} \} \tag{15}$$

By equation (9), we can know that $s_{i(n+1)}^{(m)'}$ is equal to the value of the multinomial $L_k(z)$ constructed by points $(i_\alpha, s'_{i_\alpha})(\alpha = 1,\Lambda, k)$ and $(0, s')$ when $z = n+1$. So, the value of the multinomial $I'_k(z)$ constructed by points $(n+1, s'_{i(n+1)})$and $(i_\alpha, s'_{i_\alpha})(\alpha = 1,\Lambda, k)$ is equal to s' when $z = 0$. $I'_k(z)$ is:

$$I'_k(z) = \sum_{\alpha=1}^{k} s'_{i_\alpha} (\prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}) \tag{16}$$

$$+ s_{i(n+1)}^{(m)'} \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

By $RS(d,P,func,t,r,a,b,c)$ algorithm and Step1.1 in the Minimal Authorized Subset Renewing Algorithm, we know that $\mathrm{I}_{kj_\delta}(n+1)$ is equal to the multinomial $\mathrm{I}_{kj_\delta}(z)$ constructed by $(0, u_{j_\delta}(0))$ and $(i_1, u_{j_\delta}(i_1)),\Lambda, (i_k, u_{j_\delta}(i_k))$ when $z = n+1$. So, we know that $u_{j_\delta}(0)$ is equal to the value of the multinomial $I'_{kj_\delta}(z)$ constructed by points $(\mathrm{n}+1, \mathrm{I}_{kj_\delta}(n+1))$ and $(i_1, u_{j_\delta}(i_1)),\Lambda, (i_k, u_{j_\delta}(i_k))$ when $z = 0$. $I'_{kj_\delta}(z)$ is

$$I'_{kj_\delta}(z) = \sum_{\alpha=1}^{k} u_{j_\delta}(i_\alpha) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{k} \frac{z - i_\beta}{i_\alpha - i_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)} \tag{17}$$

$$+ I_{kj_\delta}(n+1) \prod_{\beta=1}^{k} \frac{z - i_\beta}{n+1-i_\beta}$$

In the same way, we can get：

$$J'_{\theta j_\delta}(z) = \sum_{\alpha=1}^{\theta} u_{j_\delta}(j_\alpha) \prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{\theta} \frac{z - j_\beta}{j_\alpha - j_\beta} \cdot \frac{z-(n+1)}{i_\alpha - (n+1)}$$

$$+ J_{\theta j_\delta}(n+1) \prod_{\beta=1}^{\theta} \frac{z - j_\beta}{n+1-j_\beta} \tag{18}$$

And $J'_{\theta j_\delta}(0) = u_{j_\delta}(0)$.

By equation (15)、(16) and (17),

$$I_k^{(m)}(z) = I'_k(z) - \sum_{\delta=1}^{\theta} I'_{kj_\delta}(z),$$

$$I_k^{(m)}(0) = I'_k(0) - \sum_{\delta=1}^{\theta} I'_{kj_\delta}(0) = s' - \sum_{\delta=1}^{\theta} u_{j_\delta}(0) \tag{19}$$

By equation (8)、(9)、(10),

$$s' = \sum_{\alpha=1}^{\theta} (s_{j_\alpha}^{(m)} + \sum_{\delta=1}^{\theta} u_{j_\delta}(j_\alpha))(\prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{\theta} \frac{-j_\beta}{j_\alpha - j_\beta} \cdot \frac{-(n+1)}{j_\alpha - (n+1)}) +$$

$$(s_{j(n+1)}^{(m)} + \sum_{\delta=1}^{\theta} J_{\theta j_\delta}(n+1)) \prod_{\beta=1}^{\theta} \frac{-j_\beta}{n+1-j_\beta}$$

$$= \{\sum_{\alpha=1}^{\theta} s_{j_\alpha}^{(m)} (\prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{\theta} \frac{-j_\beta}{j_\alpha - j_\beta} \cdot \frac{-(n+1)}{j_\alpha - (n+1)}) + s_{j(n+1)}^{(m)} \prod_{\beta=1}^{\theta} \frac{-j_\beta}{n+1-j_\beta}\} +$$

$$\sum_{\delta=1}^{\theta} \{\sum_{\alpha=1}^{\theta} u_{j_\delta}(j_\alpha))(\prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{\theta} \frac{-j_\beta}{j_\alpha - j_\beta} \cdot \frac{-(n+1)}{j_\alpha - (n+1)})$$

$$+ J_{\theta j_\delta}(n+1) \prod_{\beta=1}^{\theta} \frac{-j_\beta}{n+1-j_\beta}\}$$

By equation (14)、(18), the above equation can be changed into:

$$s' = s + \sum_{\delta=1}^{\theta} u_{j_\delta}(0). \tag{20}$$

By equation (19)、(20), to know:

$$I_k^{(m)}(0) = s' - \sum_{\delta=1}^{\theta} u_{j_\delta}(0) = s + \sum_{\delta=1}^{\theta} u_{j_\delta}(0) - \sum_{\delta=1}^{\theta} u_{j_\delta}(0) = s$$

Proof Ends.

The above lemmas can prove that our General Dynamic Secret Sharing Algorithm no only can renew the shares and the minimal authorized subsets, but also can recover the shared secret correctly.

4. SECURITY ANALYSIS

(1) The suppose of the type of adversary and the difficulty in computation

By the character of our algorithm, we suppose there be two kind of adversary. The first one is static and strong admissible. Static means the adversary can bribe some participants before the algorithm begins; Strong admissible means that the number of the participants that can be bribed is no more than the minimal authorized subset, where the number of the members is the least. The second one is dynamic and strong admissible. Dynamic means that the adversary can confirm which participants might be bribed in accordance with the minimal authorized subsets that have been formed recently. Strong admissible means that adversary can bribe any participant but there is at least one member he can't bribe in all minimal authorized subsets in one Period and there are at least 2 minimal authorized subsets, in which none of the members can be bribed in one period.

We use a basic computing difficulty suppose, which means it is impossible to anyone to compute the discrete logarithm whose bottom is g in Z_q.

(2) Security

Lemma 5. For the first and second kind of adversary, our General Dynamic Secret Sharing is robust, that is to say the adversary can't recover the secret and can't prevent the credible minimal authorized subsets from recovering the shared secret S; has strong surveillance, that is to say the cheating behavior of the adversary and other wrong all can be monitored in time.

Proof. About the first kind of adversary. Because we just define every minimal authorized subset and don’t give the members in every minimal authorized subset. So, at first, the adversary can’t know which members can be combined to reconstruct the shared secret. Even if all the members that the adversary have bribed are all in the same minimal authorized subset, he can’t recover the shared secret for there is at lease one member in the minimal authorized subsets he hasn’t bribed.

About the second kind of adversary. Because there is at least one member P_{i_t} in all minimal authorized subsets he can’t bribe, the adversary can’t recover the shared secret in one period. Even if the adversary could bribe P_{i_t} after on period, he couldn’t recover the shared secret because the share of P_{i_t} has been renewed to the share of the next period.

In addition, because there is at least two credible minimal authorized subsets in one period, the adversary can’t prevent this kind of minimal authorized subsets recover the shared secret.

The system verifies the information sent by all the participants and can find the cheating behaviors of the dishonest participants by executing the Shares Renewing Algorithm and the Minimal Authorized Subsets Renewing Algorithm periodically. The strong and powerful surveillance can prove our General Dynamic Secret Sharing Algorithm is robust and secure. *Proof ends.*

By the above lemma and the unconditional security of ZFT_GVSS[6], we can come into conclusion that the information can be gotten by the adversary is indifferent with the shared secret S and our General Dynamic Secret Sharing algorithm is unconditionally secure.

5. CONCLUSIONS

With a view to “general” and “dynamic”, this paper has dept into discussing and studying the secret sharing algorithm and tries to make it more applicable. The members that can be combined to recover the shared secret must be announced before the algorithm begins and can’t be changed again in the threshold scheme and in ZFT_GVSS scheme. The General Dynamic Secret Sharing algorithm has broken the confine of the two schemes. We take the method of renewing the shares and the minimal authorized subsets periodically and make the adversary have no time to judge which participants can be combined to recover the shared secret in one period. By this way the security of the algorithm can be proved strongly and powerfully. Because of the character of our algorithm in security and complexity of computing, our algorithm may be used in private key management, group oriented cryptography, military affairs and electronic commerce.

6. REFERENCES

[1] T Pedersen. Non-interactive and information-theoretic secure verifiable secret sharing. Advances in Cryptology-Crypto’91. Berlin, Springer-Verlag，1991. 129～140.

[2] A Herzberg, S L Jarecki, H Krawczyk et al. Proactive secret sharing or: How to cope with perpetual leakage. Advances in Cryptology-Crypto’95. Berlin, Springer-Verlag, 1995.339～352.

[3] TENG Meng, ZOU Peng, and WANG Huai-Min. A Proactive Secret Sharing Algorithm. Journal of Computer Research and Development,2003,40(7):1008～1015.

[4] P Feldman. A practical scheme for non-interactive verifiable secret sharing. Proc of 28th Annual IEEE Symposium on Foundations of Computer Science. New York, IEEE Computer Society Press, 1987. 427～437.

[5] R Gennaro. Theory and practice of verifiable secret sharing [PhD dissertation]. Massachusetts Institute of Technology (MIT), Cambridge, 1996.

[6] ZHANG Fu-Tai and WANG Yu-Min. AN Unconditional Secure General Verifiable Secret Sharing Protocol. Journal of Computer Research and Development , 2002, 39(10): 1199～1204.

[7] ZHANG Fu-Tai and WANG Yu-Min. Distributed Key Generation Based on Generalized Verifiable Secret Sharing. Journal of Electronic Transactions, 2003, 31(4): 580~584.

Li Xiaoxin is a graduate student in School of Computer Science and Technology, Wuhan University of Technology。And his research direction is network security and he is also interested in Electronic Commerce and the skills of middle ware.

Guo Qingping is a Full Professor and a head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He is one of the DCABES international conference founder, was the chairman of DCABES 2001, co-chair of DCABES 2002, and will be the chairman of DCABES 2004. 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-commence.

Zhang Feng, male, is a master candidate in School of Computer Science and Technology, Wuhan University of Technology. His research direction is computer network security.

Central 3A Platform Based-on SSO

Zhang Yingjiang[1,2], Li Jun[2], Zheng Qiuhua[2], Li Layuan[1]
[1] Wuhan University of Technology [2] Hubei University of Technology
Wuhan, Hubei, 430068, China
Email: yjjzhang@hubpu.edu.cn **Tel.:** +86-27-88034021

ABSTRACT

This paper discusses the necessity of using single sign-on technology and summarizes the key technology of single sign-on. The design and implementation of a central 3A platform based on SSO called SSOPortal has been reported.

Keywords: Single sign-on; Central; 3A Platform; SSOPortal

1. INTRODUCTION

As the popularity of applications based on computer networks increases, users often log on many different systems everyday. Generally, each application has it's own authentication/authorization system and maintains it's own independent log. The disadvantages of this kind of security solution become obvious. First, it increases the costs of developing the application and the costs of managing the application. Second, because each application requires the user to obey some safety tactics arranged, i.e. entering user ID and password, the more applications user loges on, the more possibility of mistakes made by the user. If users forget their password, they can't login, they have to ask for help from administrator. They must wait a long time until getting the password again. To avoid this embarrass situation, users usually use simple password or same password in all of different systems instead, or write passwords down on notes. These customary actions leave the systems vulnerable for internal attacks. It will reduce the system's security also. So, how to provide users a convenient authentication and authorization method becomes the hottest issue that we need thinking over in recent years.

To resolve the above problems, Authentication/ authorization /administration/auditing service, as an infrastructure of enterprise network must separates from applications. Users need only be checked once and they can access to all authorized systems (include C/S and Web) and other resource freely. This solution will improve system efficiency, reduce cost, and enhance system security. The idea is Single Sign-On (SSO)[1]. Based on SSO, to implement Authentication /authorization /auditing service in a central platform is called 3A Platform Based-on SSO.

According to the survey made by IDC, in the near future years，the increase part of security product will be share by Security 3A(Administration, Authorization, Authentication), Encryption, Firewall/VPN, Content Security, IDS etc. Among them, Administration, Authorization, Authentication (3A) has been the hottest technique. In the recent years, Security 3A software has accounted for 50% worldwide market shares in security products, and the growth rate of 3A products has exceeded the traditional security products (See Fig. 1). Security 3A software has been the technical drive force on business for enterprise. The meaning of "security" changes from insurance to Security 3A. Security 3A application has been the focus of clients, partners, vendors.[2]

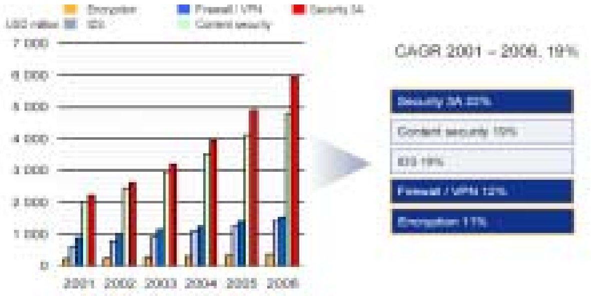


Fig. 1 Network Security Product Shares

2 SINGLE SIGN ON MODELS AND ITS TECHNIQUES

SSO implementation still hasn't a standard till now. Several different models of SSO are introduced as follows [3].

2.1 Broker-Based SSO

In a Broker-Based SSO solution, there exists one server for central authentication and user account management. The broker gives an electronic identity that can be used for requesting further access. The central database reduces administrative overhead and provides a common, single place for authentication. The advantage of a broker-based SSO solution is that one central database makes the management easily. The main problem of broker-based solution is that the end applications need to be modified.

2.2 Agent-Based SSO

In an Agent-Based solution there exists an agent program that automatically identifies the user for different applications. The agent can be designed in different ways. It can also be placed on the server side to act as an "interpreter" between the authentication system of the server and the authentication method used by the client. This solution makes the migration easier, as the software vendor supplies different agents that are designed to communicate with the legacy applications

2.3 Gateway-Based SSO

In this solution, requests of accessing protected resource and network services must pass through a "door"—gateway. A model of the solution is presented in Fig2.

The Gateway could be a firewall or a dedicated cryptographic server. In the solution, all the requested services need to reside behind the gateway, in a trusted network segment. The clients authenticate themselves cryptographically to the gateway that grants access to the services. As the services behind the gateway can be recognized based on their IP addresses, a rule base based on IP addresses can be built on the gateway. When this rule base is combined with a user database residing in the gateway, the gateway can be used for Single Sign-On. As the

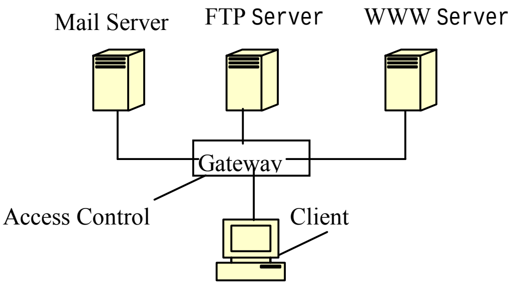


Fig. 2 Model of a Gateway-based SSO

gateway has the possibility to access all the traffic flow to the services, it can monitor and alter the data flow to the services. Thus it can replace the authentication information going to the services, making it suitable for access control without modification of the applications themselves. But if several gateways are used and the databases cannot be synchronized automatically, it will bring problems. If the client uses a proxy to access system services, it also can't provide based-user control.

2.4 Agent & Broker based SSO

Agent and Broker-Based SSO Solution is a mixed way to implement SSO, in which the agent-based solution is combined with a broker-based solution. In this Solution, both advantages of the central management of the broker-based solution and the flexibility of agent-based solution can be achieved. A graphic of the model is shown in Fig. 3

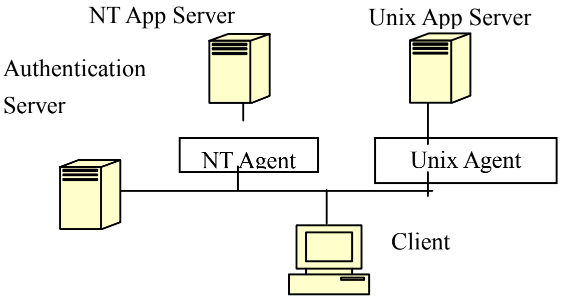


Fig. 3 Model of Agent and Broker-Based SSO

2.5 SSO Models Summary

After analyzing the above models, we can draw a conclusion, that is, Agent & Broker-based SSO is the most suitable one for implementing SSO. It can make full use of Broker-based SSO's concentrated management technique and create a center database to manage user's ID. Also it can utilize Agent-base SSO's flexibility without modification within the applications.

However, Agent & Broker based SSO solution still has some flaws.

- In Agent-based SSO, a central user database is needed. We have to immigrate the current user databases to the center database. This will lead to raise the cost.
- It is difficult to manage the center database for those kinds of corporations that have many branches distributed in different areas. For example, when an employee resigned from a corporation, the center database administrator might keep his or her account till the administrator has been asked to delete it. This will bring system the latent hole of security.
- Another flaw is that this solution only consider authentication.

3. SSOPORTAL

Based on Agent-based SSO solution, combining with other SSO technologies, we have designed a Single Sign On system — SSOPortal. SSOPortal provides central authentication services, central authorization services, policy-based user management and central account services. In addition, it uses a logical central database to implement system's central database.

The model of SSOPortal is illustrated in Fig. 4.

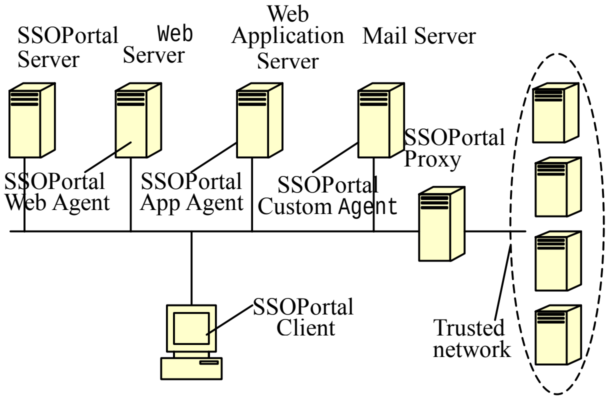


Fig. 4 The Model of SSOPortal

3.1 SSOPortal Components

The SSOPortal consists of three main components: the SSOPortal Server, the SSOPortal Agent and the SSOPortal Client.

The SSOPortal Server is played as a broker in the system that provides the following services:

- Central authentication
- Central authorization
- Policy-based user management
- Central accounting

The SSOPortal Server has two parts, prime module and secondary module. The prime module is the part of server whose implementation is through programming, and secondary module consists of some components that store system data, such as RDBMS, LDAP directories. The prime module includes authentication module, authorization module, account module, and administration module. The secondary module includes log database, user directories and policies directories.

The SSOPortal Agent is a middleware that integrates with Web applications and can communicate with SSOPortal server to protect system resources. Using an SSOPortal, the agent intercepts request for a resource, then checks whether the resource is protected or not. If a resource is unprotected, a user gains access without intervention. If the resource is protected, the SSOPortal Agent will talk with the SSOPortal server to authenticate the user and to determine whether he or she has the rights to access the resource. When an authorization is successful, the SSOPortal Agent proceeds with request to forward the request to Web Server.

The SSOPortal Client is the entrance that user accesses resource. The design is totally different for B/S application and C/S application. If user attempts to access B/S application, the client only requires a browser that supports cookie, SSL and JavaScript. If he or she attempts to access C/S application, the client needs to install an additional SSO application in client side.

3.2 How SSOPortal works

3.2.1 Procedure of accessing B/S based application or Web resources

When a user attempts to access a B/S based application or Web resources, SSOPortal processes the request as described in Fig. 5.

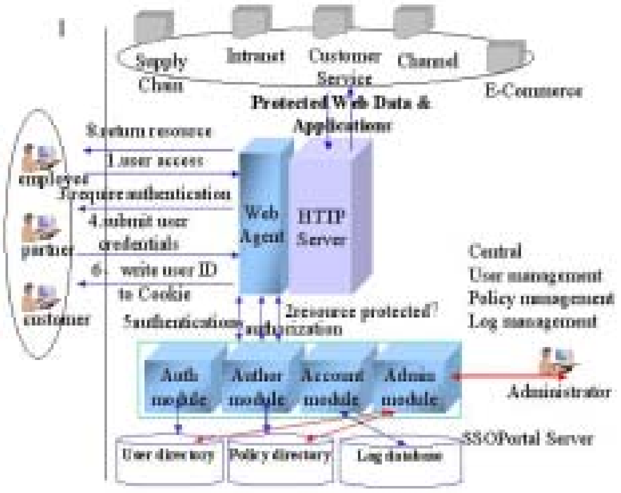


Fig.5 Procedure of Accessing B/S Based Application or Web Rresources

The steps are:

Step 1: User posts request of accessing resource by browser.

Step 2: Web agent intercepts the HTTP request to check access URL. Web agent interacts with SSOPortal server to determinate whether the resource should be protected.

Step 3: The authorization module of SSOPortal Server queries policy directories and returns the result to the agent.

Step 4: If the resource is unprotected, the SSOPortal Server instructs the Web Agent to forward user's request to Web Server.

Step 5: If the resource is protected, the SSOPortal Server instructs the Web Agent to check HTTP headers. If Web Agent can't find user identifies in cookie of HTTP headers, the Web Agent will redirect the authentication form to ask user for authentication. If Web Agent finds user's identifies, then go to step 10.

Step 6: After user posts creditable message, the Web Agent sends request to authentication module of SSOPortal Server for authentication.

Step 7: The authentication module looks LDAP user directories to determine if the user is legal and to notify the answer to the Web Agent.

Step 8: If the user is successful authenticated, the Web Agent writes a "set cookie" instruction. [4] The browser will write cookie to local. After that, all HTTP request for the domain will carry the cookie.

Step 9: If the user is failed to authenticate, the Web Agent will block the HTTP request and returns an error document as well as to send the log record to account module of SSOPortal Server for logging the event.

Step 10: After the Web Agent finds user's identifies in HTTP headers, it asks authorization module of SSOPortal Server for authorization.

Step 11: The authorization module checks LDAP policy directories to determine whether the access is allowed and then to send its decision to the Agent.

Step 12: If user is authorized, the Web Agent will forward the user's request to Web Server.

Step 13: If user is denied for accessing the resource, the Web Agent sends an error HTTP response.

3.2.2 Procedure of accessing C/S based application or Web resources

When a user attempts to access a C/S based application or Web resources, SSOPortal processes the request as described in below.

Step 1: User starts SSOPortal Client application to log on SSOPortal Server.

Step 2: The authentication module of SSOPortal Server verifies user and returns a ticket to the client.

Step 3: SSOPortal Client writes the ticket to local cookie so that it can be used for accessing C/S application.

Step 4: SSOPortal Client asks authorization module of SSOPortal Server for authorization.

Step 5: Authorization module sends a list of authorized application, scripts for auto-logon and a table of user credentials to SSOPortal Client.

Step 6: According to these data, SSOPortal Client displays these authorized application in a GUI interface with icon form.

Step 7: After that, when the user attempts to access a C/S application, what he or she only need to do is double click the icon. SSOPortal Client will retrieve user credentials from user credentials table and login the C/S application by script for auto-logon.

Because user credentials table includes some user login information such as password, it can't be stored by plain-text form [5]. SSOPortal encrypts these kind of data before storing them in policy directories, and decrypts them in user computer. The secret key for encryption and decryption is user's SSOPortal password. Through it, SSOPortal can resolve the latent secure problem. In additional, the user credentials between policy directories and C/S application must be synchronized

.

3.3 Authentication

In SSO based 3A systems, users need only be checked once and they can access to all authorized systems (include C/S and Web) and other resources freely. In these systems, authentication plays a key role. If authentication is not powerful, when hacker intrudes the system at one point, the whole system will be exposure completely. Thus, SSO based 3A systems should provide a strong authentication mechanism which makes the resource more safety, and user's access more convenient. Authentication module of SSOPortal protects the system resource by the method that returns a value of authentication level. Depending on the authentication level value, authorization module authorizes users different privilege. Authentication level is defined as the degree of trusted in user identity after user successful authentication and is determined by the security of authentication mechanism, the decrypt method in authentication mechanism.

SSOPortal provides the following authentication services:

- User name/Password
- One-Time Password
- X.509 Certificate
- X.509 Certificate and User name/Password

Here is the SSOPortal authentication level Table 1:

Table 1

Authentication method	Authentication level
User name/Password	1
One-Time Password	10
X.509 Certificate	15
X.509 Certificate and User name/Password	20

Considering for scalable, the values of authentication level are set at discrete. System administrator can modify these values.

3.4 Authorization

SSOPortal authorizes users by means of two methods. One is authorizing based-on role, which is to grants users different privileges according to the roles of users. A user can be multi-role. Another is authorizing based-on resources, that is, SSOPortal can bind resources to policies.

3.4.1 Policy objects

SSOPortal has several policy objects: domain, agent, directory, auth_scheme, admin, realm, rule, policy. In a like-BNF manner, these objects can be described as following:

- <Domain> ::= <Domain_Id> <Description > [<Dir_Id>….] [<Admin_Id>…] [<Realm_Id>…]

- <Agent> ::= <Agent_Id> <Description> <Agent_Type> <Ip_Addr> <Shared_Secret>

- <Directory> ::= <Dir_Id> <Description> <Dir_Setup> <Usr> <Pwd> [<Universal_Id> <Disabled> <Pwd_Attr> <Pwd_Data> <Email>]

- <Auth_Scheme> ::= <Auth_Scheme_Id> <Description> <Auth_Level> <…>

- <Admin> ::= <Admin_Id> <Description> <Dir_Id> <Auth_Scheme_Id> <System | [<Domain_Id>…]> <Tasks>

- <Realm> ::= <Realm_Id> <Description> <Agent Id> <Resource_Filter> <Auth_Scheme_Id> <Protected> <MaxTimeoutEnabled><MaxTimeout><IdleTimeoutEnabled > <IdleTimeout> <Synchr_Auditing> <Dir_Id> <Events>

- <Rule> ::= <Rule_Id> <Description> <Realm_Id> <Resource> <Reg_Match> <Action> <Allowed> <Enabled> <Time_Limit> <Ip_Limit>

- <Policy> ::= <Rule> <User> <Response> <Ip_Limit> <Time_Limit>

3.4.2 Authorization Process

The SSOPortal authorization process brings the components of SSOPortal together. Authorizing a user for accessing requires the Authorization module to determine which policies rule the trigger when a user attempts to access a particular resource.

The Authorization module performs two primary functions in the following order:

- **To determine whether a resource is protected**

The Steps are:

Step 1: The Agent sends the details of the HTTP request to the Authorization Module.

Step 2: The Authorization Module gets agent's ID by agent's IP.

Step 3: The Authorization Module queries the set of realm by agent's ID. The Authorization Module looks up the realm that matches longest the requested path to a resource.

Step 4: If the value of "protected" field of realm record is true, the Authorization Module returns a response to the SSOPortal Agent, which includes a flag to indicate resource's status (protected or unprotected) and realm's ID, authentication scheme's ID, authentication level, etc. and then go to step 6.

Step 5: If the value of "protected" field is false, the Authorization Module sends a response to the SSOPortal Agent, which indicates that resource is unprotected. The SSOPortal Agent forwards user's request to web server.

Step 6: If resource is protected, the SSOPortal Agent determines whether the request event is identical with response's event field. If it is different, the SSOPortal Agent forwards user's request.

Step 7: If it is identical, the SSOPortal Agent checks the HTTP request to determine whether HTTP header's cookies include the user's identity.

Step 8: If user's request doesn't include cookies, the SSOPortal Agent authenticates user according to the fields of "Auth_Scheme_ID"、 "Auth_Level"、 "Dir_ID".

Step 9: If user's request includes user's identity, the SSOPortal Agent checks whether the user's identity is expired or not according to "MaxTimeoutEnabled"、 "MaxTimeout"、 "IdleTimeoutEnabled"、 "IdleTimeout" fields.

Step 10: If the identity is expired, the SSOPortal Agent authenticates user by “Auth_Scheme_Id”、 “Auth_Level”、 “Dir_Id” fields.

Step 11: If not, the SSOPortal Agent checks if user’s authentication level is greater than response’s.

Step 12: If greater, the SSOPortal Agent sends user’s identity and realm’s ID, action, resource as well as user’s IP to authorization module for authorization.

Step 13: If less, the SSOPortal Agent authenticates user by “Auth_Scheme_ID”、 “Auth_Level”、 “Dir_ID” fields.

● **To determine whether a user is authorized**
This procedure is illustrated in Fig. 6.

Here are the steps:
Step 1: The SSOPortal Agent sends user’s identity and realm’s ID, action, resource as well as user’s IP to authorization module for authorization.

Step 2: The SSOPortal Authorization Module sets a flag to false and retrieves policies that bind to user.

Step 3 If there are some policies bind to user, the SSOPortal Authorization Module gets a policy in this set. If not, the SSOPortal Authorization Module notifies the SSOPortal Agent to deny the access.

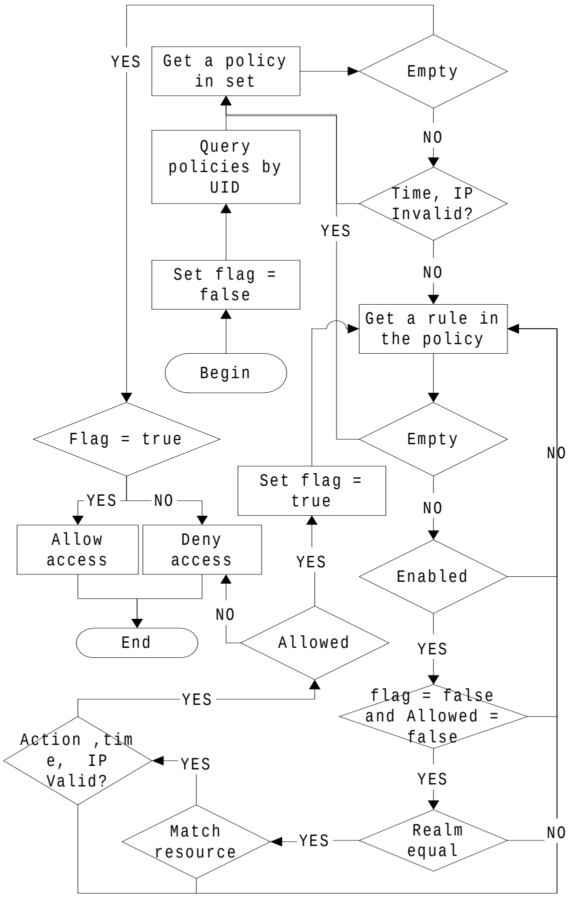


Fig. 6 Procedure of determining whether a user is authorized

Step 4: The SSOPortal Authorization Module checks the time restriction and IP restriction of this policy to determine whether policy is triggered.

Step 5: If policy doesn’t fire, the SSOPortal Authorization Module gets next policy in the set. If all policies have been checked, go to step 13.

Step 6: If the policy doesn’t contain any rule, go to step 5.

Step 7: If the rule is disabled, gets next rule in the set. If all rules have been checked, go to step 5.

Step 8: If flag is true or “Allowed” field is true, go to step 7.

Step 9: To check this rule by realm ID, if it is invalid, go to step 7.

Step 10: To check this rule by “Resource” and “Reg_Match” fields, if it is invalid, go to step 7.

Step 11: To check rule by time and IP constraint, if it is invalid, go to step 7.

Step 12: If “Allowed” value is false, the SSOPortal Authorization Module notifies the SSOPortal Agent to deny the access. If not, set flag to true, go to Step 7.

Step 13: If flag is true, SSOPortal allows the access to resource, else denies this request.

In SSOPortal, a deny access rule always takes precedence over an allow rule. This ability enables administrator to configure two different policies for resources in the same realm for different users. One policy allows certain users access, while the other denies a different group of users access .

3.5 Administration

SSOPortal’s architecture separates the system from policy domain management, so that each type of management can be performed by different administrators. By delegating management tasks, SSOPortal makes administration of large environments easier because those people in an organization who most familiar with a particular set resources and users can be assigned the privileges to manage them. In addition, it improves security by controlling who can create a modify users and policy objects. Depending on their role in an organization, SSOPortal administrators can have different privileges to manage SSOPortal. An administrator with maximum privileges can delegate the following management privileges to other managers:

- to create manage system and policy domain objects
- to manage users
- to manage keys
- to view and modify system reports

3.6 Auditing

SSOPortal can track user behavior and monitor site’s performance. SSOPortal audits all user activity, which includes all authentications and authorizations, as well as administrative activity, which includes any changes to policy stored. SSOPortal also track user sessions so administrator can monitor the resource being accessed, how often users attempt access, and how many uses are accessing the site.
SSOPortal can generate reports that include auditing

information about user activity, failed access attempts, and administrative changes. The types of reports are as follows:

● Activity reports—including information such as the type of resources that users access and how frequently they attempt to access, how many users are accessing particular resources, and whether access attempts were successful.

● Intrusion reports — including information about failed authentication and authorization attempts by a specific user, SSOPortal Agent, or both.

● Administrative reports—including administrative activity by a particular administrator or by the object. Administrative activity includes changes to policies and policy domain configurations, as well as management to user.

4. CONCLUSION

Compared to other SSO systems, SSOPortal has the following advantages:

● From desktop version SSO system to site-federation version SSO system, SSOPortal faces to enterprise system. With emphasis on conformity to enterprise resource, it provides central authentication services, central authorization services, central account services and policy management. SSOPortal is an enterprise platform for managing resource. It can secure access to resource, make the management easier, and reduce cost effectively.

● SSOPortal supports SSO not only on B/S application but also on C/S application. As to B/S application, SSOPortal even provides system-level authorization services and account services.

So far, SSOPortal has implemented resource-level central authorization services for B/S application. But for C/S application, it can only implement application-level authorization services. How to provide resource-level central authorization services for C/S application is still under further research.

5. ACKNOWLEDGMENTS

Authors would like to thank the participants in the Workshop on Computer Security for their helpful discussion and cooperation on this topic. This work was sponsored by HPPD.

6. REFERENCES

[1] Nancy Chamberlin. A Brief Overview of Single Sign–On Technology. Government Information Technology Issues 2000, A view to the Future: pp.3～7

[2] IDC. Worldwide Security 3As Software Market Forecast and Analysis, 2001-2005

[3] Jani Hursti. Single Sign-On, Proceedings of Helsinki University of Technology Seminar on Network Security 1997, Security of Corporate Networks: pp.4～8

[4] D.Kristol and L.Montuli, “HTTP State Management Mechanism,” RFC2109, Feb, 1997: pp.1～21

[5] Bruse Schneier. Applied in Cryptography: Protocols, Algorithms, and Source Code. China Machine Press, 2000-1-1

Zhang Yingjiang is a Full Professor and the president assistant of Hubei University of Technology. He obtained Master’s degree from Xi’an Jiaotong University in 1988. He was a visiting scholar of Vassa Polytechnic in 1996. He has published seven books and over 40 Journal papers. His research interests are in network security and management information system.

A New Security Solution of E-Commerce Based on Web Service

Xiang Yang[1,2] **Li Layuan**[1] **Hao Yulong**[3] **Shi Zhiqing**[2]
[1] **Department of Computer Science, Wuhan University of Technology, Wuhan 430063**
[2] **1802 mailbox 155#, Nanjing 210018**
[3]**PLA University of Science & Technology, Nanjing 210007**
Email: xynjwh@163.com jwtu@public.wh.hb.cn

ABSTRACT

This paper points out some issues about the security of the application systems based on web service by analyzing the technique architecture of web service, and proposes to build a safe electronic commerce by constructing a uniform security service layer, and do a deep research on how it is combined with PKI.

Keywords: Web Service, Security, PKI, Electronic Commerce, SOAP.

1. INTRODUCTION

Electronic commerce is a new commercial operation pattern, which in the open Internet environment, based on the browser and server application, can realize the on-line shopping, on-line trades and on-line electronic payment. Nowadays, the biggest problem to promote the electronic commerce is how to guarantee the security during the on-line trades. Because the Internet itself is open, and the on-line trades encounter various dangers, the demands for the security control should be taken into consideration. The demands for the electronic commerce security are mainly as the following:

(1). The trade information must be kept secret. The commerce information used in trade is required to keep secret. For example if the credit card's account number and user name are known to other person, the owner will face danger of embezzlement. If your rivals know the information of your ordering and payment, you may lose business opportunities. Therefore, generally during the information transfer of electronic commerce, encryption is required.

(2). The trader identity must be definite. Probably having never met with each other and separated from a long distance, both parties of on-line trade should be able to identify the counterparts' identity in order to trade successfully. As far as the sellers are concerned, they must ensure the buyers are not liars. Meanwhile, the buyers also worry if the stores on the net are the forgery. For this reason, the prerequisite for the trade is that both parties can confirm each other's identity conveniently and reliably.

(3). The trade contract can't be denied. Because the market conditions are ever changing, the trade contract can't be denied as long as it has once been reached. Otherwise, the benefit of one side will be injured inevitably.

(4). The trade document cannot be modified. If its content is changeable, the trade itself is just undependable, and so the consumers or the merchants may suffer loss due to it. For this reason, the document content of electronic trade also must be kept unchangeable so as to guarantee the trade serious and fair.

Since the emergence of XML (Extensible Markup Language), its excellent open standard makes all different systems able to exchange data with each other, and makes it become a trend to build electronic commerce system based on web service. Web service is a typical hierarchical distributed technique based on both XML and SOAP (Simple Object Access Protocol) [1]-[2]. A typical web service system will make full use of various different techniques, object models and programming language, which may include the simple Perl scripts or the web service programmed by C++ language or Java language, it also possible to have some complicated application based on the J2EE application server. With the compatibility and extensibility of XML and SOAP, web service can realize communication among different platforms [3]-[4]. Nevertheless, a certain cost should be paid: it means the security of these systems can hardly be guaranteed. It is also very difficult to find out a public security standard for all related techniques of the system.

The main problem, brought by use of the public security structure in web services, is that the security system structure has the typical distributed characteristic, and these system structures usually demand all the parts in them to realize the pivotal function and arithmetic. So no security system can ignore the regretful fact, that the security level of the whole system is the same as the security level of the weakest part of the whole system, namely the so-called" bottleneck effect" in the security system. Therefore, either we may avoid using some techniques, or we may make a compromise for the security of the whole system. Apparently, the ineffective activities will occur easily. Avoiding using some techniques is generally impractical, and also runs in opposite directions with the idea of web service (they can set up link among any technique). Therefore, this paper put forward a web service system structure based on uniform security.

2. THE UNIFORM SECURITY SERVICE SYSTEM STRUCTURE BASED ON WEB SERVICE

The main idea of the uniform security service system structure is to transfer the complexity of the security system structure to the uniform security service layer, so that others parts of the system need not undertake any security obligation.

In the uniform security service system structure, an application system based on web service will be able to be divided into many fields by various security levels and security requirements in the practical applications. In each field, the uniform security service will be realized by one or several servers, all the security arithmetic can be found in these servers and they are the exclusive discrimination places in these defined fields .For this reason, the uniform security discrimination and registration method boasts another advantage: Even if a user is faced with the interactions of a lot of different security elements in a particular field, he only needs to practice sign-on one time. The uniform security service itself, which maybe is a web service, can perform an outer wrapped function for the current security system structure, and provide much different security function such as discrimination, authorization, etc.

During an electronic commerce process, the user first applies for an ID by dispatching the uniform security service, which can identify him in a particular field. To acquire the ID, the user must provide the correct discrimination qualification information. This information has many various different forms: For example, it can be a simple user name and user password or certificate, yet it is also possible to take other forms. The uniform security service will verify the qualification information of the users by its internal security system structure, and grant the user an ID.A concrete security process is shown in figure 1:

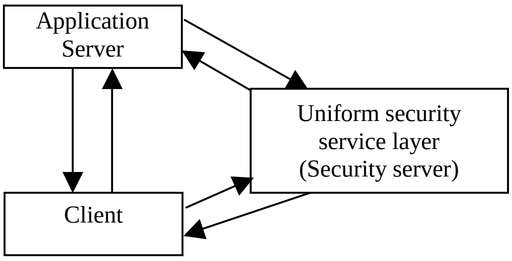


Figure1 Web service uniform security service system structure

(1). The client sends out the access request to the uniform security service layer.
(2). The uniform security service layer will return the ID of the client.
(3). The client will send out the application request and the ID to the application server.
(4). The application server will send out the ID to the uniform security service layer for certification and judgment.
(5). The uniform security service layer will return the results of certification and judgment.
(6). The application server will make the particular response according to the returned information.

3. THE WEB SERVICE TRANSFER BASED ON WS-SECURITY

In the electronic commerce environment, besides certification and judgment information of the traders, the secrecy, integrality and irreversibility of the transferred content are also under very strict requirement. Because what web service transfers are XML text based on the SOAP protocol, we can guarantee the secrecy, integrality and irreversibility of the transferring information with the extensive mechanism of the web service security.

WS-Security (Web service security) is a standard about the XML metadata container [5]. Many solutions have already been put forward to the transfer security of the network information. For example, the Kerberos and the X.509 are used for verifying the identity of the user [6], and the X.509 manages the secret key with the current PKI. The encryption and signature based on the XML describe the encryption and signature methods for the XML message content. The WS-Security adds these mechanisms in SOAP messages by appending a structure in the current standard and realizes a security extension pattern, which is irrelevant to transfer.

WS- Security defines a header element of SOAP to contain the data relative to security. If XML signature is used, the header element will contain the information of the XML signature definition, which includes the signature method of the message, the used secret key and the value of signature. In the same way, if a certain element in the message is encrypted, the header element of WS-Security can still include the encrypted information (for example, the encrypted information defined by XML encryption). The WS- Security does not specify the format of encryption or signature, but specify how to append the security information defined by other standard into the message of SOAP.

PKI (Public Key Infrastructure) is a quite mature system, which currently is used to meet the information security requirement of the open Internet. It guarantees the system information safety and checks the identity of the certificate holder by public key techniques and digital certificates [7]. Here we introduce the PKI mechanism to realize the information security during the information transfer. As a result, the uniform security service system structure based on web service will become into the form of figure 2:

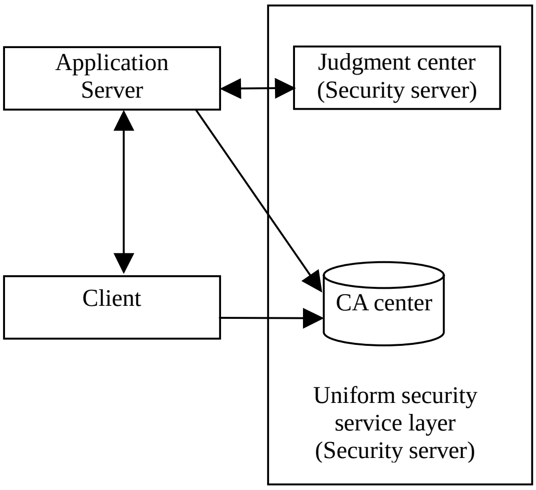

Application Server
Judgment center (Security server)
Client
CA center
Uniform security service layer (Security server)

Figure2 Web service transfer based on WS-Security

In the uniform security service system structure based on Web Service, the uniform security service layer consists of a CA (Certification Authority) center and a judgment center. The CA center can realize the identity authentication of both the client and the application server and the security and the integrality during the information transferring. Both the client and the application server should register in the CA center to apply ID certificates for themselves, and take them as their identity authentication symbols. Data in the SOAP transfer is no longer a simple XML script, but a XML file which realizes WS－Security. As an attribute in XML HEADER the ID certificate is encrypted by the user private key and the transferred content of XML BODY is also given digital signature. The application server decrypts and verifies the information transferred from the clients by their public key. Thus, the veracity, security and integrality of information can be guaranteed. The following is a SOAP message with WS － Security [8] , in which the customer token and digital signature are used.

```
<?xml version="1.0" encoding="utf-8"?>
<S:Envelope
xmlns:S="http://www.w3.org/2001/12/soap-envelope"
xmlns:ds="http://www.w3.org/2000/09/xmldsig#">
      <S:Header>
<m:path xmlns:m="http://schemas.xmlsoap.org/rp/">
<m:action>http://fabrikam123.com/getQuote</m:action>
<m:to>http://fabrikam123.com/stocks</m:to>
<m:id>uuid:84b9f5d0-33fb-4a81-b02b-5b760641c1d6</m:id>
</m:path>
<wsse:Security
xmlns:wsse="http://schemas.xmlsoap.org/ws/2002/04/secext">
<wsse:UsernameToken Id="MyID">
<wsse:Username>Zoe</wsse:Username>
</wsse:UsernameToken>
<ds:Signature>
<ds:SignedInfo>
<ds:CanonicalizationMethod
          Algorithm=
  "http://www.w3.org/2001/10/xml-exc-c14n#"/>
<ds:SignatureMethod
        Algorithm=
      http://www.w3.org/2000/09/xmldsig#hmac-sha1"/>
<ds:Reference URI="#MsgBody">
<ds:DigestMethod
      Algorithm=
  "http://www.w3.org/2000/09/xmldsig#sha1"/>
<ds:DigestValue>LyLsF0Pi4wPU...</ds:DigestValue>
</ds:Reference>
</ds:SignedInfo>
<ds:SignatureValue>DJbchm5gK...</ds:SignatureValue>
<ds:KeyInfo>
<wsse:SecurityTokenReference>
<wsse:Reference URI="#MyID"/>
</wsse:SecurityTokenReference>
</ds:KeyInfo>
</ds:Signature>
</wsse:Security>
</S:Header>
<S:Body Id="MsgBody">
<tru:StockSymbolxmlns:tru="http://fabrikam123.com/
payloads"> QQQ
               </tru:StockSymbol>
</S:Body>
</S:Envelope>
```

4. CONCLUSIONS

The web service solutions of electronic commerce based on the uniform security service enjoy the advantages as follows:

(1). Since all management mechanisms about the security aspects are concentrated on the uniform service layer, and all the units in the distributed system are not necessary to realize all function and mechanism of security solely. Therefore it will become very easy to set up and implement the security strategy, so is the maintenance of whole system.

(2). SOAP interface for the uniform security service makes this system structure very compatible. Just as what we have mentioned before, the uniform security service itself is also a web service. If the uniform security service layer can show its interface WSDL (Web Service Description Language), the API (Application Program Interface) of the uniform security service can be produced and used immediately.

(3). Because the uniform security service layer need not transfer the security credit information everywhere, it strengthens the security of the whole system. The uniform security service layer becomes the only place to accept the security credit information, so it can provide a wide range security service (beyond a particular security field). Meanwhile, the security credit information is still in this field.

5. REFERENCES

[1]. Ueli Wahli, Web service Wizardry Websphere Studio application developer, World Books Press, Oct .2002

[2]. Self-study Guide: WebSphere Studio Application Developer and Web service, http://www.ibm.com/redkooks

[3]. SOAP Version 1.2 Usage Scenarios http://www.w3.org/TR/2002/WD-xmlp-scenarios-20020626/

[4]. SOAP Version 1.2 Part 1: Messaging Framework,http://www.w3.org/TR/2003/PR-soap12-part1-20030507/

[5]. Specification: Web Services Security, http://www-106.ibm.com/developerworks/library/ws-secure/

[6]. Web Services Security: X.509 Certificate Token Profile,http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wss

[7]. PKI foundation, http://www.fanqiang.com/a5/b5/20010608/190000448.html

[8]. Web service security, http://www-900.ibm.com/developerWorks/cn/webservices/ws-secure/index.shtml.

Xiang Yang was born in 1973. He is a Ph.D. candidate in Department of Computer Science, Wuhan University of Technology. His research interests include high-performance computer networks and network simulation technology.

Firewall System Based on IPv4/IPv6

Min Lianying   Chen Jiong
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, 430063, China PR
Email: min_ly@mail.whut.edu.cn   Tel.:+86 (0)27-86554621

ABSTRACT

This Paper introduces the basic form of IPv6 protocol and the conversion mechanism about network address on IPv4/Ipv6; and then analyses the system structure of firewall and SOCKS protocol; Finally it designs a SOCKS firewall system structure based on IPv4/IPv6 protocol, points out the working principle and working procedure of it detailedly.

Keywords: Internet, IPv4, IPv6, Firewall.

1. INTRODUCTION

With the development and widely use of Internet throughout the world, for example the users of network are increasing rapidly in the developing country such as China and India, there are two important questions of the Internet development use which must be faced: the first one is that the IP address of 32 bits is becoming used up later; the second one is that the list of router is increasing as the speed of geometric series, which is far away above the speed of memory technology. In order to solve these two problems, which are serious restrict the development of Internet, Internet Engineering Task Force (IETF) have already begun studying and instituting the next protocol of network – IPv6 – in the 1990s. Because of the protocol of IPv4 is widely used throughout the world and IPv6 is developed on the IPv4, the protocol of IPv4 and IPv6 may be used in Internet nowadays until most hardware and software of network are upgraded to sustain IPv6. With the users of Internet are attach importance to the security of network, to design a firewall which sustains IPv4 and IPv6 is very impendence and necessary.

2. IPv6

The message of IPv6 is base header, and then it followed with one or more extension header, at last it is the data area. Because of extension header is the optional parameters, the least message of data is base header and data. The length of extension header may longer or shorter than base header. In many data message, the length of data area is far more longer than header. [1]

The format about Base header of IPv6 is shown in Fig. 1. It is mainly composed with version, priority, flow label, payload length, next header, hop limit, source address and destination address. The length of each portion is also shown in Fig. 1. According to this structure, we can see that IPv6 can express the IP address of 128 bits, the address space is four times than IPv4. This can solve the problem which address space is used up before long. In theory, IPv6 can connect with 2^{128}, which is very large astronomical number, equipment of network. Therefore the electric appliance such as television, refrigerator, air-condition etc. can be connected into Internet. By that time, we can control these electric appliances through Internet directly. The fancy of digital era may be realized in deed.

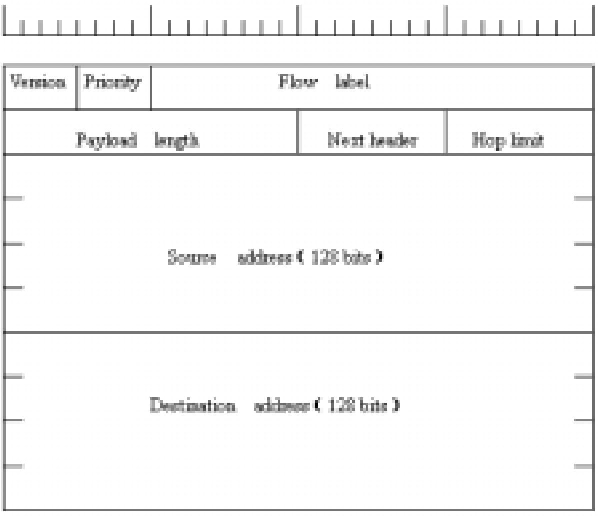


Fig.1 The Base Header of IPv6

The extension header of IPv6 can append accord to the demand, therefore it make a preparative for the wide usage of Internet. At present, IPv6 has defined six extension headers, they are:

(1) Hop—by—hop options. The use of it can stipulate the random options of the connect management by hop—by—hop.

(2) Destination options. The use of it can explain the random options of the management by the destination node.

(3) Routing. The use of it can point out the routing for the visit of destination host computer.

(4) Fragmentation. It is always used to manage the data segment, in order to whether to reinstall the signal of subsection.

(5) Authentication. It uses to offer the service of authentication and integrality for the information massage, in order to let the information is coming from the correct source node and it is not to be changed at random.

(6) Encapsulating Security Payload (ESP). It uses to offer the service of security for the information massage, in order to let the information don't be wire tapping by third party. Furthermore it is always stand at the last place of the extension header queue of IPv6. Whenever it is encrypted, it will be the last visible header.

The security of IPv6 is realized by the use of authentication and Encapsulating Security Payload.
Data area is made up of binary system code of transmitted information. It is the same of IPv4, so this paper will be omitted.

The address of IPv6 has three kinds, such as single channel broadcast address, excessive channel broadcast address and random channel broadcast address. Single channel broadcast address is used to distinguish the single interface, and excessive channel broadcast address is used to distinguish a group of interfaces. When data message transmits to a single channel broadcast address, it will be transmitted to a specify place. When data message transmits to excessive channel broadcast address, it will be transmitted to all member of excessive channel broadcast address group. Random channel broadcast address has similarity with random channel broadcast address, but data message will always be transmitted to one member of excessive channel broadcast address group (commonly may be the neatest member). Random channel broadcast address is a important expand of IPv6 protocol, it can solve some problems which IPv4 protocol cannot or difficult to solve. For example, when user connects with a group of consistent file servers, he can contact with the nearest server by random channel broadcast address without to know the material server he connects. Another example is that when the host computer of user uses random channel broadcast address as its gateway address, router may become a member of random channel broadcast address group, therefore the host computer may not know the exact address of router. When the place of router changes, data message will also transmit to the nearest number of random channel broadcast address group, user may not set a new address of router. Thus, the use of random channel broadcast address may predigest the problem that when the top structure changes the setting may also be changed in the IPv4 protocol.

3. CONWERSION ABOUT NETWORK ADDRESS ON IPv6 / IPv4

Compare with IPv4, IPv6 has improved some aspect such as extendable, routing, security, setting, reliability. But IPv4 is not compatible with IPv6, so when we use new protocol we must reset the software, which is in each of network hardware, to let it adapt the new protocol's requirement. [2] But as we all known, IPv4 is widely used in the Internet and with the induction of IPv6, IPv4 will not disappear immediately so we must set up a new conversion system in order to let the application of network hardware practicable when network hardware upgrades to IPv6. The raise of conversion about network address on IPv6/IPv4 is just based on above requirement.

In order to realize the network address and network protocol on IPv4/IPv6, we can consider two aspects as follows: one is that let the system on IPv6 accessing the system or service on IPv4; another is that let the system on IPv4 accessing the system or service on IPv6. Whether we use what kind method of above two, we must design a translator to set up an effective joint between two different network protocols. The different between these two methods is: the first one needs a group of IPv4 address which mapped the whole IPv6 address; the second one needs private IPv4 address which point to IPv6. The particular of these two methods are introduced in literature [3]. If readers have interest in these, you can refer to literature [3] directly.

4. SYSTEN STRUCTURE OF FIREWALL

Firewall is made up of one or a group of network equipment, which is used to reinforce the access control between two or more network. The form of the realization of firewall has many methods, but a common firewall system has three basic characters as follows: the first one is that all data, which are transmitted between interior network and exterior network, should get across the firewall; the second one is that only the accredited and legal data can get across the firewall; the third one is that the firewall itself will not be attacked by any other outside influence.

The system structure of firewall mainly has two kinds method as follows: [4]

⑴ Proxy Host. By this method, interior network dose not communicate with Internet directly. Interior network protocols (such as Netbios, TCP/IP, etc.) are used between interior network users and proxy gateway. TCP/IP protocol, which is the standard protocol about communication, is used between gateway and Internet.

⑵ Router and Screened Host. By this method, we use router and filter to restrict the exterior computer visits interior network by IP address or domain name, and we also can appoint or restrict interior network visits Internet. We can obstruct the deviant accessing logging in between interior network and exterior network by the use of filter host compute which is used to enforce filtration, percolation, validation and security scout.

At present, the technique of firewall mainly has several methods as follow:

⑴ Application gateway firewall;
⑵ Package filtration firewall;
⑶ Complex firewall;
⑷ SOCKS firewall;
⑸ SOCKSv5/TLS firewall.

5. FIREWALL SYSTEM STRUCTURE BASED ON IPv4/IPv6 AND SOCKS

In 1998, SOCKS is open out by NEC (U.S.A.) and use in firewall technique. The firewall system based on SOCKSv5 has a stronger authentication mechanism. Authentication method may has some characteristic such as choose or arrange, address parse proxy, expand address on IPv6 and IPv4, expand application on UDP, realize the integrity and encrypt of data, clarity to users, etc. .

SOCKS is a canonical protocol. This protocol realizes by a proxy server based on C/S mode network environment. SOCKS is independent from application layer protocol. Seen from OSI network structure, SOCKS is between transport layer and application layer. [5] In order to realize the conversion of IPv4 protocol and IPv6 protocol in SOCKS firewall, we must use the method about IPv6/IPv4 network address conversion, which discussed in literature [3], to design a translator in SOCKS server. According this theory, we design a Firewall system structure based on IPv4/IPv6 and SOCKS, which is shown in Fig. 2.

Underside we will analyze the working principle of above firewall system structure particularly

If we use direct address conversion, and set exclusive IPv4 address to SOCKS client and SOCKS server, an then the working principle of this firewall based on IPv4/IPv6 and SOCKS is: Firstly, user will input destination host address (suppose this address is IPv6 address), and then SOCKS client

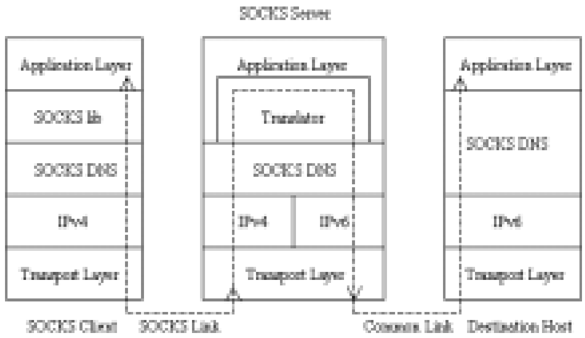


Fig.2 The System Structure of Firewall Based SOCKS Protocol and IPv4/IPv6

will call some corresponding function in SOCKS lib to set corresponding requirement to SOCKS server; Secondly, when SOCKS server receive the requirement from SOCKS client, it will call systemic function to know whether this destination address is the same type address with SOCKS client address; Thirdly, SOCKS server will operate and maintenance the address conversion mapped table, which is designed for IPv4/IPv6 in application layer specially, and get a IPv6 address of destination host. If SOCKS server can connect with destination now, it will return a success response to SOCKS client, otherwise it will not return any response to SOCKS client; Lastly, when SOCKS client receive the success linking response from SOCKS server, it will visit the application layer on destination host pellucidly on its application layer. According to the analysis above all, we can see that actually the transmission of all data will pass through SOCKS server, this is the basic thought of firewall based on SOCKS.

If we use domain name parse address conversion, the working principle of this firewall based on IPv4/IPv6 and SOCKS has some conform with the working principle when we use direct address conversion, but there are still some differences between them. Underside we will emphases to introduce the full course about how to use “DNS name parse proxy” to realize the connection between SOCKS client and destination host:

⑴ SOCKS client will call its DNS name parse function in SOCK lib (in common use is Gethostbyname( )) to try to get the IPv6 address of destination host. At the same time, the name of host (FQND) will be called to the SOCKS storeroom of SOCKS client as parameter.
⑵ The name of host (FQND) will be save to a mapped table in the SOCKS storeroom of SOCKS client, and at the same time a dummy IP address, which is matched with the function application called, will be produced.
⑶ Once application receive the dummy IP address, application will use this dump IP address as a parameter of socket to call socket function (in common use is connect( )) to initiate a connecting request. And then, application will estimate by the IP address of socket: if the address is dummy IP address, the application will take out the corresponding FQND from the mapped table in the SOCKS storeroom to match itself.
⑷ Application will use SOCKS command, which is matched with the created socket, (according to above this may use the command of CONNCET) to send FQND to SOCKS server by SOCKS linking.
⑸ SOCKS server will use the receive FQND as a parameter to call the real DNS name parse function (in common use is Getaddrinof( )).
⑹ SOCKS server get the real IPv6 address from its SOCKS DNS, and at the same time it will create the matched socket.
⑺ SOCKS server will use the real IPv6 address as a parameter to call the socket function(in common use is connect( )), in order to create a common linking and communicate with destination host.

6. CONCLUSION

This paper points out the working principle and process according to the method of the conversion about network address on IPv6 / IPv4. It will solve the problem that there is still very little hardware equipment, which sustain IPv6 directly, in the market nowadays and how to solve the matching about IPv4 protocol and IPv6 protocol. And it also provide a fine foreground for the widely use of IPv6 protocol.

7. REFERENCES

[1] Huitema C．IPv6：The New Internet Protocol [M]．Second Edition ，USA：Prentice Hall inc.，1999
[2] Malkin G，Minnear R．RIPng for IPv6[J]．INTERNET-DRAFT，1997，1：54~57
[3] HE Xiaoyan， WU Jieyi． IPv6/IPv4 Network Address Translation [J] ．Computer Engineering and Application，2000，7：122~124
[4] LU Canglong ． Network Security and Firewall System[J]．Data Traffic，2003，2：21~23
[5] YANG Chun ． Study of SOCKS Application in Firewall[J]． Transaction of University of Electronic Science and Technology，1999，26（2）：199~201

Min Lianying is an associate professor in School of Computer Science and Technology, Wuhan University of Technology. His research interests are in computer network and information system.

Chen Jiong is a graduate student in School of Computer Science and Technology, Wuhan University of Technology. His research interest is in network security.

Empirical Studies for Two Evolutionary Fuzzy Controllers

Xu Huazhong[1] Wang Pan[1,2] Xu Chengzhi[1] Zhang Jianjian[1]
[1]Wuhan Univ. of Technology (East Yuan) [2]Huazhong Univ. of S&T
Wuhan, 430070, P.R. China
Email: jfpanwang@tom.com **Tel:** 86-27-87858435

ABSTRACT

Empirical studies are given for two kinds of adaptive fuzzy control strategies based on evolutionary computation (EC)–multi-regulated factors fuzzy controller and qualitative-quantitative self-regulated fuzzy controller. Some complex and hard-to-control plants are selected such as chaotic system, nonlinear MIMO systems. Results illustrate these control strategies have satisfactory dynamic, steady and robust performance. Meanwhile, some key issues about evolutionary are discussed.

Keywords: Empirical Studies, Evolutionary Computation, Fuzzy Control.

1. INTRODUCTION

In [1], two kinds of adaptive fuzzy control strategies were presented by us based on evolutionary computation (EC). Principles, methods and steps of these two algorithms are analyzed. In these strategies, some key parameters of two self-regulated fuzzy controllers (a multi-regulated-factors fuzzy controller and qualitative-quantitative self-regulated fuzzy controller) are optimized by EC. Both linear and nonlinear quantization functions as quantized formula are employed and ITAE index is applied as fitness function. As a consequent part, empirical studies are given for the control of some complex plants. At the end, some key issues about evolutionary are discussed. Following are the outline formula of these strategies:

Strategy1 multi-regulated factors fuzzy control

$$U_i^{'} = \begin{cases} <c_{0i}E_i + (1 - c_{0i})EC_i> & E_i = 0 \\ <c_{1i}E_i + (1 - c_{1i})EC_i> & E_i = \pm 1 \\ <c_{2i}E_i + (1 - c_{2i})EC_i> & E_i = \pm 2 \\ <c_{3i}E + (1 - c_{3i})EC_i> & E_i = \pm 3 \end{cases} \tag{1}$$

Strategy2 Qualitative-quantitative self-regulated fuzzy control. Specific meanings can be found in [1].

2. EMPIRICAL STUDIES

2.1 Track Studies

Large volumes of simulations are carried out for complex systems such as MIMO systems, chaotic systems and systems with time delay. The partial results are as follows:

Model 1.Three-level system with time-delay [2]

$$U_i = \frac{\beta_{1i}|E_i|}{\beta_{1i}|E_i| + \beta_{2i}|EC_i|}E_i + \frac{\beta_{2i}|EC_i|}{\beta_{1i}|E_i| + \beta_{2i}|EC_i|}EC_i + U_{0i} \tag{2}$$

$$G(s) = \frac{k_0 e^{-\tau s}}{(T_1 s + 1)\quad(T_2 s + 1)\quad(T_3 s + 1)} \tag{3}$$

Where $k_0 = 1.0,\ \tau = 0.2,\ T_1 = 1.0,\ T_2 = 2.0,\ T_3 = 5.0$
Input : r(t)= 1(t);

Eq(4) is the correspondent difference equation:

$$y\ (k) = a_0 y\ (k-1) + a_1 y\ (k-2) + a_2 y\ (k-3) + a_3 u\ (k-m)$$
$$+ a_4 u\ (k-1-m) + a_5 u\ (k-2-m) + a_6 u\ (k-3-m) \tag{4}$$
$$m = \langle \tau / Ts \rangle,\ a_0 = 2.8362,\ a_1 = -2.6799,\ a_2 = 0.8436$$
$$a_3 = a_6 = 1.15 \times 10^{-5},\ a_4 = a_5 = 3.45 \times 10^{-5}$$
$$m = \langle \tau / Ts \rangle,\ a_0 = 2.8362,\ a_1 = -2.6799,\ a_2 = 0.8436$$
$$a_3 = a_6 = 1.15 \times 10^{-5},\ a_4 = a_5 = 3.45 \times 10^{-5}$$

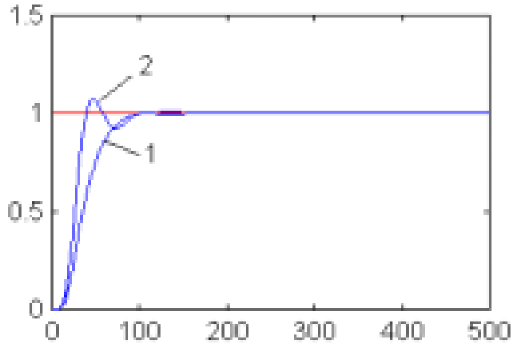


Fig. 1 Step response of Strategy 1

Fig.1 is the step response of **Strategy 1** both with linear quantization and with nonlinear quantization. Where curve 1 represents the response curve with linear quantization while curve 2 represents the response curve with nonlinear quantization.

For linear quantization: e(∞)=0.0004486,ITAE=10.0792; For nonlinear quantization: e(∞)= -0.0000117, ITAE=5.2752

$$\begin{cases} x_{k+1} = -px_k^2 + y_k + 1 \\ y_{k+1} = qx_k \end{cases} \tag{5}$$

$$p = 1.4\ ,\ q = 0.3$$

Model 2. Helon chaotic system [3]
Our goal is to control xk and make it stabilized in an unstable fixpoint:(0.6314,0.1894)。Following is the structure of chromosome:

$$\beta_1, \beta_2, Ku_1, Xe, Xec, U_0, Ku_2$$

This work is partly supported by NSF of China (Grant no.: 60174039).

Where *Ku1* 和 *Ku2* are the amplified constants before and no less than 6 steps. After optimization，the optimized parameters are found　Relative results and control curve are as follows:

Chromosome:
(0.090172243534353436; 0.86552127062706274; 0.066780740334033403; 1.0673219331933192; 0.21778016101610162; -4.3346754605460545; 0.0067546566656665665)
ITAE=21.54;

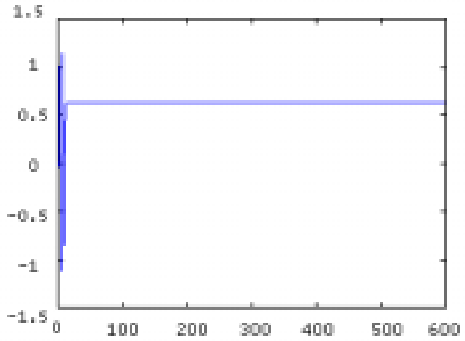


Fig. 2 Step response of **Strategy 2**

Model 3. MIMO system [4]

$$\begin{cases} y_1(k+1)=0.4y_1(k)+\dfrac{v_1(k)}{1+v_1^2(k)}+0.2v_1^3(k)+0.5v_2(k) \\ y_2(k+1)=0.4y_2(k)+\dfrac{v_2(k)}{1+v_2^2(k)}+0.4v_2^3(k)+0.2v_1(k) \end{cases} \quad (6)$$

(1)　Step response

$$R_{10}=\begin{bmatrix} r_1(k) \\ r_2(k) \end{bmatrix}=\begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

Following is the structure of chromosome:

β_{11}, β_{21}, ku_{11}, Xe_1, Xec_1, $U0_1$, ku_{21}, β_{12}, β_{22}, ku_{12}, Xe_2, Xec_2, $U0_2$, ku_{22}

Where Ku_1 和 Ku_2 are the amplified constants before and no less than 3 steps.

After optimization，the optimized parameters are found Relative results and control curve are as follows:

(0.9844; 0.329294; 0.0502485; 1.11149; 0.602321; 10.1193; 0.0993054; 0.17292; 0.892409; 0.0126661; 0.228389; 0.784035; -14.1052; 0.0145923)
ITAE=0.2221, and: α=(α1，α2)=(0.5，0.5)

(2)　Sine input response

$$R_{11}=\begin{bmatrix} r_1 \\ r_2 \end{bmatrix}=\begin{bmatrix} 0.5\sin(t) \\ 0.5\sin(t) \end{bmatrix}$$

Following is the structure of chromosome:
β_{11}, β_{21}, ku_{11}, Xe_1, Xec_1, $U0_1$, ku_{21}, β_{12}, β_{22}, ku_{12}, Xe_2, Xec_2, $U0_2$, ku_{22}

Where Ku_1 和 Ku_2 are the amplified constants before and no less than 3 steps.

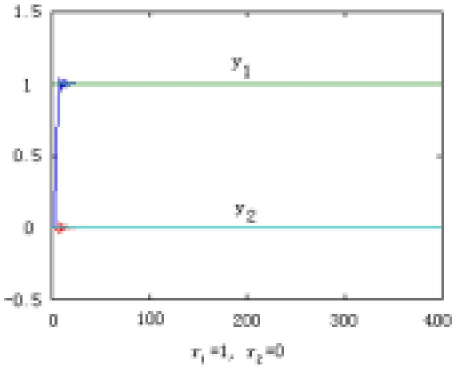


Fig. 3 Step response of **Strategy 2**

After optimization，the optimized parameters are found Relative results and control curve are as follows:

(0.138209; 0.914379; 0.0209151; 0.107254; 1.80087; -0.00168077; 0.912279; 0.978738; 0.04036; 0.177828; 1.79418; -0.00233543)

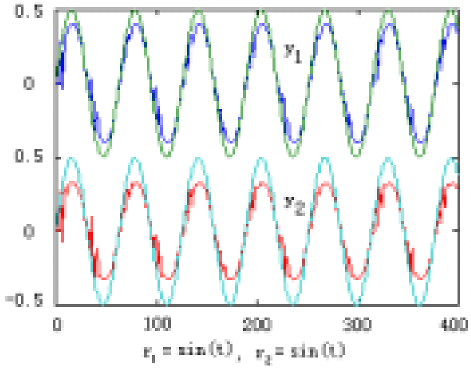


Fig. 4 Step response of **Strategy 2**

(3)　Anti-interruption experiment

$$\begin{cases} y_1(k+1)=-0.039\,y_1(k)-0.5y_2(k)- \\ 0.0498\,y_1(k-1)+0.9525\,u_1(k)+ \\ 0.6u_2(k)+0.1363\,u_1(k-1) \\ y_2(k+1)=-0.5y_1(k)+0.2y_2(k)+ \\ 0.25\,y_2(k-1)+0.25u_1(k)+ \\ 0.9u_2(k)+0.4u_2(k-1) \end{cases} \quad (7)$$

Model 4. Two input and two output system [5]
β_{11}, β_{21}, Ku_1, Xe_1, Xec_1,U_{11},U_{21}, β_{12}, β_{22},, Ku_2, Xe_2, Xec_2,U_{12},U_{22} where U_1、U_2 are two different steady- state values.

By optimization with several times, following satisfactory parameters are obtained.

(0.962231; 0.852035; 0.0509751; 0.716212; 1.94749; 89.0379; 51.3706; 0.576218; 0.683888; 0.059916; 0.625563; 1.9793; 39.4284; 26.5782)

In our anti-interruption experiment, model is polluted into:

$$
\begin{cases}
y_1(k+1) = -0.039y_1(k) - 0.5y_2(k) - \\
0.0498y_1(k-1) + 0.9525u_1(k) + \\
0.6u_2(k) + 0.1363u_1(k-1) + 0.25rand \\
y_2(k+1) = -0.5y_1(k) + 0.2y_2(k) + \\
0.25y_2(k-1) + 0.25u_1(k) + 0.9u_2(k) + \\
0.4u_2(k-1) + 0.25rand
\end{cases}
\tag{8}
$$

The relative result is shown as in **Fig. 5** with the same parameters.

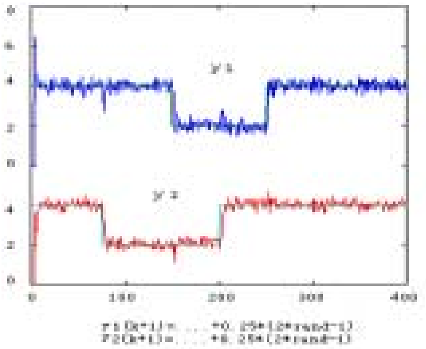


Fig. 5 Anti-interruption experiment

We also carry out the anti-interruption experiment while the signal is interrupted. For the length limit of this paper, the results will not be reported.

(4) Robustness experiment
We carry out robustness experiments, following is the result while the model is changed into:

$$
\begin{cases}
y_1(k+1) = -0.039y_1(k) - 0.5y_2(k) - \\
0.0498y_1(k-1) + 0.9525u_1(k) + \\
0.5u_2(k) + 0.1363u_1(k-1) \\
y_2(k+1) = -0.5y_1(k) + 0.2y_2(k) + \\
0.25y_2(k-1) + 0.25u_1(k) + \\
0.9u_2(k) + 0.5u_2(k-1)
\end{cases}
\tag{9}
$$

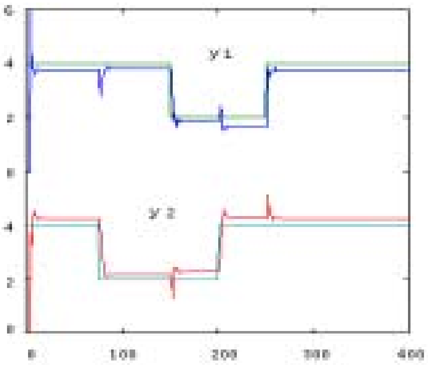


Fig. 6 Anti-Interruption Experiment

Fig. 6 shows the well robustness of the presented algorithm(s).

3. CONCLUSIONS

Empirical studies are given for two kinds of adaptive fuzzy control strategies based on evolutionary computation (EC) – multi-regulated factors fuzzy controller and qualitative-quantitative self-regulated fuzzy controller. Some complex and hard-to-control plants are selected such as chaotic system, nonlinear MIMO systems. Results illustrate these control strategies have satisfactory dynamic, steady and robust performance.

Generally speaking, for many specific problems, the strategies put forward in this paper may not be the best ones, but those could obtain globally satisfactory results. Based on these results, people could further find parameters conveniently to gain excellent performance, realize effective integration of man-machine. However, the presented strategies have also some deficiencies and face some challenges. Simply speaking: 1) How to integrate ECs (or hybrid ECs) into on-line controllers is an important problem deserving to be solved thoroughly. 2) The problem of selection of performance criterion is a key issue to be addressed. 3) They don't have some remarkable advantages of other branches of computational intelligence methods and it's necessary to integrate the advantages of these different branches. All the above would be the direction of our next work.

4. REFERENCES

[1] Wang Pan, Xu Chengzhi, Feng Shan, et al, "Two Kinds of Novel Evolutionary Fuzzy Controllers."DCABES'2004 (to appear).

[2] Wang Pan. Applicational Researches with Soft Computing for Some Decision and Control Issues, PhD dissertation, Huazhong University of Science and Technology, 2003.

[3] Fang Jian'an., Shao Shihuang, "Control of A Kind of Chaotic System Using Genetic Algorithm and Fuzzy Logic", Proceeding of Industrial Technology, 1996, Shanghai.

[4] Wang Pan, Xu Chengzhi, "A New Evolutionary Fuzzy Control Algorithm and Simulation Studies", Technical Report2001-01-01, WHUT, 2001.

[5] Jin Qibing., Gu Shusheng, Niu Yujiao, "Parameter-Converged Rapidly Neural PID Control for Multivariable Systems", Chinese J. Control and Decision, Vol. 13 (Suppl), 1998, pp. 448-452.

Wang Pan is a Full Associate Professor and a head of Institute of Control and Decision, Wuhan University of Technology. He received the B.S. degree in industrial automation from Wuhan University of Technology, Wuhan, P. R. China, and the M.S. and Ph. D. degrees in systems engineering from Huazhong University of Science and Technology, Wuhan, P. R. China. He has published over 30 Journal papers, 15 Conference papers. His research interests are intelligent control, decision analysis, and biomedical intelligent information systems.

The General M Set and Julia Sets Generated by Complex Iteration $Z_{N+1}=Z_N^{-2} + C$

Zhe Xu, X.G. Deng, X.D. Liu, W.Y. Zhu,
School of Information and Engineering, Northeastern University
Shenyang 110006 China
Email: xuzhe@263.net dengxuegong@sohu.com Tel.: 13066622225/024-23842017 & 13998158421

ABSTRACT

In this paper, the General Mandelbrot set and Julia sets of complex iteration $z_{n+1} = z_n^{-2} + c$ were studied. The number and classification of periodic buds were given and the growth rules of Julia set in a bud were described. According to the classification of the buds, the similarity of the Julia sets that are in the same bud is described, and the difference of the Julia sets that the construction parameter is in different buds are described, attached with many computer images.

1 INTRODUCTION

The chaos and fractal images generated by complex iteration $z_{n+1} = z_n^{-2} + c$ have been widely studied in many papers[1−8] by the method of escape time algorithm. In this paper, we mainly studied the inner structure of the General Mandelbrot set (will be noted as GM below) of complex iteration $z_{n+1} = z_n^{-2} + c$ and the distribution of the buds in the set. We will show how many buds of the same period, how to distinguish them and the Julia set with parameter in it, and the growth of the Julia set when the parameter changes in a bud as well.

The GM set of f (z, c) = z $^{-2}$+c

The General Mandelbrot set of complex iteration $z_{n+1} = z_n^{\alpha} + c$ is often defined as $M = \{c \in C: \quad z_n \not\to \infty, when \quad n \to \infty\}$ [1-5]. This definition needs a compelling escape criterion. When α=2, or the iteration function is a polynomial, the image of the GM set is very intelligible, but when $\alpha = -2$, it is quite difficult to get a escape criterion to make the image clear. Fig 1 is the image of the GM of $z_{n+1} = z_n^{\alpha} + c$, many papers have drawn lots of interesting conclusions by studying this image. But it is not very convenient for our study, because we are interested in the periodic and topological property of the GM. So we give the following definition, which is equal to the definition above, in the meaning of measurement.

$M_f = \{c \in C$:there is a number N_c, $\lim\limits_{n\to\infty} f^{nN_c}(c,c)$ exist}

N_c is the period of c. This definition can be modified but be equal to the image as following:

$$M_P(c) = \begin{cases} RGB(255,255,255) & if \quad no \quad N_c \\ RGB(a(n),b(n),c(n)) & if \quad N_c = k \end{cases}$$

The functions of $a(n),b(n),c(n)$ are only color functions with no actual meaning but the visual impression, and they cannot all be 255 at the same time.

$$M_f = \{c \in C : M_p(c) \neq (255,255,255)$$

We get the image of M_f by computer as fig2.

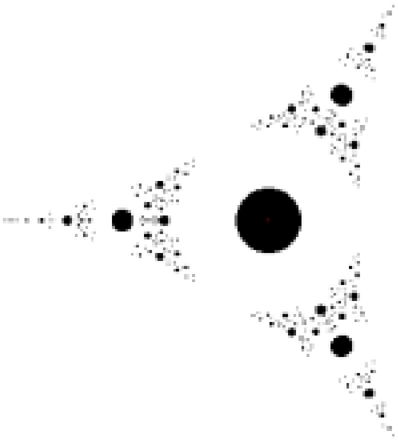


Fig1 M_f by escape time algorithm

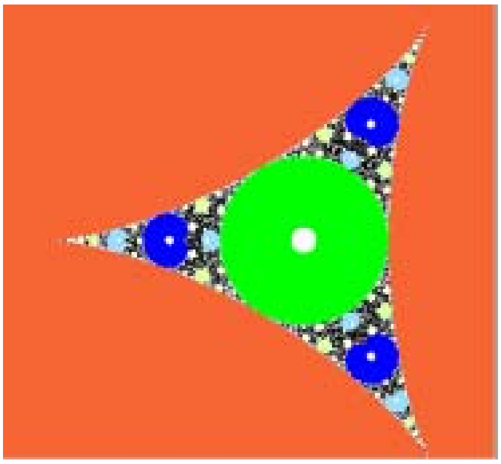


Fig2 M_f by our definition

Buds and the classification

In Fig2, the petal-like parts with a single color are called buds (single connected subsets). The part with a color of red is the 1-periodic bud (it is also a petal-like bud when we see it in Riemann sphere), the green one is the 2-periodic bud and so on. In bud B, for all $c \in B$, $f^n(c,c)$ has an n-periodic attractive periodic orbit, so we call it n periodic bud. On the boundary of any bud, there are just 3 points, which are tac-points, or tangent points with other buds that have lower period. We call these points dividing points. On the real-axis, we can see a bud sequence with a digressive scale from the 2-period bud to the left. The period of these buds are 3, 4, 5…. in turn. We can also see many buds attached to this bud and there are many buds in the set. First of all, we give a classification of the buds as following

1) Main bud: the bud attached to the 1-period bud two

times or more,

2) K-attaching bud: the bud attached to other buds at only one point and has k-time period,
3) Isolated bud: the bud that does not meet any larger bud,
4) Scale biggish bud: the bud which is larger than other ones in certain family of buds, attached to one bud.

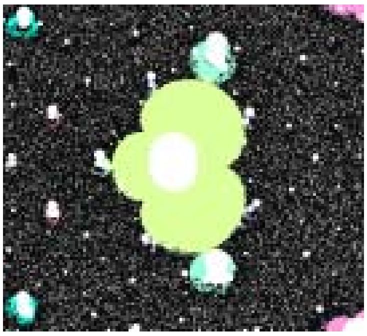


Fig3 Isolated bud with period 5 between 3 and 4 main bud

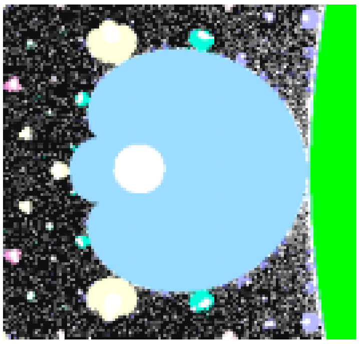


Fig4 2- attaching bud to main bud with period 2

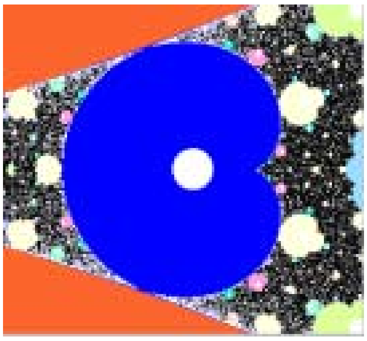


Fig5 Main bud with period 3 on real direction

For any n-periodic bud, there are three 2n-periodic bud attached to it at the three parts of the boundary respectively. From these tangent points to the dividing point, the scale biggish bud has periods of 3n, 4n, 5n …in turn. Since a majority of buds are isolated buds or attaching to isolated buds, the description we mentioned above cannot perfectly describe the bud distribution of the GM set.

2 THE NUMBER OF THE BUDS

Any bud in the GM set has a kernel, which implies there is an infinite attracting periodic orbit in it. This property can be used to calculate the number of buds of certain periods. Let Q(n) be the number of n-periodic bud, P(n) be the number of the solution of equation fοfο οΛ Κ fοf (0, c) = $f^n(0, c)=0$ (n>1) and { n_i } be the set of all nontrivial factor of n.

$$\text{As we know } f^2(0,c) = \frac{1-c^3}{c^2}$$

$$\text{If } f^n(0,c) = \frac{\sum_{i=1}^{P(n)} a_i z^i}{\sum_{i=1}^{M} b_i z^i}$$

$$\text{Then } f^{n+1}(0,c) = \left(\frac{\sum_{i=1}^{P(n)} a_i c^i}{\sum_{i=1}^{M} b_i c^i}\right)^{-2} + \mathrm{c} = \frac{\left(\sum_{i=1}^{M} b_i c^i\right)^2 + c\left(\sum_{i=1}^{P(n)} a_i c^i\right)^2}{\left(\sum_{i=1}^{P(n)} a_i c^i\right)^2}$$

so P(n+1) = 2P(n) +1 and we have known P(2) = 1，P(3) = 3 Since any bud has a kernel in which there is a point c which has a super attractive periodic orbit with ∞, and this c is a solution of c = f^n (0, c), so

$$\mathrm{Q(n) = P(n)} - \sum_{n_i} \mathrm{Q(n_i)}$$

Table1 is the list of the number of buds of period n.

Table1 the number of n-period bud

n	1	2	3	4	5	6	7	8	9	10	11	12
P(n)		1	3	7	15	31	63	127	255	511	1023	2047
Q(n)	1	1	3	6	15	27	63	120	252	495	1023	2037

The boundary of 1-period bud

Let us look at the image of real function $f(x,c) = x^{-2} + c$, When two curves $y = x^{-2} + c$ and $y + x$ meet at a tangent point (x_0, y_0), the complex c is just the left acme of the triangle like part of GM set. we can easily get the value c by

$$\begin{cases} y'(x_0) = (-2x_0)^{-1/3} = 0 \\ \quad x_0 = x_0^{-2} + c \end{cases}$$

so: $c = x_0 - x_0^{-2} = (-1/2)^{-1/3} + (-1/2)^{-2/3}$

The Fibonacci sequences in the GM

In the GM, we have defined the scaling biggish bud, for example, Fig.3 is the local part near the tangent point of 1-periodic bud and 2-periodic bud. We can find that all the buds have a common property that

periodic number of scaling biggish satisfy Fibonacci rule. In Fig.3 we can see a bud between the 2-periodic bud and the 3-periodic bud there are a comparatively bigger bud with period 5 attached to the 1-periodic bud. This phenomenon can be found at all bud's boundary in the GM. So we have following formula, let P(a) and P(b) denote the period of two buds a an b on the same part of the boundary of a bud and there are no larger buds between them and c is the scale biggish bud between a and b, then P(c) = P(a) + P(b).

The Julia set

Let $\lambda(w) = \prod_{i=0}^{p-1} f'(f^i(w))$ be the characteristic value of w, if $0 \le |\lambda| < 1$, w is an attractive point, if $|\lambda| > 1$, w is a repulsive point. The Julia set is usually defined as $J_f = \overline{\{w \in C : |\lambda(w)| > 1\}}$ but this definition is not very convenient for computer program. In the meaning of measure, we give the following definition of Julia set.

Let $\{ z_i \}$ is an attractive period orbit for parameter c. $J_f = \bigcup_i \partial Q(i)$ in which Q(i)= {z: $f^p(z, c)$ converge to a fixed point z_i }. Since there is no other attractive periodic orbit in all buds of the GM set, this definition is reasonable. As a matter of fact, when the mapping has more than one attractive period orbit, the Julia set can also be defined in this way.

In this section, we mainly discuss the changing behavior of the Julia set in one bud of the GM. Firstly, we find that the Julia set of the parameter in the main periodic buds has an obvious character. (see fig 8-11). This character merely belongs to the main periodic buds. At intersecting points, the neighborhood region is divided into k parts if the bud has a period k, and that behavior can be found at all detailed parts of the Julia set. When the parameter lies in a k-attaching bud, we can see other interesting phenomena, at the intersecting points that the neighborhood region is divided into 2k-parts and these parts are alternately chaotic region and open attractive region (see fig15). When the parameter lies in an isolated bud: the case is quite different, we can no longer find any periodic information of the bud by merely observing the geometric structure of the Julia set (but the number of color, see fig12, 13). Certainly, for all parameters, we can carefully count how many colors appeared in the Julia set and the number is just the periodic number of the Julia set for all parameters. But this property is not the geometric property of the Julia set.

All the Julia sets, with parameter in the same bud, are equivalent in the meaning of topology, but the evolvement of Julia set by the changing of the parameters in a bud is quite interesting. If the parameter in 1-periodic bud, the Julia set is dust like set, the structure is not quite clear for us. When the parameter lies in other buds, The changing of the image of the Julia set is well-regulated when the parameter changes in a bud with period larger than 1. We can see that the parameter near the kernel the Julia set is relatively regular, but when the parameter moves to the boundary of the bud the growth of the Julia set is evoked. Many horn-like parts grow from the Julia set and the number of the legible horns depends on the point on the boundary that the parameter moves to. While the parameter near the tangent point of n-time periodic bud, we can find that the number of horns is just n, and their tines move to a single point bit by bit and converge at one point when the parameter is just the tangent point.

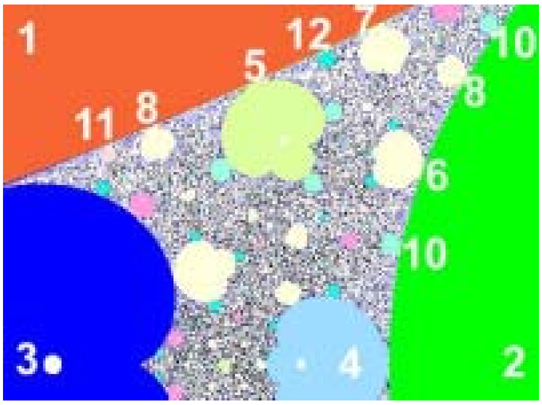


Fig 6 Picture for scale biggish bud

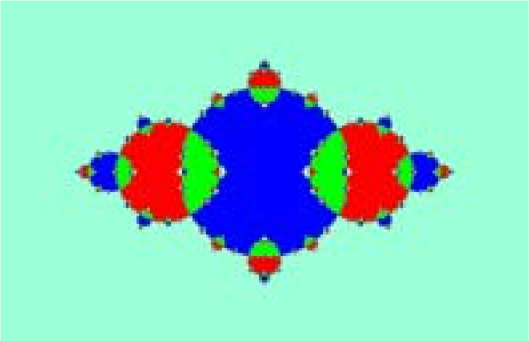


Fig7 The Julia set of -1.380948+0i in 4-main bud near the center

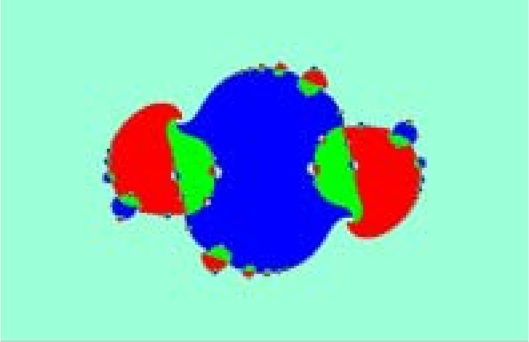


Fig8 The Julia set of -1.325016+0.033 1 i in 4-main bud near the 1-periodic bud

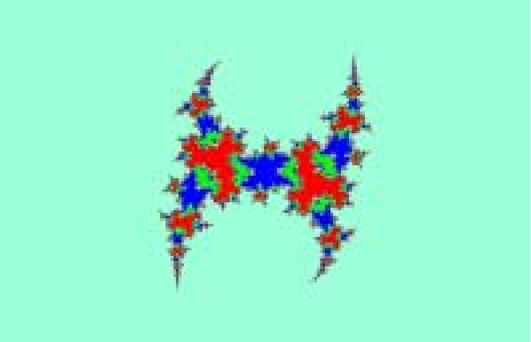


Fig9 The Julia set of -1.380948+0i in 4-main bud near the 2- attaching bud

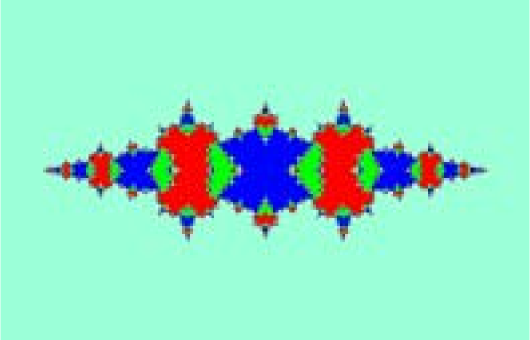


Fig10 The Julia set of -1.345479+0 i in 4-main bud near the tac-point

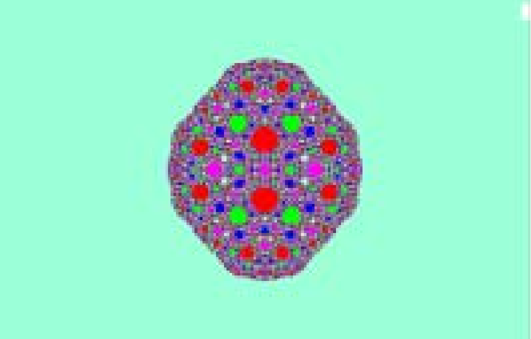


Fig11 The Julia set of -1.380948+0i in an isolated bud between 3 and 4 main bud

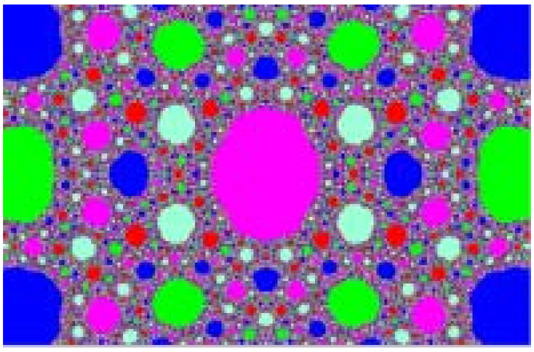


Fig12 The enlarged center of fig12

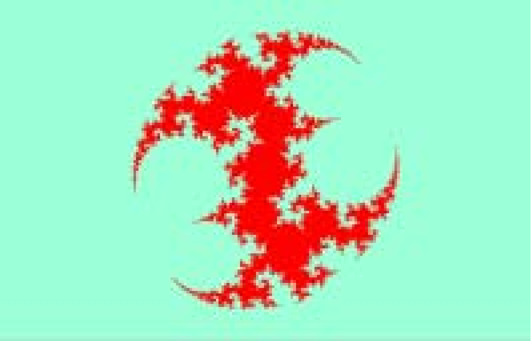


Fig13 The Julia set of -0.5917 +0.21537i in the 2-main bud near the tangent point with the 3-attaching bud

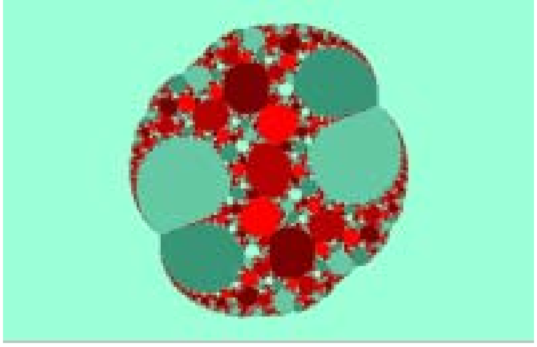


Fig14 The Julia set of -0.592 +0.21547i in the 3-attaching bud of the 2-main bud near the tangent point

The phenomena occurring in fig14-15 can be explained by Fig16, which is obtained by limiting the iteration $f(z,c) = z^{-2} + c$ at line $y = \frac{8}{9}x$ (the tangent point of the 2- attaching bud of 2-periodic bud in this line). The iteration $x_n = x_{n-1}^{-2}$ + c can be written as following: c=a+bi, x_0 =cosk, y_0 =sink. $x_n = \frac{x_{n-1}^2 - y_{n-1}^2}{(x_{n-1}^2 + y_{n-1}^2)^2} + a$, $y_n = \frac{-2x_{n-1}y_{n-1}}{(x_{n-1}^2 + y_{n-1}^2)^2} + b$ y_n =cosk, and we can get two interesting multi- bifurcation diagram Fig11 and Fig12. The horizontal lines are all x=0.

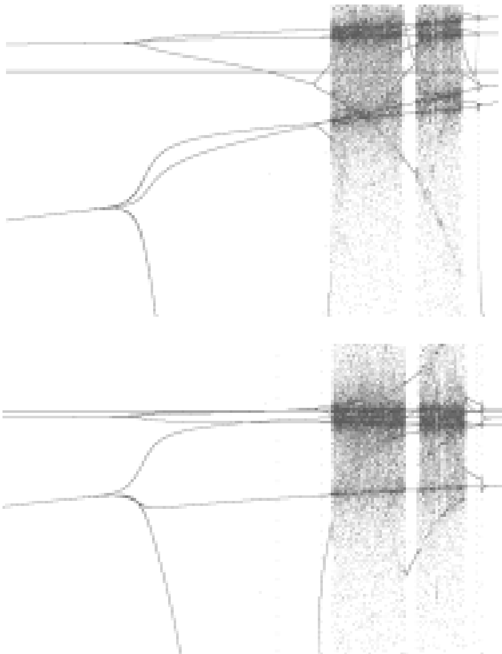


Fig 15 the multi- bifurcation diagrams

This multi- bifurcation diagram illuminates that, near the tangent points of k-attaching bud, one attracting point can separate two or more attracting points, and the attracting regions are divided into n parts.

3 CONCLUSION AND CONJECTURES

In this paper, we have given the number of the periodic buds in the GM, and showed that the buds in this set can be classified. And moreover, we have indicated that the scaling biggish buds satisfy the Fibonacci sequence. We have also studied the growth of Julia sets in a bud. These works may be helpful for the study of the dimension of the GM.

4 ACKNOWLEDGEMENT

Our research is partially supported by the Doctoral Program Foundation of the State Education Commission of China (2000014512) and the Natural Science Foundation of Liaoning Province, P R China. (69974008)

5 REFERENCES

[1] Mandelbrot B.B. The Fractal geometry of Nature [M], San Fransisco: Freeman WH, 1982.

[2] Gujar U G, Bhavsar V C. Fractals from $z \leftarrow z-\alpha+c$ in the complex c-plane [J], Computers & graphics, 1991, 15(3):441-449.

[3] Gujar U G, Bhavsar V C. and Vangala N., Fractals Images from $z \leftarrow z^{-\alpha}+c$ in the complex c-plane Faculty of Computer Science. University of New Brunswick, Fredericton, Canada (5 1990)

[4] Dhurandhar S V. Bhavar V C. Gujar U G. Analysis of z-plane fractals images from $z \leftarrow z^{-\alpha}+c$ $\alpha<0$[J]. Computers & graphics

[5] Ken W Shirriff An investigation of fractals Generated by $z \leftarrow z^{-\alpha}+c$[J] Computers &graphics1993 17 603-607

[6] Lakhtakia, Varadan V V, Messier R. On the symmetries of the Julia sets for the process $z=z^{-\alpha}+c$ J.phy.A math.Gen. 20, 3533(1987)

[7]chen N. Zhu W Y.Bud-sequences conjecture on M fractal image and M-J conjecture between c and z planes from $z \leftarrow z\text{-}w+c$ ($w=a+bi$)[J] Computers &graphics1998 22(4) 537-546.

[8] Yan D J. Liu X D, Zhu W Y, A Study of Mandelbrot and Julia Sets Generated From a General Complex Cubic Iteration[J] Fractals 1999, 7(4)433-437

Xu Zhe is the student of School of Information and Engineering, Northeastern University. He graduated from Shenyang Polytechnics University in 1995. And in March 2002, he entered into School of Information and Engineering of NEU, majored in Chaos and Fractal.

Generating Algorithm of IAGO Generating Space

Xiao Xinping, Tang Weiqing
College of Sciences, Wuhan University of Technology
Yu Jia Tou, Campus Wuhan, Post code 430063, China
Email： xiaoxp@mail.whut.edu.cn **Tel：** 86-27-62825755

ABSTRACT

Grey generating provides a theoretical foundation of data processing in grey modeling. Based on mathematical system theory, the generating algorithm of IAGO generating space is developed, the inverse linear relation between r-AGO transformation and r-IAGO transformation is proved, other available results related to the relationship between r-IAGO series and the moving operator are also obtained, which are of benefit to grey modeling.

Keywords: AGO; r-IAGO; Negative power AGO; Moving operator

1. INTRODUCTION

Since the grey theory was proposed by Deng in 1982 [1], the grey model (GM) has been widely applied to many fields such as marketing, communications, power engineering, reliability engineering, image compression and so forth [2-4]. The basis of grey modeling is data handling, such as accumulated generating operation (AGO), inversed accumulated generating operation (IAGO), and class ratio generating operation. In the conventional contents of grey theory, studying the properties of the AGO series is of interest to researchers. Wen [5] obtained some properties of the AGO for invariant series, Dai [6] studied the frame of AGO generating space, we proposed the generalized results on r-AGO series in 2001 [7]. The negative power AGO is also a concept related to AGO and IAGO, Deng [8] discussed the negative power. In this paper, we take aim at the negative power AGO, r-IAGO series and moving operator, and study its properties.

2. MAIN RESULTS

Let $x^{(0)}, x^{(-r)}$ be series in raw and r-IAGO, i.e.,

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \Lambda , x^{(0)}(n))$$
$$x^{(-r)} = (x^{(-r)}(1), x^{(-r)}(2), \Lambda , x^{(-r)}(n))$$
$$x^{(-r)}(k) = x^{-(r-1)}(k) - x^{-(r-1)}(k-1)$$

and let $\Phi^{(1)}, \Phi^{(r)}$ be the generating basis matrix in IAGO and r-IAGO respectively,

$$\Phi^{(r)} = (\Phi_1^{(r)}, \Phi_2^{(r)}, \Lambda , \Phi_n^{(r)})^T \ \ (r = 1,2,\Lambda)$$

The moving operator is a new concept dealing with and converting the discrete events, for the digit α, suppose $\Delta(k)$ is a symbol to describe the existing in point *k*, $\Delta(k)\alpha$ implies that the data in *k* is α, C_n^m is the evolutional coefficient of k order , We suppose that

(1) $x^{(-r)}(k) = 0,\ for\ k \le 0;$

(2) $C_n^m = 0,\ for\ m < 0\ or\ m > n$

Theorem 1 $\Phi_k^{(r)}(i) = (-1)^{i-k} C_r^{i-k}$

Proof: Omitted.

Theorem 2 $\Phi_k^{(r)} = \sum_{m=1}^{k-1} \Delta(m)0 + \sum_{m=k}^{n} \Delta(m)(-1)^{m-k} C_r^{m-k}$

Proof: Omitted.

Theorem 3 $x^{(-r)}(k) = \sum_{m=0}^{j} (-1)^m C_j^m x^{-(r-j)}(k-m)$

Proof: To prove the theorem via complete induction.

First: To prove *j*=1. By the definition of $x^{(-r)}$, there is

$$x^{-(r)}(k) = x^{-(r-1)}(k) - x^{-(r-1)}(k-1)$$
$$= \sum_{m=0}^{1} (-1)^m C_1^m x^{-(r-1)}(k-m)$$

which shows that Theorem 3 is true for *j*=1.
Second: Suppose Theorem 3 holds for *j-1*.

$$x^{(-r)}(k) = \sum_{m=0}^{j-1} (-1)^m C_{j-1}^m x^{-(r-j+1)}(k-m)$$

Owing to

$$x^{-(r-j+1)}(k) = x^{-(r-j)}(k) - x^{-(r-j)}(k-1)$$

there are

$$x^{(-r)}(k) =$$
$$\sum_{m=0}^{j-1} (-1)^m C_{j-1}^m [x^{-(r-j)}(k-m) - x^{-(r-j)}(k-m-1)]$$
$$= \sum_{m=0}^{j-1} (-1)^m C_{j-1}^m x^{-(r-j)}(k-m)$$
$$- \sum_{m=0}^{j-1} (-1)^m C_{j-1}^m x^{-(r-j)}(k-m-1)$$
$$= \sum_{m=0}^{j-1} (-1)^m C_{j-1}^m x^{-(r-j)}(k-m)$$
$$- \sum_{m=1}^{j} (-1)^{m-1} C_{j-1}^{m-1} x^{-(r-j)}(k-m)$$
$$= C_{j-1}^0 x^{-(r-j)}(k) + \sum_{m=1}^{j-1} (-1)^m [C_{j-1}^m + C_{j-1}^{m-1}] x^{-(r-j)}(k-m)$$
$$+ (-1)^j C_{j-1}^{j-1} x^{-(r-j)}(k-j)$$
$$= C_j^0 x^{-(r-j)}(k) + \sum_{m=1}^{j-1} (-1)^m C_j^m x^{-(r-j)}(k-m)$$
$$+ (-1)^j C_j^j x^{-(r-j)}(k-j)$$
$$= \sum_{m=0}^{j} (-1)^m C_j^m x^{-(r-j)}(k-m)$$

This completes the proof of Theorem 3.

Theorem 4
$x^{(-r)}(k) = x^{-(r-j)}\Phi^{(j)}(k), or\ x^{(-r)} = x^{-(r-j)}\Phi^{(j)}$

Proof:

$$x^{-(r-j)}\Phi^{(j)}(k)$$
$$= (x^{-(r-j)}(1), x^{-(r-j)}(2), \Lambda\ , x^{-(r-j)}(n)) \cdot$$
$$(\Phi_1^{(j)}(k), \Phi_2^{(j)}(k), \Lambda\ , \Phi_n^{(j)}(n))^T$$
$$= \sum_{m=1}^{n} x^{-(r-j)}(m)\Phi_m^{(j)}(k)$$

Case 1 k-j>1
From Theorem 1, there is $\Phi_m^{(j)}(k) = (-1)^{k-m} C_j^{k-m}$. If $1 \le m \le k - j - 1,$ that is, $k - m \ge k - (k - j - 1) = j + 1$, then $C_j^{k-m} = 0$, which implies $\Phi_m^{(j)}(k) = 0$

Similarly, there are $\Phi_m^{(j)}(k) = (-1)^{k-m} C_j^{k-m}$ for $k - j \le m \le k$; $\Phi_m^{(j)}(k) = 0$ for $k+1 \le m \le n$.

We have

$$\sum_{m=1}^{n} x^{-(r-j)}(m)\Phi_m^{(j)}(k)$$
$$= \sum_{m=1}^{k-j-1} x^{-(r-j)}(m)\Phi_m^{(j)}(k) + \sum_{m=k-j}^{k} x^{-(r-j)}(m)\Phi_m^{(j)}(k) + \sum_{m=k+1}^{n} x^{-(r-j)}(m)\Phi_m^{(j)}(k)$$
$$= \sum_{m=k-j}^{k} x^{-(r-j)}(m)(-1)^{k-m} C_j^{k-m}$$
$$= \sum_{t=0}^{j} (-1)^t C_j^t x^{-(r-j)}(k-t) \quad (let\ k - m = t)$$
$$= x^{(-r)}k)$$

Case 2 $k \le j + 1$

$$\sum_{m=1}^{n} x^{-(r-j)}(m)\Phi_m^{(j)}(k)$$
$$= \sum_{m=1}^{k} x^{-(r-j)}(m)\Phi_m^{(j)}(k) + \sum_{m=k+1}^{n} x^{-(r-j)}(m)\Phi_m^{(j)}(k)$$
$$= \sum_{m=1}^{k} x^{-(r-j)}(m)\Phi_m^{(j)}(k)$$
$$= \sum_{m=1}^{k} (-1)^{k-m} C_j^{j-m} x^{-(r-j)}(m)$$
$$= \sum_{m=2}^{k} (-1)^{k-m} C_j^{j-m} x^{-(r-j)}(m) + \sum_{m=k-j}^{1} (-1)^{k-m} C_j^{j-m} x^{-(r-j)}(m)$$
$$= \sum_{m=k-j}^{k} (-1)^{k-m} C_j^{j-m} x^{-(r-j)}(m)$$
$$= \sum_{t=0}^{j} (-1)^t C_j^t x^{-(r-j)}(k-t) \quad (let\ k - m = t)$$
$$= x^{(-r)}k)$$

Theorem 5 $x^{(-r)} = \sum_{k=1}^{n} x^{(0)}(k) \sum_{m=k}^{n} \Delta(m)(-1)^{m-k} C_r^{m-k}$

Proof: Since $x^{(-r)} = x^{(0)}\Phi^{(r)} = \sum_{k=1}^{n} x^{(0)}(k)\Phi_k^{(r)}$ (Dai, 2001), we have

$$x^{(-r)} = \sum_{k=1}^{n} x^{(0)}(k)(\sum_{m=1}^{k-1}\Delta(m)0 + \sum_{m=k}^{n}\Delta(m)(-1)^{m-k} C_r^{m-k})$$
$$= \sum_{k=1}^{n} x^{(0)}(k) \sum_{m=k}^{n} \Delta(m)(-1)^{m-k} C_r^{m-k}$$

So we have

$$\sum_{k=1}^{n} x^{(0)}(k) \sum_{m=k}^{n} \Delta(m)(-1)^{m-k} C_r^{m-k}$$
$$= \sum_{k=1}^{n} x^{(0)}(k)[\Delta(k)(-1)^0 C_r^0 + \Delta(k+1)(-1)^1 C_r^1$$
$$+ \Lambda + \Delta(n)(-1)^{n-k} C_r^{n-k}]$$
$$= \sum_{k=1}^{n} x^{(0)}(k)\Delta(k)(-1)^0 C_r^0$$
$$+ \sum_{k=1}^{n} x^{(0)}(k)\Delta(k+1)(-1)^1 C_r^1$$
$$+ \Lambda + \sum_{k=1}^{n} x^{(0)}(k)\Delta(n)(-1)^{n-k} C_r^{n-k}$$

Due to $\Delta(n+i) = \phi$, $i = 1,2,\Lambda$, there are

$$x^{(-r)}(k) = \sum_{m=0}^{r} (-1)^m C_r^m x^{(0)}(k-m)$$

$$\sum_{k=1}^{n} x^{(0)}(k) \sum_{m=k}^{n} \Delta(m)(-1)^{m-k} C_r^{m-k}$$
$$= (-1)^0 C_r^0 x^{(0)}(1)\Delta(1) + [(-1)^1 C_r^1 x^{(0)}(1)$$
$$+ (-1)^0 C_r^0 x^{(0)}(2)]\Delta(2) + \Lambda + [(-1)^{n-1} C_r^{n-1} x^{(0)}(1) +$$
$$(-1)^{n-2} C_r^{n-2} x^{(0)}(2) + \Lambda + (-1)^0 C_r^0 x^{(0)}(n)]\Delta(n)$$
$$= x^{(-r)}(1)\Delta(1) + x^{(-r)}(2)\Delta(2) + \Lambda + x^{(-r)}(n)\Delta(n)$$
$$= (x^{(-r)}(1), x^{(-r)}(2), \Lambda\ , x^{(-r)}(n))$$

Let $X^{(r)}$ be the series of r-IAGO of raw series $X^{(0)}$, $\psi_k^{(r)}$ be the basis vector of r-AGO generating space, then we can obtain the following similar results:

Theorem 6

(1) $\psi_k^{(r)} = \sum_{m=1}^{k-1}\Delta(m)0 + \sum_{m=k}^{n}\Delta(m)C_{m-k+r-1}^{r-1}$

(2) $x^{(r)} = \sum_{k=1}^{n} x^{(0)}(k) \sum_{m=k}^{n} \Delta(m)C_{m-k+r-1}^{r-1}$

(3) $x^{(r)}$ = r-AGO ($x^{(0)}$), $x^{(-r)}$ = r-IAGO ($x^{(0)}$)

Proof: Omitted.

Theorem 7 Let R be the real set, then r-AGO and r-IAGO are mutually inverse linear transformation in linear space $R^{(n)}$.

Proof: For any $x^{(0)}, y^{(0)} \in R^{(n)}, k, l \in R$, we have

r-AGO ($kx^{(0)} + ly^{(0)}$)

= r-AGO $z^{(0)}$ = $z^{(r)}$ = $z^{(0)}\Psi^{(r)}$

= $[kx^{(0)} + ly^{(0)}]\Psi^{(r)}$

$= kx^{(0)}\ \Psi^{(r)} + ly^{(0)}\ \Psi^{(r)}$

$= kx^{(r)} + ly^{(r)}$

$= k$ r-AGO ($x^{(0)}$) $+ l$ r-AGO ($y^{(0)}$)

This means that r-AGO is a linear transformation in $R^{(n)}$. Similarly we can prove that r-IAGO is also a linear transformation in $R^{(n)}$.

By Theorem 6, there is

(r-IAGO · r-AGO) $x^{(0)}$

= r-IAGO [r-AGO ($x^{(0)}$)]

= r-IAGO ($x^{(r)}$)

= r-IAGO ($x^{(0)}\ \Psi^{(r)}$)

= [r-IAGO ($x^{(0)}$)] $\Psi^{(r)}$

$= x^{(-r)}\ \Psi^{(r)}$

$= x^{(0)}\ \Phi^{(r)}\ \Psi^{(r)}$

$= x^{(0)}$

So we have

r-AGO · r-IAGO = **E,**

r-IAGO = (r-AGO) $^{-1}$.

where **E** is the unit transformation.

3. CONCLUSIONS

The properties of the negative power AGO and r-IAGO series are discussed, the results related to the relationship between r-IAGO series and the moving operator is obtained. All of these are of benefit to revealing the latent law of data handling in grey modeling.

Acknowledgement The authors are very grateful to Prof. Liu S.F. for many suggestions concerning the manuscript.

4. REFERENCES

[1] Deng Julong, Course on Grey System Theory, Wuhan: HUST Press , 1990

[2] Jou JM, Shiau YH. A low-cost gray prediction search chip for motion estimation. IEEE Transactions on Circuits and Systems, Vol.49, No.7, 2002, pp. 928-938

[3] Su S F, Lin C B, Hsu YT. A high precision global prediction approach based on local prediction approaches. IEEE Transactions on Systems Man and Cybernetics, Vol.32, No.4, 2002, pp. 416-425

[4] Lin, Chan-Ben; Su, Shun-Feng; High-precision forecast using grey models, International Journal of Systems Science, Vol.32, No.5, 2001, pp. 609-619

[5] Kun-Li Wen, AGO for invariant series. The Journal of Grey System, Vol.10, No.1, 1998, pp.17~21

[6] Dai Debao, Chen Rongqiu et al. Frame of IAGO Generating Space. The Journal of Grey System, Vol.13, No.1, 2001, pp.9~12

[7] Xiao Xinping, Li Feng, Novel Results On AGO Space. The Journal of Grey System, Vol.13, No.4, 2001, pp.325~330

[8] Deng Julong, Moving operator in grey theory. The Journal of Grey System, Vol.11, No.1, 1999, pp.1~5

Xiao Xinping is a Full Professor and deputy dean of College of Science, Wuhan University of Technology. He graduated from Huazhong University of Science and Technology and got a PH.D in 2002. He was the co-chair of CGSTA 1996 and will be the chairman of CYORM 2004. He has published four books, over 50 Journal papers. His research interests are in system engineering, graph theory and control technology.

Type Checking for Software System Specifications in Real-Time Process Algebra

Chuanwen Liu Xinming Tan
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei 430063, China
Email: chwliu@mail.whut.edu.cn tanxm@mail.whut.edu.cn **Tel.:** +86-27-86551711

ABSTRACT

This paper presents the development of a type checker for formal specifications of software systems described in Real-Time Process Algebra (RTPA). The grammar of RTPA is formally described by using the EBNF convention. Design and implementation techniques are presented that keep the RTPA syntax as close to its original mathematical notations as possible, and allows it be easily parsed as well.

The tasks of type checking for RTPA specifications can be classified into three categories: (a) identifier type compliancy, (b) expression type compliancy, and (c) process constraint consistency. The RTPA type checker has been designed and implemented to support system architects and system analysts to ensure the correctness and consistency of system specifications to a maximum extent. And it shows that with the grammar defined, the type checker can be integrated with the parser and do both parsing and type checking in one pass.

Keywords: Software engineering, formal methods, RTPA, parser, type checker, code generation.

1. INTRODUCTION

The Real-Time Process Algebra (RTPA) is designed to describe real-time and distributed systems [6, 7]. As a mathematical notation system, RTPA possesses sixteen built-in types, sixteen meta-processes, and sixteen process relations. The fundamental concepts of RTPA are processes and component logic models (CLMs). System architectures are modeled by the CLMs, while system static and dynamic behaviors are described by the processes. RTPA has been used to specify large scale real-time and distributed systems [5, 7-9]. To accelerate applications of RTPA in practical use, an RTPA-based code generator project has been established [4, 10].

This paper presents the design and implementation of a type checker for formal specifications of software systems described in RTPA. The paper starts with the description of the RTPA grammar, then the type checking for RTPA syntax, and finally the implementation of the type checker.

The grammar of RTPA is formally described by using the EBNF convention adopted by ANTLR [3]. The EBNF description of the RTPA grammar is a foundation for the design and implementation of the RTPA type checker. Three categories of functions in type checking for RTPA specifications, known as (a) identifier type compliancy, (b) expression type compliancy, and (c) process constraint consistency, have been studied. The first category of type checking verifies defining and applied occurrences of identifiers and their scopes. It guarantees that every identifier is uniquely defined in its scope and every reference is applied to defined identifier. For expression type check, each expression is assigned with a result type that is derived from its components and their relations. Type compliancy among expressions in the context of a RTPA process is verified, in order to prevent operations between incompliant expressions and variables. The process constraint check is not considered as a traditional part of type checking, but is a feature of RTPA for verifying process constraints based on defined process combination rules. For example, a process should not be put onto the base level and interrupt level in the same system.

The RTPA type checker has been implemented with ANTLR and Java in such a way that both parsing and type checking can be carried out in one pass and to support system architects and analysts for developing consistent and correct specifications of large-scale real-time and distributed systems.

2. THE GRAMMAR OF RTPA

The grammar of RTPA is specified using the EBNF convention adopted by ANTLR. The grammar rules of RTPA can be grouped into four groups in term of syntax categories, namely architecture rules, static behaviors rules, dynamic behaviors rules, and expression rules.

2.1. The EBNF Definition of the RTPA Grammar

A system specification in RTPA specifies the system architecture, static behaviors, and dynamic behaviors of a given system as defined in rule *R1* below.

```
rtpa_specification
:       top_schema    architecture    static_behaviors
        dynamic_behaviors    EOF;
                                                      (R1)
```

The first part of a specification is the top_schema as defined in rule *R2*.

```
top_schema
:       system_declaration DEFINITION_SYMBOL
        system_architecture_declaration
        PARALLEL_SYMBOL
        system_static_behaviors_declaration
        PARALLEL_SYMBOL
        system_dynamic_behaviors_declaration
;                                                  (R2)
```

The system declaration defines the system's name while the system architecture declaration, system static behaviors declaration, and system dynamic behaviors declaration introduce respective specification parts.

There are over hundreds of rules for the complete definition of the grammar of RTPA. Only those high level and RTPA featured rules are given here due to the limited space.

2.1.1. RTPA Grammar Rules for System Architecture Specifications: The architecture of a system comprises subsystem declaration, component logic models declaration, and component logic model schemata declaration, as shown in rule *R3*.

```
architecture
:       system_architecture_declaration
        DEFINITION_SYMBOL
        (subsystem_declaration)? component_logic_models
        clm_schemata
;                                                   (R3)
```

The subsystem declaration, as shown in rule *R4-7*, defines how many subsystems in the system and their structural relations.

```
subsystem_declaration
:       (subsystem_expression    EQUALITY_SYMBOL)+
;                                                   (R4)
subsystem_expression
:       simple_subsystem_expression    (subsystem_relation
        simple_subsystem_expression)*
;                                                   (R5)
simple_subsystem_expression
:       system_name
|       LEFT_PARENTHESIS subsystem_expression
        RIGHT_PARENTHESIS
;                                                   (R6)
subsystem_relation
:       PARALLEL_SYMBOL
|       SEQUENCE_SYMBOL
|       PIPELINE_SYMBOL
|       CALL_SYMBOL
;                                                   (R7)
```

The component logic models declaration lists all the CLMs and their respective numbers used in the system as shown in rule *R8-10*.

```
component_logic_models
:       clm_declaration    (PARALLEL_SYMBOL
        clm_declaration)*
;                                                   (R8)
clm_declaration
:       simple_clm_declaration    (PARALLEL_SYMBOL
        simple_clm_declaration)*
;                                                   (R9)
simple_clm_declaration
:       LEFT_ANGLE_BRACKET      clm_name
        (BAR_SYMBOL    LEFT_SQUARE_BRACKET
        DECIMAL RIGHT_SQUARE_BRACKET)?
        RIGHT_ANGLE_BRACKET
|       LEFT_PARENTHESIS clm_declaration
        RIGHT_PARENTHESIS
;                                                   (R10)
```

The component logic model schemata rules define the details of all the CLMs, constants, statuses, and events used in the system by the clm schemas, access model schema, constant schema, event schema, and status schema respectively, as shown in rule *R11-22*.

```
clm_schemata
:       clm_schemas    (access_model_schema)?
        (constant_schema)?    event_schema    status_schema
;                                                   (R11)
clm_schemas
:       (clm_schema    (EQUALITY_SYMBOL
clm_object
        (PARALLEL_SYMBOL      clm_object)*)? )+
;                                                   (R12)
```

The clm schema describes the clm's structure using *record*, which is a collection of *fields*; while the clm object is used to initialize the clm, as shown in rule *R13-16*. The access model schema is used to abbreviate accessing dynamically allocated objects in a form of starting point (represented by a record name) plus an offset (calculated from fields of type *pointer*), as shown in rule *R17*. The constant schema defines all the symbolic natural and Boolean constants used in the system as shown in rule *R18*.

```
clm_schema
:       (IDENTIFIER DOT_SYMBOL "Architecture"
        DOT_SYMBOL )?
        clm_name DEFINITION_SYMBOL record
;                                                   (R13)
field
:       (iteration_expression)? LEFT_ANGLE_BRACKET
        field_name    (BAR_SYMBOL      logical_expression)?
        RIGHT_ANGLE_BRACKET
;                                                   (R14)
record
:       (iteration_expression)?    record_name
        COLON_SYMBOL
        LEFT_PARENTHESIS field (COMMA_SYMBOL
        field)* RIGHT_PARENTHESIS
;                                                   (R15)
clm_object
:       clm_name    COLON_SYMBOL      tuple_assignment
;                                                   (R16)
```

```
access_model_schema
:       (IDENTIFIER      DOT_SYMBOL     "Architecture"
        DOT_SYMBOL!)?
        "AccessModel" SYSTEM_SUFFIX
        DEFINITION_SYMBOL      record_name
        LEFT_PARENTHESIS      field    (COMMA_SYMBOL
        field)*    RIGHT_PARENTHESIS       type_suffix
;                                                   (R17)
```

```
constant_schema
:       (IDENTIFIER      DOT_SYMBOL     "Architecture"
        DOT_SYMBOL)?
        "Constant"      SYSTEM_SUFFIX
        DEFINITION_SYMBOL    single_assignment
        (BAR_SYMBOL    single_assignment)*
;                                                   (R18)
```

The event schema defines all the events used in the system as shown in rule *R19-20*.

```
event_schema
:       (IDENTIFIER      DOT_SYMBOL     "Architecture"
        DOT_SYMBOL)?
        "Event"      SYSTEM_SUFFIX
        DEFINITION_SYMBOL
        event    (BAR_SYMBOL      event)*
;                                                   (R19)
event
:       AT_SYMBOL    IDENTIFIER    STRING_SUFFIX
;                                                   (R20)
```

The status schema defines all the statuses used in the system as shown in rule *R21-22*.

```
status_schema
:       (IDENTIFIER      DOT_SYMBOL     "Architecture"
DOT_SYMBOL!)?
"Status"      SYSTEM_SUFFIX      DEFINITION_SYMBOL
status    (BAR_SYMBOL      status)*
;                                                   (R21)
status
:       AT_SYMBOL    IDENTIFIER    BOOLEAN_SUFFIX
;                                                   (R22)
```

2.1.2. RTPA Grammar Rules for System Static Behaviors

Specifications: The static behaviors part of a system consists of process declarations and process definitions for all the processes in the system as shown in rule *R23-25*.

```
static_behaviors
:       process_declarations    process_definitions
;                                                        (R23)
process_declarations
:       system_static_behaviors_declaration
        DEFINITION_SYMBOL
        process_declaration    (PARALLEL_SYMBOL
        process_declaration    )*
;                                                        (R24)
process_definitions
:       (process_definition)*
;                                                        (R25)
```

A process declaration describes the interface of the process, which includes the process name, input and output arguments, and operated CLMs, as shown in rule *R26-31*.

```
process_declaration
:       process_name    LEFT_PARENTHESIS
        formal_input_output    RIGHT_PARENTHESIS
;                                                        (R26)
formal_input_output
:       formal_input    SEMICOLON_SYMBOL    formal_output
        SEMICOLON_SYMBOL    formal_clm
;                                                        (R27)
formal_input
:       LEFT_ANGLE_BRACKET
        INPUT_ARGUMENT_SYMBOL
        (formal_arguments)?    RIGHT_ANGLE_BRACKET
;                                                        (R28)
formal_output
:       LEFT_ANGLE_BRACKET
        OUTPUT_ARGUMENT_SYMBOL
        (formal_arguments)?    RIGHT_ANGLE_BRACKET
;                                                        (R29)
formal_clm
:       LEFT_ANGLE_BRACKET
        CLM_ARGUMENT_SYMBOL
        (formal_arguments)?    RIGHT_ANGLE_BRACKET
;                                                        (R30)
formal_arguments
:       ((iteration_expression)?    variable | status)
        (COMMA_SYMBOL
        ((iteration_expression)?    variable | status))*
;                                                        (R31)
```

A process definition, as shown in rule *R32-36*, describes the behavior of the process apart from its interface.

```
process_definition
:       (IDENTIFIER    DOT_SYMBOL    "StaticBehaviors"
        DOT_SYMBOL! )?    process_declaration
        DEFINITION_SYMBOL
        LEFT_CURLY_BRACKET    process_expression
        RIGHT_CURLY_BRACKET
;                                                        (R32)
process_expression
:       factor_process    (sequencial_process_relation
        factor_process)*
;                                                        (R33)
factor_process
:       simple_process    (concurrent_process_relation
        simple_process)*
;                                                        (R34)
sequencial_process_relation
:       SEQUENCE_SYMBOL
|       CALL_SYMBOL
|       RECURSION_SYMBOL
|       PIPELINE_SYMBOL
;                                                        (R35)
```

```
concurrent_process_relation
:       PARALLEL_SYMBOL
|       CONCURRENCE_SYMBOL
|       INTERLEAVE_SYMBOL
|       INTERRUPT_SYMBOL
;                                                        (R36)
```

The simple processes, as shown in rule *R37*, are syntactic building unit for constructing compound processes by using process relations.

```
simple_process
:       system_process
|       single_assignment
|       tuple_assignment
|       addressing_process
|       input_process
|       output_process
|       read_process
|       write_process
|       timing_process
|       duration_process
|       memory_allocation_process
|       memory_release_process
|       increase_process
|       decrease_process
|       exception_detecting_process
|       skip_process
|       stop_process
|       branch_process
|       switch_process
|       for_process
|       repeat_process
|       while_process
|       process_instance_expression
|       name_process
|       jump_process
|       time_driven_dispatch_process
|       event_driven_dispatch_process
|       interrupt_expression
|       labelled_process
|       LEFT_CURLY_BRACKET    process_expression
        RIGHT_CURLY_BRACKET
;                                                        (R37)
```

2.1.3. RTPA Grammar Rules for System Dynamic Behaviors Specifications:

The dynamic behaviors part of a system consists of the processes classification, processes deployment, and processes dispatch, as shown in rule *R38-41*.

```
dynamic_behaviors
:       (process_classification)?    (process_deployment)?
        (process_dispatch)?
;                                                        (R38)
process_classification
:       system_dynamic_behaviors_declaration
        DEFINITION_SYMBOL
        LEFT_CURLY_BRACKET    process_expression
        RIGHT_CURLY_BRACKET
;                                                        (R39)
process_deployment
:       process_deployment_declaration
        DEFINITION_SYMBOL
        LEFT_CURLY_BRACKET    process_expression
        RIGHT_CURLY_BRACKET
;                                                        (R40)
```

```
process_dispatch
:       process_dispatch_declaration
        DEFINITION_SYMBOL
        LEFT_CURLY_BRACKET    process_expression
        RIGHT_CURLY_BRACKET
;                                              (R41)
```

The processes classification puts all the processes in the system into four scheduling priority groups: *base level*, *high level*, *low interrupt level*, and *high interrupt level*. The processes deployment describes how the processes are deployed at the run time. The processes dispatch specifies which event will trigger which process.

2.1.4. RTPA Grammar Rules for Expressions Specifications: The rules in this part describe all the expressions used in the system.

```
expression
:       string_expression
|       time_expression
|       numerical_expression
|       logical_expression
;
numerical_expression
:       integer_expression
|       real_expression
|       pointer_expression
|       subscript_expression
|       real_expression
|       real_expression
;
logical_expression
|       relational_expression
|       boolean_expression
;
relational_expression
:       set_relational_expression
|       numerical_relational_expression
|       time_relational_expression
```

Due to the limited space, only gives the expression categories above not the actual grammar rules.

2.2. RTPA Types

RTPA is strongly typed in the sense that each identifier in a RTPA expression possesses a suffix as its type indicator, and operations crossing maximal types are not allowed. RTPA predefines 16 built-in data types and 19 type suffixes. In order to do expression type checking, maximal types are introduced.

2.2.1. RTPA Data Types: RTPA has 16 built-in data types for modeling data, events, and architectures, as shown in Table 1.

Table 1 RTPA Data Types

No.	Data Type	Syntax
1	Integer	N
2	Real	R
3	String	S
4	Boolean	BL
5	Byte	B
6	Hexadecimal	H
7	Pointer	P
8	Short Time	hh:mm:ss
9	Long Time	hh:mm:ss:ms
10	Date	yyyyy:mm:dd
11	Short DateTime	yyyyy:mm:dd:hh:mm:ss
12	Long DateTime	yyyyy:mm:dd:hh:mm:ss:ms
13	Run-time Determinable Type	RT
14	System Architecture Type	ST
15	Event	@S
16	Status	@BL

2.2.2. RTPA Type Suffixes: RTPA uses 19 type suffixes (Table 2) to indicate the types of all identifiers. In addition to the 16 data types, three other type suffixes are introduced to denote symbolic constants. More accurately, prefix and suffix are combined to identify events and statuses in consent to the original RTPA notation convention. For example, a system event *TimeOut* can be denoted as @TimeOut_S.

Table 2 RTPA Type Suffixes

No.	Type	Type Suffix
1	Integer	_N, _C (used for natural constants)
2	Real	_R
3	String	_S
4	Boolean	_BL, _T (used for Boolean constant true), _F (used for Boolean constant false)
5	Byte	_B
6	Hexadecimal	_H
7	Pointer	_P
8	Short Time	_hh:mm:ss
9	Long Time	_hh:mm:ss:ms
10	Date	_yyyyy:mm:dd
11	Short DateTime	_yyyyy:mm:dd:hh:mm:ss
12	Long DateTime	_yyyyy:mm:dd:hh:mm:ss:ms
13	Run-time Determinable Type	_RT
14	System Architecture Type	_ST
15	Event	@ _S (@ used as prefix)
16	Status	@ _BL (@ used as prefix)

2.2.3. RTPA maximal types: RTPA maximal types are introduced to help type checking. There are nine maximal types defined for RTPA (Table 3). Operations within the same maximal type are allowed while operations crossing maximal types are forbidden unless explicit type casting is carried out. The only exception is when the Run-time Determinable Type is operated with another maximal type; the result maximal type is determined by that non-RT maximal type. When determining the types of expressions during expression type check, all RTPA data types involved will be converted to their respective maximal types.

Table 3 RTPA Maximal Types

No.	Maximal Type	Types will converted to
1	Integer	N, B, H, P
2	Real	R
3	String	S
4	Boolean	BL
5	Date	hh:mm:ss, hh:mm:ss:ms, yyyyy:mm:dd, yyyyy:mm:dd:hh:mm:ss, yyyyy:mm:dd:hh:mm:ss:ms
6	Run-time Determinable Type	RT
7	System Architecture Type	ST
8	Event	@S
9	Status	@BL

3. TYPE CHECKING TECHNOLOGIES FOR RTPA

This section describes techniques and algorithms of the RTPA type checker. Type compliance constraints and checking rules for identifiers, expressions, and processes will be defined and explained.

3.1. Identifier Type Compliance

Identifiers are used to represent process names, variables, and constants in RTPA specifications. The identifier type compliance checks are used to maintain the rules of both identifier declarations and usages.

3.1.1. The Scope of Identifiers: An identifier in RTPA specifications is only visible and valid within a defined scope. The identifier identity is defined as of the same name, type, and scope. Identifiers with the same name and type would be different if they are defined in different scopes. In this case, the identifier with a smaller scope will override the identifier with a larger scope. There are four kinds of scopes specified in RTPA known as the *global*, *system*, *CLM*, and *process* scopes.

- ***Global*** – visible in all different system specifications. System names are of global scope.
- ***System*** – visible in the whole system specification in which it be defined. Subsystem names, CLM names, and process names are of system scopes.
- ***CLM*** – visible in the CLM in which it be defined. Fields are of CLM scopes.
- ***Process*** – visible in the named process in which it be defined. Variables defined in named processes are of process scopes. Named processes are those defined in the static behaviors part.

3.1.2. Checking Identifier Type Compliance: Identifiers in RTPA specifications should be declared and then used. Identifiers may be classified as *defined*, *undefined*, or *redundantly defined* in a given scope. Only defined identifiers can pass the compliance checking for identifier types.

The algorithm of identifier type compliance checking for RTPA can be described as shown in Algorithm 1.

Algorithm 1. Identifier Type Compliance Checking

```
(a)  Create an identifier table to hold the information
     of name, type, and scope of each identifier in a
     given specification.
(b)  Parse through the specification, change the
     scope of identifiers as the context changes, and
     scan each identifier and identify its type:
     (i)   When a declaration of an identifier is
           recognized, check if there is an identical
           identifier that has already been registered
           in the identifier table. If so, it is redundant;
           otherwise, it is a newly declared identifier,
           and its name, type, and scope have to be
           added into the identifier table.
     (ii)  When an identifier is referenced, check it
           up in the identifier table. If it has not been
           defined in the table, it is an invalid
           reference to an undefined identifier.
(c)  Repeat Step (b) until the end of the specification
     has been reached.
```

3.2. Expression Type Compliance

Type compliance checking for expressions focuses on the following syntactic problems: (a) The types of identifiers in an expression should be compliant with each other, and (b) The left-hand-side (LHS) expression should be compliant with the right-hand-side (RHS) expression in an assignment-like process.

3.2.1. Checking Expression Type Compliance: Expression type checking assesses if the types of identifiers in an expression are compatible and can be resolved to a definite maximal type, which means the calculation of value of the expression can be carried out. The algorithm for calculating the maximal type of an expression in RTPA can be described as shown in Algorithm 2.

Algorithm 2. Expression Type Compliance Checking

```
(a)  Parse through the expression, set the maximal
     type of the first identifier in the expression as
     the type of the expression.
(b)  For each following identifier in the expression:
     (i)   Conduct any type casting if it is explicitly
           specified, and Change the type to the
           corresponding maximal type;
     (ii)  Compare its type with the type of the
           expression. If they are not the same types,
           report a type conflict; otherwise, continue;
     (iii) Set the resulted type as Boolean when a
           relational expression is parsed.
(c)  Repeat Step (b) until the end of the expression.
```

3.2.2. Checking Assignment Type Compliance: Assignment type compliances need to be checked in RTPA assignment-like processes, such as assignment, input, output, read, write, timing, duration, and function call. A general pattern of assignments is that the LHS of an assignment is a variable and the RHS is an expression. It is required that both sides of an assignment should be type compatible and can be resolved to a definite maximal type. An exceptional case is that when the type of one side of an assignment is **RT**, it will be always succeed.

For a given assignment-like process, the maximal type of the variable or expression at its LHS is compared for type compatibility with the maximal type of the expression at its LHS. The expressions will be evaluated by Algorithm 2 as defined in Section 3.2.1.

For a function call process, the identifier table discussed in Section 3.1.2 needs to be extended. Additional information for identifiers' syntactic category or kind [2], such as input and output arguments, should be identified, in order to match the actual arguments against the formal arguments in the process interface.

3.3. Process Constraint Consistency

Three categories of process constraints in RTPA specifications have been identified known as the *interna*l, *relational*, and *dynamic* constraints. Checking consistencies of these constraints can greatly help to ensure the correctness of RTPA specifications.

3.3.1. Checking the Consistency of Internal Process Constraints: There are two situations for checking internal process constraints as follows:

- ***CLM reference constraints*** - CLMs operated in a process should be consistent with those declared in the process' interface.
- ***Event-driven process constraints*** - Events used in an event-driven process should not be duplicated.

The consistency check for the former is to match each CLM used in the process against the CLMs declared in the interface of the process. When the process definition is scanned, all CLM declarations will be collected and held in a table. Then, when analyzing the process body, all identifiers will be checked against the CLMs registered in the table.

The consistency check for the event-driven process constraints is to detect if any duplicated event exists in the event-drive processes.

3.3.2. Checking the Consistency of Relational Process Constraints: Process relations are used in RTPA to connect two or more processes in order to form a complex process. The relational process constraints can be classified as follows:

- When applying the *function call* relation to two processes, the second process should be a different named process.
- When applying the *recursion relation* to two processes, they should be identical named processes.
- When applying the *concurrence, parallel*, or *interleave relation* to two processes, they should be named processes.
- When applying the *pipeline relation* to two processes, they should share the same CLMs as declared in their interfaces.

To check the above consistencies, two adjacent processes need to be checked for what kind of processes they are, and what relation in between them is defined by using program variables when parsing the process_expression rule. If one of the RTPA relations as listed above is identified, corresponding constraints should be checked by verifying the statuses of the variables. The pipeline relation between a pair of processes needs special attention, for which both process interfaces have to be checked as that of matching actual arguments against formal ones as discussed in Section 3.2.2.

3.3.3. Checking the Consistency of Dynamic Process Constraints: In the dynamic behavior specification of RTPA, duplications of processes in process classification and duplication of events in process dispatching should be avoided.

The technologies used to check the dynamic process consistency are the same as that of event-driven processes as discussed in Section 3.3.1.

4. IMPLEMENTATION OF THE TYPE CHECKER

Based on the rules, algorithms, and techniques developed in the previous sections, the type checker for RTPA is implemented by using ANTLR and Java as illustrated in Fig.1.

Corresponding to Fig. 1, the RTPA type checker has been developed in the following processes.

(a) To define the RTPA grammar in EBNF.
(b) To convert the RTPA grammar into a set of LL(k) [1] parsing rules, which can be used to generate the RTPA parser. There are rules that can not be described by LL(k). Syntactic predicates of ANTLR have been used to make them determinable within a fixed depth of look-ahead [3].
(c) To develop a set of Java programs to implement the RTPA type checker, in terms of those algorithms discussed in Section 3.
(d) To embed the checker programs into the RTPA parser rules as special semantic actions.
(e) To load the RTPA parser rules with type checking actions to ANTLR in order to generate the RTPA type checker that results in the RTPA *lexer* and *type checker* in executable Java classes.

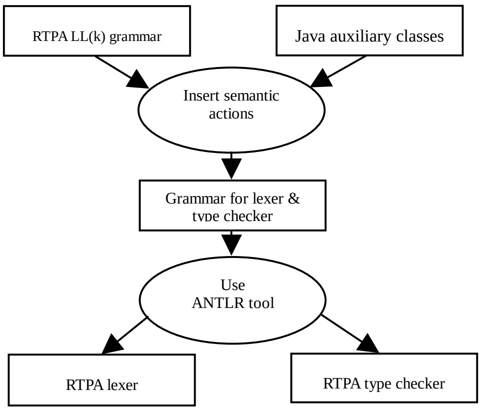


Fig. 1 The building process of the RTPA type checker

5. CONCLUSIONS

This paper has described the development of the type checker for Real-Time Process Algebra. The RTPA grammar has been defined in such a way that makes the RTPA syntax as close to its original mathematical notations as possible; and at the same time, allows it be easily parsed and transformed. The design and implementation of the RTPA type checker has not only achieved the traditional type checking functions for a formal language in traditional sense, but also the specific consistency checking of RTPA process constraints in a way that fulfills both parsing and type checking together in one pass.

Future work will be focused on identifying and implementing more process constraints for the RTPA type checker. On this basis, a fully developed tool for automatic code generation from the formal RTPA specifications to C++ or Java code will be completed.

6. REFERENCES

[1] A.V. Aho, R. Sethi, and J.D. Ullman, Compilers: Principles, Techniques, and Tools, Addison Wesley, 1986.
[2] D. Grune, H.E. Bal, C.J.H. Jacobs, and K.G. Langendoen, Modern Compiler Design, John Wiley & Sons, Ltd, England, 2000.
[3] T. Parr, ANTLR, http://www.antlr.org/, Jan. 2003.
[4] X. Tan and Y. Wang, “Specification of the RTPA Grammar and Its Recognition”, Proc. of the 2004 IEEE International Conference on Cognitive Informatics (ICCI'04), IEEE CS Press, Canada, August, 2004.
[5] X. Tan and Y. Wang, “Specification of Abstract Data Types using Real-Time Process Algebra (RTPA)”, Proc. of the 2003 Canadian Conference on Electrical and Computer Engineering (CCECE’03), IEEE CS Press, Montreal, Canada, May, 2003, pp.1293-1296.
[6] Y. Wang, “The Real-Time Process Algebra (RTPA)”, The International Journal of Annals of Software Engineering, Vol.14, USA, 2002, pp. 235-274.

[7] Y. Wang, "Real-Time Process Algebra (RTPA) and Its Applications", in B.K. Aichernig and T. Maibaum eds., Formal Methods at the Cross Roads: From Panacea to Foundational Support, Springer LNCS 2757, Berlin, 2003, pp.322-336.

[8] Y. Wang and C.F. Noglah, "Formal Description of Real-Time Operating Systems using RTPA", Proc. of the 2003 Canadian Conference on Electrical and Computer Engineering (CCECE'03), IEEE CS Press, Montreal, Canada, May, 2003, pp.1247-1250.

[9] Y. Wang and C.F. Noglah, "Formal Specification of a Real-Time Lift Dispatching System", Proc. of the 2002 IEEE Canadian Conference on Electrical and Computer Eng. (CCECE'02), Winnipeg, Canada, May, 2002, Vol. 2, pp. 669-674.

[10] J. Zhao and Y. Wang, "Design of a Parser for Real-Time Process Algebra (RTPA)", Proc. of the 2003 Canadian Conference on Electrical and Computer Engineering (CCECE'03), IEEE CS Press, Montreal, Canada, May, 2003, pp. 1259-1262.

Chuanwen Liu is a lecturer and a PhD candidate of the School of Computer Science and Technology, Wuhan University of Technology. He graduated from Wuhan University of Technology in 1992 with his BSc and in 1999 with his MSc both in computer science and its application. His research interests are in distributed systems, e-commence, and web application development.

A Definition and Study of a New Kind of Similarity Measure of Fuzzy Sets *

Tong Xiaojun[1], Gao Zunhai[2], Yuan Zhiyong[3]
[1,2]Department of Mathematics and Physics, Wuhan Polytechnic University
Wuhan, Postcode 430023, China
[1,2,3]Department of Control and Engineering,
Huazhong University of Scienceand Technology
Wuhan, Hubei, Postcode 430074, China
Email: tongxiaojun1998@yahoo.com.cn **Tel.:** +86(0)2762097718(+862783937411)

ABSTRACT

It is well known that the similarity measure can be induced from entropy of fuzzy sets, we prove that this existing kind of similarity measure all satisfies the proposition, that is, the similarity measure of two fuzzy sets equals to that of their complementary sets. But the mutual subset hood does not content this proposition. We construct a new kind of similarity measure depending on entropy, and it doesn't meet this proposition.

Keywords: fuzzy set, similarity measure, entropy.

1. INTRODUCTION

P. Z. Wang introduced the definition of similarity measure scaling the equality degree between two fuzzy sets [1]; it is used widely in pattern recognition, clustering analysis, Image and information process. Zeng brought forward more logical definition and study the character and the retaliation between similarity and entropy. Kosko induced the mutual subsethood theorem based on the factor of proportionality; in fact, this mutual subsethood is a similarity measure. Wang [5] showed the limitation of the calculating formula suggested by Kosko by the counterexample according to l^p -distance, and then proposed the modified formula. And other scholars study similarity measure by other idea, for example, Liu proposed the definition of $\sigma -$ similarity measure and $\sigma -$ entropy. Lin [10] directly brought l^1 -distance into the definition of similarity measure and gave a series of properties of this similarity measure. He renewedly has upbuilded the formula from σ -entropy to σ -similarity measure according on l^1 -distance. If we analyses the arithmetic of similarity measure $S(A,B)$, it can see, the arithmetic has the follow recourses: one is the method by using l^p （or L^p ） distance to construct. For example,

$$S(A,B)=1-\frac{1}{n}(\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|^p)^{1/p} \qquad (1)$$

Another is based on the factor of proportionality. Kosko and Wang obtain following similarity measures respectively.

$$S(A,B)=\frac{\sum Count(A\cap B)}{\sum Count(A\cup B)} \qquad (2)$$

$$S(A,B)=\frac{1}{n}\sum Count(\frac{A\cap B}{A\cup B}) \qquad (3)$$

The third is from entropy to construct. For example:

$$S(A,B)=e(\frac{A-B}{2}+F) \qquad (4)$$

Where $e(A)$ is an entropy of fuzzy set A.

Liu give the follow calculating method:

Let $X=\{x_1,x_2,\Lambda,x_n\}$, $A,B,C\in F(X)$, $F(X)$ is the class of all fuzzy sets of X , $\mu_A(x)$ is the membership function of the fuzzy set A. For $A,B\in F(X)$, define $D_i\in P(X)(i=1,2,3,4)$ as

$$D_1=\{x\in X;\mu_A(x)\geq\frac{1}{2},\mu_B(x)\geq\frac{1}{2}\},$$

$$D_2=\{x\in X;\mu_A(x)\geq\frac{1}{2},\mu_B(x)<\frac{1}{2}\},$$

$$D_3=\{x\in X;\mu_A(x)<\frac{1}{2},\mu_B(x)\geq\frac{1}{2}\},$$

$$D_4=\{x\in X;\mu_A(x)<\frac{1}{2},\mu_B(x)<\frac{1}{2}\}$$

Then define $R_i\in F(X)(i=1,2,3,4,5,6,7,8)$ as

$$\mu_{R_1}(x)=\begin{cases}\mu_A(x)\vee\mu_B(x), & x\in D_1\\ \frac{1}{2}, & x\notin D_1\end{cases};$$

$$\mu_{R_2}(x)=\begin{cases}\mu_A(x)\wedge\mu_B(x), & x\in D_1\\ \frac{1}{2}, & x\notin D_1\end{cases};$$

$$\mu_{R_3}(x)=\begin{cases}\mu_A(x), & x\in D_2\\ \frac{1}{2}, & x\notin D_2\end{cases};$$

$$\mu_{R_4}(x)=\begin{cases}\mu_B(x), & x\in D_2\\ \frac{1}{2}, & x\notin D_2\end{cases};$$

$$\mu_{R_5}(x)=\begin{cases}\mu_A(x), & x\in D_3\\ \frac{1}{2}, & x\notin D_3\end{cases};$$

$$\mu_{R_6}(x)=\begin{cases}\mu_B(x), & x\in D_3\\ \frac{1}{2}, & x\notin D_3\end{cases};$$

* This work is supported by National Natural Science Foundation under Grant 79970025 of China, Foundation of Hubei Provincial Department of Education under Grant 2003X130 and 2004D005, and Scientific Research Team of Wuhan Polytechnic University under Grant 03-T-06.

$$\mu_{R_7}(x)=\begin{cases}\mu_A(x)\vee\mu_B(x), & x\in D_4\\ \frac{1}{2}, & x\notin D_4\end{cases};$$

$$\mu_{R_8}(x)=\begin{cases}\mu_A(x)\wedge\mu_B(x), & x\in D_4\\ \frac{1}{2}, & x\notin D_4\end{cases};\qquad(5)$$

Then define

$$S(A,B)=\frac{1}{2}(e(R_1)-e(R_2)+e(R_3)+e(R_4)$$
$$+e(R_5)+e(R_6)-e(R_7)+e(R_8))-1$$

The discussion of the first and second method is relatively much. We can prove this existing kind of similarity measure, which induced from entropy of fuzzy sets, all that satisfies the proposition, that is, $S(A,B)=S(A^c,B^c)$. Do all these kinds of similarity measure satisfy this proposition? We will get a kind of similarity measure that does not satisfy this proposition.

2. PRELIMINARY

Definition 2.1. [1,2] A real function S: $F^2\to R^+$ is called a similarity measure of fuzzy set on F if S satisfies the following properties:

(D1) $E(A,B)=E(B,A)$, $\forall A,B\in F$;

(D2) If $\mu_A(x)\in\{0,1\}$, then $E(A,A^c)=0$;

(D3) $E(C,C)=1,\quad\forall C\in F(X)$;

(D4) If $A\subseteq B\subseteq C$, then are

$$E(A,C)\le\min\{E(A,B),E(B,C)\}.$$

Definition 2.2. [1,2] A real function $e:F^2\to R^+$ is called an entropy on F if e has the following properties:

(1) $e(D)=0,\forall D\in\wp(x)$;

(2) $e([\frac{1}{2}]_X)=1$, where $\mu_{[\frac{1}{2}]_X}(x)=\frac{1}{2}$;

(3) $\forall A,B\in F$, if $\mu_B(x)\ge\mu_A(x)$ when $\mu_A(x)\ge\frac{1}{2}$ and $\mu_B(x)\le\mu_A(x)$ when $\mu_A(x)\le\frac{1}{2}$, then $e(A)\ge e(B)$;

(4) $\forall A\in F,\quad e(A)=e(A^c)$.

Definition 2.3. [5] Let E be a similarity measure on F. We call a σ –similarity measure on F, if for any $\forall A,B\in F(X)$ and $D\in P(X)$, there holds

$$E(A,B)=E(A\cap D,B\cup D^c)+E(A\cap D^c,B\cup D)$$

3. ANALYSIS OF ARITHMETIC OF SIMILARITY MEASURE INDUCED BY ENTROPY

According to the analysis of the distance measure arithmetic that construct from entropy, many scholar get the following fact:

Proposition2.1. [2,5,6] If define $S(A,B)=e(\frac{A-B}{2}+[\frac{1}{2}]_X)$, so $S(A,B)$ is a similarity measure.

Proposition2.2. If define $S(A,B)=e(\frac{A-B}{2}+[\frac{1}{2}]_X)$, so $S(A,B)=S(A^c,B^c)$;

Proof. $S(A^c,B^c)=e(\frac{1}{2}(A-B)+[\frac{1}{2}]_X)=S(A,B)$

$$\mu_{A^c-B^c}(x)=|\mu_{A^c}(x)-\mu_{B^c}(x)|=|\mu_A(x_i)-\mu_B(x_i)|=\mu_{A-B}(x)$$

Hence $S(A,B)=S(A^c,B^c)$.

Functions $R_i(i=1,2,3,4,5,6,7,8)$ are defined as Introduction, the essay give the follow characters:

Proposition2.3. If define

$$S(A,B)=\frac{1}{2}(e(R_1)-e(R_2)+e(R_3)+e(R_4)$$
$$+e(R_5)+e(R_6)-e(R_7)+e(R_8))-1$$

Then $S(A,B)$ is a similarity measure.

Proposition2.4. The above similarity measure $S(A,B)$ satisfy $S(A,B)=S(A^c,B^c)$.

Proof. For $A,B\in F(X)$ and $i=1,2,3,4$, let $D'_i\in P(X)$ as

$$D'_1=\{x\in X;\mu_{A^c}(x)\ge\frac{1}{2},\mu_{B^c}(x)\ge\frac{1}{2}\},$$
$$D'_2=\{x\in X;\mu_{A^c}(x)\ge\frac{1}{2},\mu_{B^c}(x)<\frac{1}{2}\},$$
$$D'_3=\{x\in X;\mu_{A^c}(x)<\frac{1}{2},\mu_{B^c}(x)\ge\frac{1}{2}\},$$
$$D'_4=\{x\in X;\mu_{A^c}(x)<\frac{1}{2},\mu_{B^c}(x)<\frac{1}{2}\}$$

Then define $R'_i\in F(X)(i=1,2,3,4,5,6,7,8)$ as in the formula (5), so：

$$R_1=R_8^{'},R_2=R_7^{'},R_3=R_6^{'},R_4=R_5^{'},$$
$$R_5=R_4^{'},R_6=R_3^{'},R_7=R_2^{'},R_8=R_1^{'}$$

now give the process of proving $e(R_1^{'})=e(R_8)$, others can also educe analogously。

because

$$\mu_{R_1'}(x)=\begin{cases}\mu_{A^c}(x)\vee\mu_{B'}(x) & x\in D_1^{'}\\ \frac{1}{2} & x\notin D_1^{'}\end{cases}$$

so

$$\mu_{R_1'^c}(x)=\begin{cases}1-(\mu_{A^c}(x)\vee\mu_{B^c}(x)) & x\in D_1^{'}\\ \frac{1}{2} & x\notin D_1^{'}\end{cases}$$
$$=\begin{cases}\mu_A(x)\wedge\mu_B(x) & x\in D_4\\ \frac{1}{2} & x\notin D_4\end{cases}$$
$$=\mu_{R_8}(x)$$

and from $e(R_1^{'})=e(R_1^{'c}),$ we can get $e(R_1^{'})=e(R_8)$. So we have

$$S(A^c,B^c)=\frac{1}{2}(e(R'_1)-e(R'_2)+e(R'_3)+e(R'_4)$$
$$+e(R'_5)+e(R'_6)-e(R'_7)+e(R'_8))-1$$
$$=\frac{1}{2}(e(R_1)-e(R_2)+e(R_3)+e(R_4)$$
$$+e(R_5)+e(R_6)-e(R_7)+e(R_8))-1$$
$$=S(A,B)$$

4. A NEW DEFINITION OF SIMILARITY MEASURE

Let $e(A)$ be the entropy of fuzzy set A, we have the following conclusion.

Theorem4.1. If define

$$S(A,B)=1-\left[e(\frac{A\cup B}{2})-e(\frac{A\cap B}{2})\right]$$

Where the membership function of the fuzzy set $\frac{A}{2}$ is $\mu_{\frac{A}{2}}(x)=\frac{\mu_A(x)}{2}$, and then $S(A,B)$ is a similarity measure.

Proof. From definition2.1，we have

(1) Evidently, we can get S(B,A)=S(A,B)；

(2) If $D\in\wp(X)$，so $D\cup D^c=X, D\cap D^c=\varnothing$ and $S(D,D^c)=1-\left[e(\frac{D\cup D^c}{2})-e(\frac{D\cap D^c}{2})\right]=0$；

(3) $S(A,A)=1-(e(\frac{A}{2})-e(\frac{A}{2}))=1$；

(4) If $A,B,C\in F(X)$ and $A\subseteq B\subseteq C$, then we have

$$S(A,B)=1-(e(\frac{AUB}{2})-e(\frac{AI\ B}{2}))$$
$$=1-(e(\frac{B}{2})-e(\frac{A}{2}))$$
$$S(A,C)=1-(e(\frac{AUC}{2})-e(\frac{AI\ C}{2}))$$
$$=1-(e(\frac{C}{2})-e(\frac{A}{2}))$$

and from $e(\frac{B}{2})\le e(\frac{C}{2})$, we can get $S(A,B)\ge S(A,C)$

Proposition4.1. Deluca and Termini [7] proposed Shannon's probability as following.

$$e(A)=\frac{1}{n\ln 2}\sum S_n(\mu_A(x_i))$$

Where

$S_n(\mu_A(x_i))=-\mu_A(x_i)\ln(\mu_A(x_i))-(1-\mu_A(x_i))\ln(1-\mu_A(x_i))$,

we define

$$S(A,B)=1-[e(\frac{AUB}{2})-e(\frac{AI\ B}{2})]$$

Then $S(A,B)\ne S(A^c,B^c)$.

Proof. Suppose $A=X, \mu_B(x)=\frac{1}{2}$, then

$$A^c=\varnothing, \mu_{B^c}(x)=\frac{1}{2}$$

And

$$S(A,B)=e(\frac{B}{2})$$
$$S(A^c,B^c)=1-e(\frac{B^c}{2})=1-e(\frac{B}{2})$$
$$e(\frac{B}{2})=\frac{1}{nln2}.\sum S_n(\mu_{\frac{B}{2}}(x_i))$$
$$S_n(\mu_{\frac{B}{2}}(x_i))=-\frac{1}{4}\ln\frac{1}{4}-\frac{3}{4}\ln\frac{3}{4}$$
$$=\frac{1}{2}\ln 2+\frac{3}{4}\ln\frac{4}{3}$$

Hence

$$e(\frac{B}{2})=\frac{1}{\ln 2}(\frac{1}{2}\ln 2+\frac{3}{4}\ln\frac{4}{3})>\frac{1}{2}\cdot$$

The hypothesis is not true, so

$$S(A,B)\ \ne S(A^c,B^c).$$

5. REFERENCES

[1]. P.Z.Wang, Theory of fuzzy sets and their Applications(Shanghai Science and Technology Publishing House,1982).

[2]. Zeng Wenyi and Li Hongxing, Research on the Relation between Degree of Fuzziness and Degree of Similarity. XiTong LiLun Yu ShiJian, 1995,14(6) :76-79.

[3]. B. Kosko, Fuzzy entropy and conditioning, Infor. Sci. 1986, 40:165-174.

[4]. CHUA-CHIN WANG and HON-SON DON, A Modified Measure for Fuzzy Subsethood, Infor. Sci. 1994, 79:223-232.

[5]. Liu Xuecheng, Entropy, distance measure and similarity measure of fuzzy sets and their relations, Fuzzy Sets and Systems, 1992, 52:305-318.

[6]. Lin Zhong-zhen, σ -entropy, σ -similarity and σ -distance, journal of Jinan University, 2002, 3:8-14.

Tong Xiaojun was born in Shanxi Province, China, on November 12, 1967, and earned Engineering Doctor Degree in Control Theory and Control Engineering in 2002, at the Department of Control Science and Engineering, Huazhong University of Science and Technology, China; the major fields of study are fuzzy theory and its application. He occupied Lecturer and vice professor at the Department of Mathematics, University of Petroleum between 1993 and 2002. Now, He is a full professor at the Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan, Hubei Province, China. His research interests are in fuzzy theory and its application, and research Biomolecular Compute. He dealt with the research of Partial Differential Equation.

The Grade Difference Format and MGM ^{p} (1,n) Optimization Model *

Zhang Shemin[1], Tong Xiaojun[2]

[1] School of Information, Xi’an University of Finance and Economics
Xi’an, Shannxi, Postcode: 710061, the People's Republic of China
[2]Department of Mathematics and Physics Wuhan Polytechnic University
Wuhan, Hubei, Postcode: 430074, the People's Republic of China
[1,2]Department of Control Science and Engineering, Huazhong University of Science and Technology
Wuhan, Hubei, Postcode: 430074, the People's Republic of China
Email: zhang_she_min@163.net **Tel.:** +86(0)2982348390

ABSTRACT

There are some applications that need using the gray system techniques in distributed computing. In the basis of the grade difference format, MGM ^{p} (1,n) model is proposed and some precision grade difference formats are obtained. Their relations and the mathematics’ mechanism are discussed. The error will be interjecting to them. In the cases of absolute and comparative error, the gray model, which is a new gray GMG ^{p} (1,n) model, is expounded and discussed. The models have nice anti-interference; the example manifest that the model has good effect.

Keywords: the gray model, the grade difference format, poor information, error.

1. INTRODUCTION

In 1982, Grey Systems Theory (GST)[1,2] appeared in the world. In 1981, a seminar of two-persons came into existence and thus organized research began. During the following twenty years, great progress has been made on both theory and application. Nowadays GST has entered Library Classification of China [3] and become one of the important contents of Systems Science (SS)(N94 SS…N941.5 Grey Systems Theory…). So far, the most widely used grey model is GM(1,1)[4,5]. In December of that year, Chinese Society Future Studies held a countrywide conference on theory and technology of forecasting in Xiamen City. At the conference, two papers [5] and [6] were published, which marked the birth of GM (1,1), gray quantity and etc.

J. Zhai constructed the multidimensional model similar to the method of $GM(1,1)$. We have expressed its gray differential equation by the grade difference format (GDF). And we have constructed MGM ^{p} (1,n) optimization model using GDF, but there are some undetermined parameter in this model, which lead the complicated calculate. Now, we try to construct a new model that may be avoiding the undetermined parameter, and the models have nice anti-interference, the example manifest that the model has good effect.

2. THE GRADE DIFFERENCE FORMATS OF MULTIFACTOR GRAY MODEL

Just for simple and convenient, the mark I denotes unit square matrix, and the mark 0 denotes zero matrix, the mark $x^{(0)}(i)$, $x^{(1)}(i)$ are n -dimensional vectors. Considering an investigated object with a multifactor, only a few of discrete data is observed are obtained, denoted as follows:

$$x^{(0)}=[x_1^{(0)},x_2^{(0)},\cdots,x_n^{(0)}]^T$$
$$x_i^{(0)}=\{x_i^{(0)}(k)\,|\,x_i^{(0)}(k)\ge 0,k=1,2,\cdots,l\},\ i=1,2,\cdots,n \qquad (1)$$

Definition 2.1. [10] The time data sequence $x^{(0)}$ in (1) is called an original sequence. Through the agency of the AGO on $x^{(0)}$, so

$$x^{(1)}(k)=AGOx^{(0)}(k)=\sum_{m=1}^{k}x^{(0)}(m),$$

we can obtain

$$x^{(1)}=[x_1^{(1)},x_2^{(1)},\cdots,x_n^{(1)}]^T$$
$$x_i^{(1)}=\{x_i^{(1)}(k)\,|\,x_i^{(1)}(k)\ge 0,k=1,2,\cdots,l\} \qquad (2)$$

The sequence $x^{(1)}$ is called accumulated generating sequence.

Definition 2.2. [10] Through the agency of the Trend Relational Analysis (TRA)[7] on $x^{(1)}$, if we can find the latent law with exponential sequence, and its gray model

$$\frac{dx^{(1)}}{dt}=Ax^{(1)}+U \qquad (3)$$

Then (4) is defined as the GDF of (3).

$$x^{(1)}(k+1)-x^{(1)}(k)\approx\begin{pmatrix} f_1^k(x^{(1)},A,U)\\ \vdots\\ f_n^k(x^{(1)},A,U)\end{pmatrix} \qquad (4)$$

Definition 2.3. [10] If Eq. (4) holds precisely, and then it is called the precise GDF. If f is related to another parameter p and when $p\to\infty$, (4) holds precisely, then (4) is called

* This work is supported by National Natural Science Foundation under Grant 79970025 of China, Foundation of Hubei Provincial Department of Education under Grant 2003X130 and Scientific Research Team of Wuhan Polytechnic University under Grant 03-T-06.

gradual precise GDF.Analysis of the present similarity measure of fuzzy sets

3. THE PRECISE GDF OF MGM (1,N) AND MGM ^{p} (1,N) OPTIMIZATION MODEL

3.1 The precise GDF of MGM (1,n)

Lemma 3.1. For the sequence

$x^{(0)}=\{x^{(0)}(k)\,|\,x^{(0)}(k)=e^{-Ak}B+(-1)^{k}C, k\in N; B,C\neq 0\}$

Where A、 B、 C denote matrix $(a_{ij})_{n\times n}$ 、 $(b_i)_{n\times 1}$ 、 $(c_i)_{n\times 1}$ respectively, the matrix φ is

$$\varphi=\begin{bmatrix} 1 & x_1^{(1)}(p) & x_2^{(1)}(p) & \mathrm{L} & x_n^{(1)}(p) & x_1^{(1)}(p+1) & \mathrm{L} & x_n^{(1)}(p+m) \\ 1 & x_1^{(1)}(p+1) & x_2^{(1)}(p+1) & \mathrm{L} & x_n^{(1)}(p+1) & x_1^{(1)}(p+2) & \mathrm{L} & x_n^{(1)}(p+1+m) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ 1 & x_1^{(1)}(p+l) & x_2^{(1)}(p+l) & \mathrm{L} & x_n^{(1)}(p+l) & x_1^{(1)}(p+l+1) & \mathrm{L} & x_n^{(1)}(p+l+m) \end{bmatrix} \quad (6)$$

$nm\leq l, m\geq 2$. Then

(1) When $n\leq 2$, $\varphi^{T}\varphi$ is invertible.

(2) When $n>2$, $\varphi^{T}\varphi$ is not invertible.

Proof. We can only prove when $n=2$, the rank of matrix φ equals to 2. Play row and column operations for the matrix φ , we can get

$$\begin{bmatrix} 1 & x_1^{(1)}(p) & x_1^{(1)}(p+1) & \mathrm{L} & x_1^{(1)}(p+n) & x_2^{(1)}(p) & \mathrm{L} & x_n^{(1)}(p+2n) \\ 1 & x_1^{(1)}(p+1) & x_1^{(1)}(p+2) & \mathrm{L} & x_1^{(1)}(p+n+1) & x_2^{(1)}(p+1) & \mathrm{L} & x_n^{(1)}(p+2n+1) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ 1 & x_1^{(1)}(p+l) & x_1^{(1)}(p+l+1) & \mathrm{L} & x_1^{(1)}(p+n+l) & x_2^{(1)}(p+l) & \mathrm{L} & x_n^{(1)}(p+l+2n) \end{bmatrix}$$

$$\to\begin{bmatrix} 1 & x_1^{(1)}(p) & x_1^{(1)}(p+1) & \mathrm{L} & x_1^{(1)}(p+n) & x_2^{(1)}(p) & \mathrm{L} & x_n^{(1)}(p+2n) \\ 0 & x_1^{(0)}(p+1) & x_1^{(0)}(p+2) & \mathrm{L} & x_1^{(0)}(p+n+1) & x_2^{(0)}(p+1) & \mathrm{L} & x_n^{(0)}(p+2n+1) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ 0 & x_1^{(0)}(p+l) & x_1^{(0)}(p+l+1) & \mathrm{L} & x_1^{(0)}(p+n+l) & x_2^{(0)}(p+l) & \mathrm{L} & x_n^{(0)}(p+l+2n) \end{bmatrix}$$

$$=\begin{bmatrix} 1 & x_1^{(1)}(p) \quad x_1^{(1)}(p+1) \quad \mathrm{L} \quad x_1^{(1)}(p+n)\, x_2^{(1)}(p) \quad \mathrm{L} \quad x_n^{(1)}(p+2n) \\ 0 & \\ \mathrm{M} & \Xi \\ 0 & \end{bmatrix}$$

Hence we can only prove that the rank of matrix Ξ is $2n$, that is, its $2n$ rows are linear independence. Suppose they aren't linear independence, and then we have

$$\begin{cases} \sum\limits_{i=1}^{n}\lambda_i e^{A(k+i-1)}b+\sum\limits_{i=1}^{n}\lambda_i(-1)^{k}c=0 & (equ1) \\ \sum\limits_{i=1}^{n}\lambda_i e^{A(k+i)}b+\sum\limits_{i=1}^{n}\lambda_i(-1)^{k+1}c=0 & (equ2) \end{cases}$$

Let $p_A=\sum\limits_{i=1}^{n}\lambda_i e^{A(k+i-1)}, q_\lambda=\sum\limits_{i=1}^{n}\lambda_i(-1)^{k}$, rewrite above system of equations as following:

$$\begin{cases} p_A b=q_\lambda & (equ3) \\ e^{A}p_A b=-q_\lambda & (equ4) \end{cases}$$

According to the matrix e^{A} is invertible, then equation equ3 and equation equ4 is conflicted. Hence we have our conclusion.

(2) We omit its proof.
We can similarly get the following lemma as above.

Lemma 3.2. For the sequence

$$x^{(0)}=\{x^{(0)}(k)|x^{(0)}(k)=e^{-Ak}(B+(-1)^{k}C),k\in N;B,C\neq 0\}$$

Where A、 B、 C denote matrix $(a_{ij})_{n\times n}$ 、 $(b_i)_{n\times 1}$ 、 $(c_i)_{n\times 1}$ respectively, the matrix φ is

$$\varphi=\begin{bmatrix} 1 & x_1^{(1)}(p) & x_2^{(1)}(p) & \mathrm{L} & x_n^{(1)}(p) & x_1^{(1)}(p+1) & \mathrm{L} & x_n^{(1)}(p+m) \\ 1 & x_1^{(1)}(p+1) & x_2^{(1)}(p+1) & \mathrm{L} & x_n^{(1)}(p+1) & x_1^{(1)}(p+2) & \mathrm{L} & x_n^{(1)}(p+1+m) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ 1 & x_1^{(1)}(p+l) & x_2^{(1)}(p+l) & \mathrm{L} & x_n^{(1)}(p+l) & x_1^{(1)}(p+l+1) & \mathrm{L} & x_n^{(1)}(p+l+m) \end{bmatrix}$$

$nm\leq l, m\geq 2$. Then

(1) When $n\leq 2$, $\varphi^{T}\varphi$ is invertible.

(2) When $n>2$, $\varphi^{T}\varphi$ is not invertible.

Theorem 3.1. [10] If exponential sequence is hidden in $x^{(1)}$, that is, $x^{(1)}(k)=e^{-A(k-1)}b$, then (7) is a precise GDF of Eq. (7').

$$\frac{dx^{(1)}(t)}{dt}+Ax^{(1)}(t)=u \qquad (7')$$

$$x^{(1)}(k+1)-x^{(1)}(k)=-(I-e^{-A})x^{(1)}(k)+(I-e^{-A})c$$

That is, $x^{(0)}(k+1)=-A_1x^{(1)}(k)+A_2$ (7)

$$x=\begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \mathrm{M} \\ x^{(0)}(l) \end{pmatrix} \quad \phi=\begin{pmatrix} -x^{(1)}(1) & 1 \\ -x^{(1)}(2) & 1 \\ \mathrm{M} & \mathrm{M} \\ -x^{(1)}(l-1) & 1 \end{pmatrix} \quad \Lambda=\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}=\left(\phi^{T}\cdot\phi\right)^{-1}\phi^{T}x$$

Where $A=-\ln(I-A_1)$, $c=A_1^{-1}A_2$, $u=Ac$, and I is the unit matrix.

Proof.

$$x^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)=e^{-A}x^{(1)}(k)-x^{(1)}(k)+(I-e^{-A})c$$
$$=-(I-e^{-A})x^{(1)}(k)+(I-e^{-A})c$$

Theorem 3.2. The condition as Theorem 3.1, then Eq. (3) has the following precise GDF.

$$x^{(1)}(k+1)-x^{(1)}(k)=(I-e^{A})x^{(1)}(k+1)+(I-e^{A})c$$

That is, $x^{(0)}(k+1)=A_1{}'x^{(1)}(k+1)+A_2{}'$ (8)

Where $A=\ln(I-A_1{}')$, $c=A_1{}'^{-1}A_2{}'$, $u=A'c$ and

$$x=\begin{pmatrix}x^{(0)}(2)\\x^{(0)}(3)\\\vdots\\x^{(0)}(l)\end{pmatrix}\quad \phi=\begin{pmatrix}-x^{(1)}(1)&1\\-x^{(1)}(2)&1\\\vdots&\vdots\\-x^{(1)}(l-1)&1\end{pmatrix}\quad \Lambda=\begin{pmatrix}A_1\\A_2\end{pmatrix}=\left(\phi^T\cdot\phi\right)^{-1}\phi^T x$$

Similar to [9], we have

Corrolly 3.1 The condition as Theorem 2.1, for any $n\times n$ order matrix α, a precise GDF of equation (7') is following.

$$x^{(0)}(k+1)$$
$$=-(I-\alpha)A_1x^{(1)}(k)+\alpha A_1{}'x^{(1)}(k+1)+(I-\alpha)A_2+\alpha A_2{}'$$
$$=-(I-\alpha)(I-e^{-A})x^{(1)}(k)+\alpha(I-e^{A})x^{(1)}(k+1)+(\alpha e^{A}-I+\alpha)(I-e^{-A})c \quad (9)$$

Theorem 3.3. The condition as Theorem 2.1, for any $n\times n$ order matrix α, a precise GDF of equation $\dfrac{dx^{(1)}(t)}{dt}+Ax^{(1)}(t)=u$ is following.

$$x^{(0)}(k)=-(I-\alpha)(e^{2A}-e^{A})x^{(1)}(k+1)+\alpha(I-e^{A})x^{(1)}(k)$$
$$+(e^{A}-\alpha e^{A}+\alpha)(e^{A}-I)c$$
$$=p_1x^{(1)}(k)+p_2x^{(1)}(k+1)+p_3 \quad (10)$$

$$\begin{cases}p_1=\alpha(I-e^{A})\\p_2=-(I-\alpha)(e^{2A}-e^{A})\\p_3=(I-e^{-A})(e^{A}-\alpha e^{A}+\alpha)c\\c=A^{-1}\text{u}\end{cases} \quad (11)$$

Or

$$x^{(0)}(k+1)=(I-\alpha)(e^{-2A}-e^{-A})x^{(1)}(k-1)$$
$$+\alpha(e^{-A}-1)x^{(1)}(k)-(e^{A}-\alpha e^{A}+\alpha)(e^{-A}-I)c$$
$$=p_1x^{(1)}(k-1)+p_2x^{(1)}(k)+p_3 \quad (12)$$

$$\begin{cases}p_1=(I-\alpha)(e^{-2A}-e^{-A})\\p_2=\alpha(e^{-A}-1)\\p_3=-(e^{A}-\alpha e^{A}-\alpha)(e^{-A}-1)c\\c=A^{-1}\text{u}\end{cases} \quad (13)$$

Proof. Because this conclusion can be obtained similarly to corrolly 3.1, we omit it.

Theorem 3.4. [10] Consider the following original data $x^{(0)}$, where ε is an arbitrary real and ε is not zero.

$$x^{(0)}=\{x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(l)\}$$
$$=\{b+\varepsilon,e^{-A}b-\varepsilon,\cdots e^{-A(k-1)}b+(-1)^{k-1}\varepsilon,\cdots,e^{-(l-1)A}b+(-1)^{l-1}\varepsilon\}$$

Using the solution of LS of theorem 3.3, we can get the function of $x^{(0)}(k)$: $\hat{x}^{(0)}(k)=e^{-A(k-1)}b'$.

Theorem 3.5. [10] Consider the following original data $x^{(0)}$, where ε is an arbitrary real and ε is not zero.

$$x^{(0)}=\{x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(l)\}$$
$$=\{(I+\varepsilon')b,(I-\varepsilon')e^{-A}b,\cdots,(I+(-1)^{k-1}\varepsilon')e^{-A(k-1)}$$
$$\cdots,(I+(-1)^{l-1}\varepsilon')e^{-(l-1)A}\}$$

Using the solution of LS of theorem 3.3, we can get the function of $x^{(0)}(k):\hat{x}^{(0)}(k)=e^{-A(k-1)}b''$.

Note. The method of literature [10], we need seek some undermined parameter through search method, which lead the complicated calculate. Now, we try to construct a new model that may be avoiding the undetermined parameter.

3.2 MGM ^{p} (1,n) Optimization Model

According to theorem 3.3-theorem 3.5, we can get the following MGM ^{p} (1,n) model.

$$x^{(0)}(k)-2x^{(1)}(k)=p_2\left(x^{(1)}(k)+x^{(1)}(k+1)\right)+p_3 \quad (14)$$

Where $p_2=-e^{A}$ or

$$x^{(0)}(k)-x^{(1)}(k)=p_2x^{(1)}(k+1)+p_3 \quad (15)$$

Where $p_2=-e^{2A}$。

The following is computation method of model MGM ^{p} (1,n)

1. (1) For original data sequence $x^{(0)}$, with absolute errors, using the solution of LS p_2,p_3 of equations group (14') and $p_2=-e^{A}$, we can get parameter matrix A.

$$\begin{pmatrix} x^{(1)}(1)+x^{(1)}(2) & 1 \\ x^{(1)}(2)+x^{(1)}(3) & 1 \\ \vdots & \vdots \\ x^{(1)}(l-1)+x^{(1)}(l) & 1 \end{pmatrix}\begin{pmatrix} p_2 \\ p_3 \end{pmatrix}=\begin{pmatrix} x^{(0)}(1) \\ x^{(0)}(2) \\ \vdots \\ x^{(0)}(l-1) \end{pmatrix}-2\begin{pmatrix} x^{(1)}(1) \\ x^{(1)}(2) \\ \vdots \\ x^{(1)}(l-1) \end{pmatrix} \qquad (14')$$

(2) For original data sequence $x^{(0)}$, with relative errors, using the solution of LS p_2, p_3 of equations group (15') and $p_2=-e^{2A}$, we can get parameter matrix A.

$$\begin{pmatrix} x^{(1)}(2) & 1 \\ x^{(1)}(3) & 1 \\ \vdots & \vdots \\ x^{(1)}(l) & 1 \end{pmatrix}\begin{pmatrix} p_2 \\ p_3 \end{pmatrix}=\begin{pmatrix} x^{(0)}(1) \\ x^{(0)}(2) \\ \vdots \\ x^{(0)}(l-1) \end{pmatrix}-\begin{pmatrix} x^{(1)}(1) \\ x^{(1)}(2) \\ \vdots \\ x^{(1)}(l-1) \end{pmatrix} \qquad (15')$$

2. For $x^{(0)}(k)=e^{-A(k-1)}b_1$, use data sequence $x^{(1)}$ to perform linear fit to get b_1, and furthermore to get the model we want: $\hat{x}^{(0)}(k)=e^{-A(k-1)}b_1$.

Note. For above second step, let $\hat{x}^{(0)}(k)=\{p_1(k), p_2(k), \cdots, p_n(k)\}$, we can define $\angle x^{(0)}(k)$ as following.

$$\angle x^{(0)}(k)=\{p_1(k)/x_1^{(0)}(k), p_2(k)/x_2^{(0)}(k), \cdots, p_n(k)/x_n^{(0)}(k)\}$$

Thus we can get b_1 through linear fit.

4. DEMONSTRATING EXAMPLE FUZZY SET

Example. [10] We can regard the number of national corporation employment from 1980 to 1990 as sequence $x_1^{(0)}$, and regard townish collectivity employment as sequence $x_2^{(0)}$. We will use the data of 1980 to 1990 as historic data (size of the sample is 11), and the data of 1991 to1992 will be used to test prediction effect. The results of l GM(1,1) model, MGM(1,2) model, and please see paper [6]. Compute fitting error and prediction error for every model to compare (using average relative error). The results are listed in table 1 and table 2.

$$x_1^{(0)}=\{481.8, 509.3, 532.9, 544.5, 554.2, 576.7, 617.3, 618.2, 623.5, 614.7, 621.0\}$$

$$x_2^{(0)}=\{166.2, 175.9, 200.1, 258.1, 293.5, 334.8, 376.4, 405.9, 421.3, 390.1, 389.7\}$$

The generating sequences $\hat{x}^{(0)}$ of three models are as following.

GM(1,1) model

$$\begin{cases} \hat{x}_1^{(0)}(1)=x_1^{(0)}(1), \\ \hat{x}_1^{(0)}(k)=512.212e^{0.0226(k-1)}, k=2, \cdots, \end{cases}$$

$$\begin{cases} \hat{x}_2^{(0)}(1)=x_2^{(0)}(1), \\ \hat{x}_2^{(0)}(k)=210.158e^{0.0759(k-1)}, k=2, \cdots, \end{cases}$$

MGM (1,2) model

$$\begin{cases} x^{(0)}(1)=\{x_1^{(0)}(1), x_2^{(0)}(1)\}^T \\ x^{(0)}(k)=e^{A_1(k-2)}b_1, \text{k=2,3,}\cdots \end{cases}$$

Where

$$A_1=\begin{pmatrix} 0.097 & -0.1289 \\ 0.2477 & -0.3503 \end{pmatrix}, b_1=\begin{pmatrix} 505.294 \\ 158.926 \end{pmatrix}$$

Table 1

k	$x_1^{(0)}(k)$	Fitting or Prediction			With weight added (%)		
		GM (1,1) Model	MGM (1,2) Model	MGM^p (1,2) Model	GM (1,1) Model	MGM (1,2) Model	MGM^p (1,2) Model
1	481.8	481.0	481.8	465.90	0	0	0
2	509.3	523.92	505.30	498.99	-2.87	0.78	2.02
3	532.9	535.90	531.00	525.63	-0.56	0.36	1.36
4	544.5	548.15	551.91	547.38	-0.67	-1.36	-0.53
5	554.2	560.69	569.15	565.44	-1.17	-2.70	-2.03
6	576.7	573.51	583.57	580.70	0.55	-1.19	-0.69
7	617.3	586.62	595.82	593.86	4.97	3.48	3.80
8	618.2	600.03	606.41	605.43	2.94	1.91	2.07
9	623.5	613.76	615.74	615.82	1.56	1.24	1.23
10	614.7	627.79	624.10	625.33	-2.13	-1.53	-1.73
11	623.0	642.14	631.73	634.18	-3.41	-1.73	-2.12
12	638.9	656.83	639.00	642.55	-2.81	0.02	0.57
13	681.2	671.85	645.45	650.55	1.37	5.25	-4.50

Table 2

k	$x_2^{(0)}(k)$	Fitting or Prediction			With weight added (%)		
		GM (1,1) Model	MGM (1,2) Model	MGM^p (1,2) Model	GM (1,1) Model	MGM (1,2) Model	MGM^p (1,2) Model
1	166.2	166.2	166.2	166.2	0	0	0
2	175.9	226.73	158.90	132.07	-28.89	9.66	24.92
3	200.1	244.61	220.15	202.31	-22.25	-10.17	-1.10
4	258.1	263.91	268.63	256.73	-2.25	-4.08	0.53
5	293.5	284.74	306.52	299.13	2.98	-4.44	-1.92
6	334.8	307.20	336.49	332.42	8.24	-0.50	0.71
7	376.4	331.44	360.36	358.80	11.94	4.26	4.68
8	405.9	357.59	379.55	379.93	11.90	6.49	6.40
9	421.3	385.81	395.13	397.10	8.42	6.21	5.74
10	390.1	416.25	407.95	411.26	-6.70	-4.58	-5.42
11	389.7	449.09	418.65	423.15	-15.24	-7.43	-8.58
12	419.4	484.53	427.71	433.33	-15.5	-1.98	-3.32
13	476.3	522.76	435.52	442.22	-9.75	8.56	7.15

MGM ^{p} (1,2) Optimization Model

$$\begin{cases} x^{(0)}(1)=\{x_1^{(0)}(1), x_2^{(0)}(1)\}^T \\ x^{(0)}(k)=e^{A_2(k-2)}b_2, \text{k=2,3,}\cdots \end{cases}$$

Where

$$A_2 = \begin{pmatrix} 1.2455 & -0.2761 \\ 0.6163 & 0.1677 \end{pmatrix}, b_2 = \begin{pmatrix} 465.902 \\ 40.972 \end{pmatrix}$$

5. REFERENCES

[1]. Julong Deng. Control problems of grey systems. Systems & Control Letters, 1982, 1 (5) 288-294.

[2]. Mianyun Chen. Grey dynamics of the system of a boring machine. Journal of Huazhong University of Science and Technology(in Chinese), 1982,10(6):7-11.

[3]. Julong Deng. The GM Model of Grey Systems. Fuzzy Mathematics(in Chinese), 1986,5(2):23-31

[4]. Julong Deng, Mianyun Chen, Guozhong Peng et al. Grey Block Theory and Long-term Forecasting Model. Preprints of the CSFS Symposium on Forecast Theory and Technical Method (in Chinese), Dec. 1-5,1983, Xiamen, China. Proceedings of the Futurenology (in Chinese), Hubei Wuhan Society of Futurenology studies, 1984, 1: 41-46.

[5]. Mianyun Chen. Grey Systems Theory is a New Direction in Researches. Preprints of the CSFS Symposium on Forecast Theory and Technical Method (in Chinese). Dec. 1-5,1983, xiamen, China. Proceedings of the Futurenology (in Chinese), Hubei Wuhan Society of Futurenology studies, 1984, 1: 26-32.

[6]. Jun Zhai, Jianming Sheng. Yingjun Feng. The Grey Modelling MGM(1,n) and Its Application. Systems Engineering—Theory & Practice (in Chinese). 1997, 5:109-113

[7]. Qingyan Xiao. Practical Forecasting Technique and Its Applications. Wuhan: Huazhong University of Science and Technology Press (in Chinese), 1993. 283-299.

[8]. Jiazu Yuan. Grey Systems Theory and Its Applications(in Chinese). Beijing: Science Press. 39-189.

[9]. Xiaojun. Tong and Mianyun. Chen. Gray Logistic model based on grade difference format. Control and Decision (in chinese), 2002, Vol.17, No 5. 554~558.

[10]. Xiaoxia. Li, Xiaojun. Tong and Mianyun. Chen. MGM ^{p} (1,n) Optimization Model. XITONG GONGCHENG LILUN YU SHIJIAN 2003, Vol 23, No4, 47~51.

[11]. Editorial Board of the Library Classification of China. Library Classification of China, Fourth Edition, Beijing: Beijing Library Publishing House. 1999. 247-249..

Zhang shemin, was born in April,1962. He is now an associate professor in the school of information at Xi’an University of Finance and Economics, and is reading his doctor’s degree in the Department of Control Science and Engineering at Huazhong University of Science and Technology. He graduated in 1987 with the Degree of Master of Science in Computer Department at Shandong Normal University. In the past few years he has published over 10 articles in the field. His research interests are in distributed DNA computing，graph theory and combinatorial optimization.

Tong Xiaojun was born in Shanxi Province, China, on November 12, 1967, and earned Engineering Doctor Degree in Control Theory and Control Engineering in 2002, at the Department of Control Science and Engineering, Huazhong University of Science and Technology, China; the major fields of study are fuzzy theory and its application. He occupied Lecturer and vice professor at the Department of Mathematics, University of Petroleum between 1993 and 2002. Now, He is a full professor at the Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan, Hubei Province, China. His research interests are in fuzzy theory and its application, and research Biomolecular Compute. He dealt with the research of Partial Differential Equation

An Object-oriented Knowledge Representation Method

Xu Yong[a] **Zhong Luo**[a] **Yang Ke**[b]
[a]**School of Computer Science and Technology, Wuhan University of Technology, Wuhan, 430070, P.R.China**
[b]**College of Architecture & Civil Engineering, Wenzhou University, Wenzhou, 325027, P.R.China**
Email: hfxing@sohu.com Tel.: +86 (0)27-85758496

ABSTRACT

In this paper, a kind of OO knowledge representation method has been presented, which is named OORL. In the process of designing and developing the geotechnical engineering security inspection expert system development tool, in order to represent the domain knowledge still better, the OO technology has been used to rebuild the production rule and construct the OORL. In OORL, logic tree, production rules, methods, engineering examples, theoretical criterions and explaining engine are all encapsulated. The form of OORL is a binary. A single OORL can solely finish a relatively absolute reasoning procedure. It is convenient for OORL to be used to represent the procedural knowledge and the judgmental knowledge. OORL is designed aiming at the actual character of using the inspection reference to analyze and evaluate the safety condition of the geotechnical engineering. Compared with the traditional knowledge representation methods, OORL has more pertinence in this domain.

Keywords: Geotechnical engineering, Object-oriented, Knowledge representation.

1. INTRODUCTION

The choice and design of the knowledge representation method is very important for building an expert system. The knowledge representation methods used in the expert system should be brief and definite. And the knowledge represented using the methods should be easy to be modified and expanded. There is no perfect theory of evaluating the knowledge representation methods. So it is difficult to determine which domain a kind of knowledge representation method applies to. The most important facts used for evaluating a knowledge representation method are what it expresses in practice, including the ability of representing knowledge, the ability of deducing the new information from the existent information, etc. When we use the traditional knowledge representation methods, such as semantic network, OAV, frame and production rule [1, 2], to represent the knowledge, along with the augment of the objects and rules, the alternation between the object and the rule will become very complex because of the lack of the modularity of the knowledge representation. The difficulty of developing and maintaining the expert system is increased greatly [3]. So it is necessary to improve the traditional knowledge methods and make them more fit the domain which we research on.

The OO technology [4] is a new method for knowledge representation. In the OO method, object and message are used for expressing the entity and the entity relation, class and inheritance are used for simulating human thought mode. The object is a kind of profound abstract of the things. Using the data type and a set of procedures encapsulated in each object, we can give a uniform presentation of the various knowledge in the correlative domains. So it is easy for the users to understand and accept the domain knowledge.

In the domain of geotechnical engineering security inspection, usually the same instruments have been placed in many different measuring points. The approaches or steps of processing the inspection data are same or similar on the whole. The difference of data processing mostly consists in using the different data and criterions for the different measuring points. So in order to describe the problem solving model of this domain still better, we design a new kind of knowledge representation method using the OO technology. This new representation method is based on the production rule which is widely used in many expert systems. Because of the OO idea used in this method, we name it OORL, object-oriented rule, in which logic tree, methods, instances, criterions and explaining engine are all encapsulated. Although there are many kinds of OO knowledge representation methods [5, 6], OORL is designed aiming at the concrete character of analyzing and evaluating safety condition of the geotechnical engineering using the inspection reference. It has more pertinence in this domain.

2. THE CONSTRUCTION OF OORL

OORL consists of production rules, engineering examples, theoretical criterions and formulas, as in Fig.1. The form of OORL is a binary tree. The rule nodes are divided into end nodes and procedure nodes. The end node describes the inference result of the rule. The procedure node is the method node, in which the inspection data and the methods used for analyzing these data are encapsulated, and can be called in the inference procedure. An OORL combines some associated production rules and can solely finish a relatively absolute reasoning procedure. In an OORL, from the entrance to each end node, it is a production rule. Because of the criterions, formulas and logic tree, which are all encapsulated in OORL, OORL is not a simple combination of production rules, and can express ‘logic AND’, ‘logic OR’ and ‘logic NOT’. OORL can be predigested into a production rule, so it is natural for OORL to represent the judgmental knowledge. Each OORL is also a relative integrate inference chain, therefore OORL is fit for representing the procedural knowledge. And so the problem that the production rule can not naturally represent the procedural knowledge is solved.

OORL is divided into rule class and rule object. Rule class is the abstract inference methods or inference chains, describes the commonness of measuring points or instruments of a kind, and can not be directly used for reasoning. While the rule object, which is the instantiation object of rule class, is built according to the practical problem, describes one or more concrete instruments or measuring points, and can be directly used in the inference procedure.

Formula, the method of OORL, finishes calculation and database operation in the inference procedure. The data used

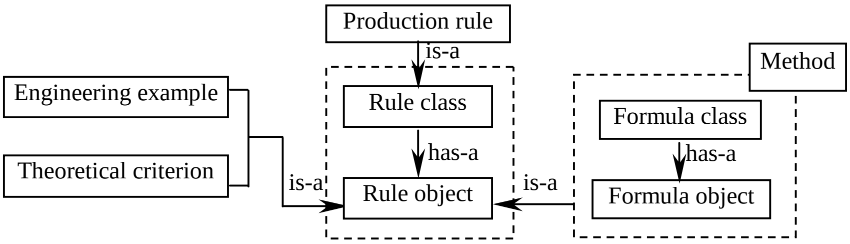


Fig.1. the structure of OORL

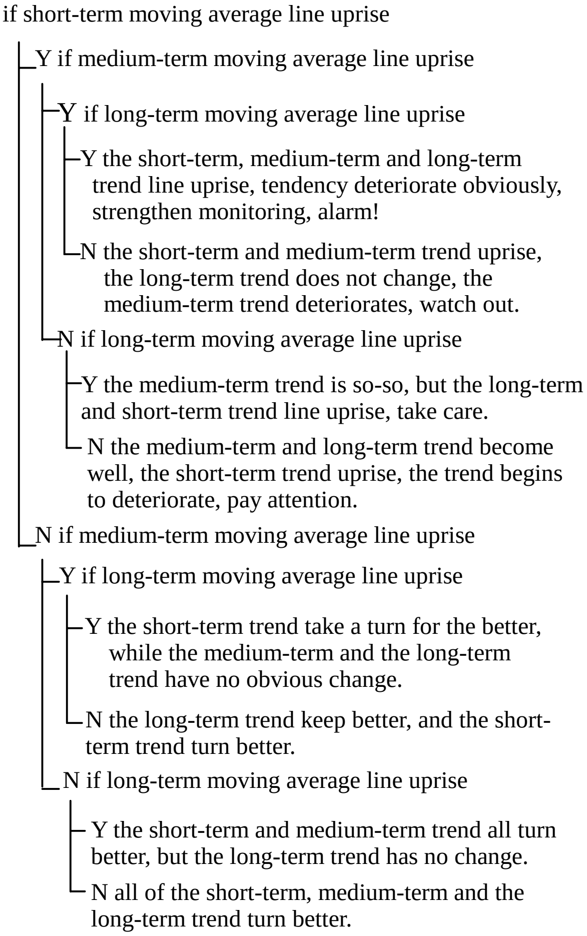
in reasoning is encapsulated in OORL when the source of the data is appointed in formula. Formula is also constructed using the OO method, and divided into formula class and formula object. The commonness of the formulas of a kind is described in formula class. The formula class can not be used in the rule. The formula object is the instantiation object of the formula class and can be directly called by the rule.

An OORL can be built in accordance with the next steps. Firstly, an inference template which includes a strict logic inference chain is constructed in the system. This template is just the tree structure of OORL. And then, we use the template to build some rule classes about the different kinds of instruments. The rule class will inherit the inference chain of the template. Furthermore, the judgmental words, the methods and the relative interpretation, which can be generally used in the rule class, will be also added to the rule class. Finally, the rule objects aimed at the concrete measuring point are built. When building the rule object, we must reify what can not be confirmed in the rule class, such as the name of measuring points related to the rule object, the source of the data, the type of the criterion, and so on. An inference template can be used for building lots of rule classes. And a rule class can be easily used for building lots of rule objects. So a great deal of repetitive work that the same kinds of rules are written can be avoided.

Now we use an example presented in Fig.2 to give more explanation of OORL. This example is a trend judge rule, in which we use the relative relation between the short-term moving average line, the middle-term moving average line and the long-term moving average line to judge the trend of current status. Here we don’t explain the theory base of the rule. In order to conveniently illuminate the problem, we assume that data descending means trend become better. The meaning of trend uprising is that the latter moving average value is bigger than the prior value. It does not mean the trend uprising will change to trend descending that the current inspection data is bigger than the prior data, so we use the relative relation of moving average line to judge the trend instead of using the relative proportions of the two successive data. In the rule, each Y(Yes) or N(No) is a node. Each node may have the son nodes. The nodes which have no son nodes are the end nodes. The others are the procedure nodes. The letter of the procedure node is the judgment needed to be processed. And the letter of the end node is the conclusion and the management advice. The conclusion includes the direct result gotten from the judgment and the extended meaning of the project.

When the rule class is built using the inference template of this example, short-term, middle-term and long-term can be altered according to the type of the instruments or the inspection frequency. For example, they can be altered using quarter, month and year. At the same time, the formulas, or the methods are added to the nodes. And the setting and letter about the judgment and conclusion of the nodes are adjusted relatively. And then, we can rapidly build a series of rule classes, which have the similar inference procedure and will be used on the other conditions.

In this example, eight production rules have been combined together. From the Fig.2, we can clearly know the relation between these eight rules and the logic tree of the whole OORL.


if short-term moving average line uprise

- Y if medium-term moving average line uprise
    - Y if long-term moving average line uprise
        - Y the short-term, medium-term and long-term trend line uprise, tendency deteriorate obviously, strengthen monitoring, alarm!
        - N the short-term and medium-term trend uprise, the long-term trend does not change, the medium-term trend deteriorates, watch out.
    - N if long-term moving average line uprise
        - Y the medium-term trend is so-so, but the long-term and short-term trend line uprise, take care.
        - N the medium-term and long-term trend become well, the short-term trend uprise, the trend begins to deteriorate, pay attention.
- N if medium-term moving average line uprise
    - Y if long-term moving average line uprise
        - Y the short-term trend take a turn for the better, while the medium-term and the long-term trend have no obvious change.
        - N the long-term trend keep better, and the short-term trend turn better.
    - N if long-term moving average line uprise
        - Y the short-term and medium-term trend all turn better, but the long-term trend has no change.
        - N all of the short-term, medium-term and the long-term trend turn better.

Fig.2. the sketch map of an OORL

3. THE INFERENCE MACHINE AND EXPLAINING ENGINE OF OORL

The inference machine encapsulated in OORL is the bottommost machine of our system, which can be called by other inference machine on the upper level. The reasoning algorithm of OORL is as follows:

```
AppRuleInference (int NodeNum) {
    if (Current node is end node) Output the result;
    else {
        Get the number of application formula from the
        structure of rule node;
        Run the formula according to the number which
        have been found;
        Search the next rule node and store its number to
        the variate, NextNodeNum;
        AppRuleInference(NextNodeNum); }
}
```

In the inference procedure of OORL, each condition and the relative judgment are processed only once. It just like to rightly select a production rule to reason for the first time and give up the other relative production rules in terms of the reasoning result. So the calls of the unrelated rules are decreased tremendously and the inference efficiency is consumedly increased. On the above example, using OORL, it is only needed to call a traditional rule and process three judgments and relative compute to get the result. While using traditional rule, even if we merely call the eight rules encapsulated in the OORL under some control strategy, to get the result we have to call these eight rules one by one and process twelve judgment at the worst. The reasoning result of one OORL will be corporate with the result of the other rules and called by the inference machine on the upper level to determine which OORL will be called next and which operation will be processed.

Using any exit of OORL as the start, we can easily reverse the inference procedure and confirm the inference chain of the result, and then find the cause of the result. If there are not the instances conflicting with the known reality, the entrance of the rule is just the result of the backward reasoning. And so the interpretation of the inference procedure is convenient to be given. The drawbacking that inference procedure of the production rule is difficult to be understood is conquered.

4. THE MAINTENANCE OF OORL

The relationship between the traditional rules is relative independency, which can only be indirectly expressed through the context. Usually, the traditional rules are considered that they are free to be appended, deleted and modified. But we think, the independence of the production rule is more shown in form, instead of being shown in logic, and the free of append, deletion and modification should be aimed at not only the rule itself but also the collectivity of the problems which we research on. In the development process of the expert system based on the traditional production rule, the conflicts resulted from appending or deleting the rule often disturb the developer much. And the inference procedure of production rule is also generally considered fuzzy. It is difficult to find the efficient methods to solve the conflicts. Somewhere conflict can be barely solved, but the other conflicts may be brought because of the solution.

In OORL, a number of production rules and relative formulas are encapsulated together according to the compact logical relationship between the production rules. The structure of OORL is not absolutely but relatively unified. The OORL is saved using the rule object as the store unit. It is different from the parallel storage pattern of the production rule. When the inference chain of OORL is modified, the relative chain will be found and asked to be modified accordingly, and then the conflicting conclusion will not be brought on this level. So the conflicts created when the traditional rules are maintained are immensely avoided. When the OORL is being modified, all the content encapsulated in it, such as the formulas, criterions, will be also asked to be modified using the compute formula edit module and the database operation edit module provided in our system. Furthermore, it is very clear that when and where an OORL is called. So before modifying the OORL, we can know the possible effect caused because of the modification through querying the call note list.

The user can complete building and maintaining the rules through the visual operations. The rules can be written with pseudo-code, for example, some compute formulas and the database operations can be written with the grammar of the similar C language. But usually, pseudo-code is automatic generated by the system. So the construction and maintenance work of OORL are very easy.

5. CONCLUSION

OORL is an object-oriented knowledge representation method, which is designed to represent the knowledge in the domain of geotechnical engineering security inspection. In OORL, logic tree, methods, engineering examples, theoretical criterions and explaining engine are encapsulated together. Compared with the structure of the traditional production rule, the structure of OORL is not absolute unification but relative unification. So OORL can be conveniently used to represent the judgmental knowledge and the procedural knowledge. And it is easy to be built, managed and maintained. The quantity of the rules and the times of repetitive judgment are decreased consumedly, the inference efficiency is increased immensely. Therefore, we think that OORL adapts to the domain of geotechnical engineering security inspection and its design method is value for the relative domain.

6. REFERENCES

[1] ZHONG Luo, LI Gui-qing, LIU Gang, “Expert System Shell Development Environment”, Mini-Micro Systems, Vol.19, No.3, March 1998, pp.212-215.

[2] M. M. O. Owrang, F. H. Grupe, “Database tools to acquire knowledge for rule-based expert systems”, Information and Software Technology, Vol.39, Issue.9, September 1997, pp. 607-616.

[3] ZHONG Luo, CHEN Cai-jun, “The Design and Implementation of the Pumpcrete Technique Expert System”, Journal of Wuhan University of Technology, Vol.23, No.10, October 2001, pp.35-40.

[4] Steven Walczak, “Knowledge acquisition and knowledge representation with class: the object-oriented paradigm”, Expert Systems with applications, Vol.15, Issues.3-4, October 1998, pp.235-244.

[5] V. Devedzic, “A survey of modern knowledge modeling techniques”, Expert Systems with Applications, Vol.17, Issue.4, November 1999, pp. 275-294.

[6] Zuo Bing Chen, Li D Xu, “An object-oriented

intelligent CAD system for ceramic kiln", Knowledge-Based Systems, Vol.14, Issues. 5-6, August 2001, pp. 263-270.

Xu Yong is the lecturer of Centre of Information and Network in Hubei Administration Institute. He graduated from Wuhan University of Technology in 1996; obtained his master's degree in computer application technology from Kunming University of Technology in 2000. Presently he is specializing in doctor's degree in Wuhan University of Technology. His research interests are in artificial intelligence, network security and e-government.

Zhong Luo is a Full Professor .He graduated from Wuhan University in 1982. His research interests are in intelligent technology, software engineering, and image graphic.

Mutual Subset Hood of the Fuzzy Set

Wei Yiliang
Wuhan Polytechnic University
Wuhan, Hubei, Postcode 430023, The People's Republic of China
Email: weiyiliang@yahoo.com.cn Tel.: +86(0)2783955676

ABSTRACT

A generic conclusion in l^p meaning is presented. This conclusion shows that the improved reason of mutual subset hood theorem is not sufficient by Wang and Don, and it is impossible to exist a mutual subsethood arithmetics which makes the monotonous decrease with l^p distance at arbitrary $p \geq 1$ as well. It also presents the calculating formula of mutual subsethood at general l^p distance.

Keywords: fuzzy set, mutual subset hood, l^p -distance.

1. INTRODUCTION

Suppose that X is an n member set, $X = \{x_1, x_1, \cdots, x_n\}$. Kosko marked the sethood A on X as the vector $\{m_A(x_1), m_A(x_2), \cdots, m_A(x_n)\}$, where $m_A(x_i)$ is subjection degree of sethood A for the element x_i in X. Wang P. Z. firstly presented the notion scaling the equality degree between two fuzzy sets – axiom definition of similarity measure. It is very spontaneous to define the similarity measure by L^p -distance. Later some scholars improved the definition of the similarity measure. Liu [4] defined σ –similarity measure, which has a good σ -addivity, and then he brought l^1 -distance into the definition of similarity measure and presented a series of properties of the similarity measure. Based on the relationship of set combination, Kosko et al. [6] firstly defined the combination measure of fuzzy set. The relationship of the classic set equation is equivalent the mutual inclusion of two sets, and then the mutual subset hood was derived from the equation degree of fuzzy set by the equivalent relation and the calculating formula of the mutual subset hood was derived from the arithmetic of inclusion degree. It is not difficult to validate that mutual subset hood is a similarity measure but it is not dependent completely on the distance. Wang [7] showed the limitation of the calculating formula suggested by Kosko by the reverse example according to l^p -distance, and then presented the modified formula (See the second part of this paper). Based on the idea that the fuzzy sets regarded as fuzzy information by Kosko et al, Bhandari and Pal [8] constructed fuzzy entropy, which is similar with Shannon's probability entropy, which have some excellent properties. Using the relationship between fuzzy measure and similarity measure, another kind of similarity measure of fuzzy set can be obtained from this fuzzy measure. Similarly, Yan [9] showed the limitation of expressing method by Wang through the reverse example according to l^p -distance. However, whether did the expressing method by Yan [9] exist the limitation and whether can it obtain a similarity measure which satisfied all l^p -distance？Hereby, firstly we obtained a conclusion at l^p, and then explained that it is impossible to exist a mutual subsethood arithmetic, which makes descend monotonously with l^p -distance at arbitrary $p \geq 1$. All these suggest that the improved reason by Wang and Yan is not sufficient.

2. SIMILARITY MEASURE OF FUZZY SET (MUTUAL SUBSETHOOD)

For the two fuzzy sets A, B on $X = \{x_1, x_1, \cdots, x_n\}$, their l^p -distance is：

$$l^p(A,B) = \left(\sum_{i=1}^{n} | m_B(x_i) - m_A(x_i) |^p\right)^{1/p}$$

The set $F(X)$ is the class of all fuzzy sets of X ; $P(X)$ is the class of all crisp sets of X ; $\left[\frac{1}{2}\right]_X$ is the fuzzy set of X for which $m_{[\frac{1}{2}]_X}(x) = \frac{1}{2}$, $\forall x \in X$. The set F is a sub-class of $F(X)$ with (1) $P(X) \subset F$, (2) $\left[\frac{1}{2}\right]_X \in F$, (3) $A, B \in F \Rightarrow A \cup B \in F, A^c \in F$, where A^c is the complement of $A \in F$, i.e. $m_{A^c}(x) = 1 - m_A(x)$, $\forall x \in X$. The number n is the member of the element of set X.

Definition 2.1. A real function S : $F^2 \to R^+$ is called a distance measure on F if S satisfies the following properties:

(D1) $S(A,B) = S(B,A)$, $\forall A, B \in F$;

(D2) If $A(x) \in \{0,1\}$, then $S(A, A^c) = 0$;

(D3) $S(A,B) = 1 \Leftrightarrow A = B$;

(D4) If $A \subseteq B \subseteq C$, then
$E(A,C) \leq \min\{E(A,B), E(B,C)\}$.

Theorem 2.1 For $\forall A, B \in F$, there exists none set function $S(A,B)$, $n \geq 2$, making for $\forall p \geq 1$, if $l^p(A,B) < l^p(A,C)$, then $S(A,B) \geq S(A,C)$.

Proof. Supposing that there exist a fuzzy set function $S(A,B)$, which makes for $\forall p \geq 1$, if

$l^p(A,B) < l^p(A,C)$, then $S(A,B) \ge S(A,C)$. Distinguishingly, if

$$A=\{0,0,\cdots,0\},\quad B=\{\frac{1.1}{n},\frac{1.1}{n},\frac{1.1}{n},\cdots,\frac{1.1}{n}\},$$
$$C=\{1,0,0,\cdots,0\}$$

Then

$$l^1(A,B)=1.1>l^1(A,C)=1$$

From the properties of $S(A,B)$, it can be obtained that $S(A,B)\le S(A,C)$. However, also because

$$l^{\infty}(A,B)=\max\{|m_A(x)-m_B(x)|\}$$

When p is large enough , $l^p(A,C)>l^p(A,B)$ so $S(A,B)\ge S(A,C)$ .As a result, $S(A,B)=S(A,C)$. Similarly, it can be obtained that $S(A,C)=S(A,D_k)=S(A,B)$ when $k\ge 2$, $D_k=\{\frac{k-1}{k},\frac{k-1}{k},\frac{k-1}{k},\cdots,\frac{k-1}{k}\}$, because $\lim\limits_{k\to\infty}D_k=\{1,1,\cdots,1\}=X$,

$$S(A,C)=S(A,X)=S(A,B).$$

Similarly, when $E_k=\{\frac{1.2}{n},\frac{1}{k},\cdots,\frac{1}{k}\}$,

$$S(A,C)=S(A,E_k)=S(A,B),$$

and also

$$E=\lim_{k\to\infty}E_k=\{\frac{1.2}{n},0,\cdots,0\}$$
$$S(A,C)=S(A,E)=S(A,B).$$

Subsequently,

$$S(A,C)=S(A,X)=S(A,A)=S(X,X).$$

Therefore, for $\forall A,B\in F$, only if $S(A,B)$ is constant，for $\forall p\ge 1$, when $l^p(A,B)<l^p(A,C)$, then $S(A,B)\ge S(A,C)$ .This is contradictory with the proposition qualification.

Kosko presented the following "mutual subsethood theorem" according to the subsethood degree of fuzzy set：

Theorem 2.2. [6] Two fuzzy sets $A,B\in F$, their mutual subsethood degree is:

$$E(A,B)=\frac{c(A\cap B)}{c(A\cup B)}$$

Where $c(A)=\sum_{x_i\in X}m_A(x_i)$, $A\cup B$ is the union of fuzzy sets A and B and $A\cap B$ is their intersection. $m_{A\cup B}(x_i)=\max\{m_A(x_i),m_B(x_i)\}$, $m_{A\cap B}(x_i)=\min\{m_A(x_i),m_B(x_i)\}$.

Wang [7] proposed three fuzzy messages A,B,C and $l^1(A,B)=l^1(A,C)$, but when $p>1$, $l^p(A,B)<l^p(A,C)$, and $E(A,B)=E(A,C)$. So, he proposed another mutual subsethood measure: $S(A,B)=\frac{1}{n}\sum c(\frac{A\cap B}{A\cup B})$. As to the improvement of Wang, Yan [9] also showed that the improvement of Wang don't inosculate with by reverse example. By Theorem 2.1, we think that there exist no a arithmetics of similarity measure, which inosculate with all l^p -distance. So this shows that the improved reasons of both Wang and Yan are not sufficient.

Noticing that in mutual subsethood theorem, $A\cap B$ indicates the equivalent component of fuzzy sets A and B , but $A\cup B$ plays a role on unionization and scale. So, we may present the following arithmetic of similarity measure according to l^p -distance:

Theorem 2.3. Two fuzzy sets A,B and $A,B\in F$, then their mutual subsethood is:

$$E(A,B)=\deg ree(A=B)=\frac{l^p(A\cap B,\varnothing)}{l^p(A\cup B,\varnothing)}$$

Or

$$E(A,B)=\deg ree(A=B)=1-\frac{l^p(A,B)}{l^p(A\cup B,\varnothing)}.$$

Proof. We only validate that $E(A,B)=\deg ree(A=B)=\frac{l^p(A\cap B,\varnothing)}{l^p(A\cup B,\varnothing)}$ is a similarity measure.

(1) Obviously, $E(A,B)=E(B,A)$;

(2) When $A\in P(X)$, and $A^c=A$, because $l^p(A\cap B,\varnothing)=0$, we have $E(A,B)=0$;

(3) When $A=B$, one can in evidence get $E(A,B)=1$; and when $E(A,B)=1$, i.e. $\frac{l^p(A\cap B,\varnothing)}{l^p(A\cup B,\varnothing)}=1$, we have $l^p(A\cap B,\varnothing)=l^p(A\cup B,\varnothing)$. So from the strict monotony of $l^p(A,\varnothing)$-distance，we have $A\cap B=A\cup B$, that is $A=B$. Hence we have $E(A,B)=1\Leftrightarrow A=B$.

(4) If $A\subseteq B\subseteq C$, then

$$l^p(A\cap B,\varnothing)=l^p(A\cap C,\varnothing)=l^p(A,\varnothing);$$
$$\le l^p(B\cap C,\varnothing)=l^p(B,\varnothing)$$

And

$$l^p(A\cup B,\varnothing)=l^p(B,\varnothing)\le l^p(A\cup C,\varnothing)$$
$$=l^p(B\cup C,\varnothing)=l^p(C,\varnothing)$$

So we have

$$\frac{l^p(A\cap C,\varnothing)}{l^p(A\cup C,\varnothing)}=\frac{l^p(A,\varnothing)}{l^p(C,\varnothing)}\quad ,$$
$$\le\frac{l^p(A\cap B,\varnothing)}{l^p(A\cup B,\varnothing)}=\frac{l^p(A,\varnothing)}{l^p(B,\varnothing)}$$

And

$$\frac{l^p(A\cap C,\varnothing)}{l^p(A\cup C,\varnothing)}=\frac{l^p(A,\varnothing)}{l^p(C,\varnothing)}$$
$$\le\frac{l^p(B\cap C,\varnothing)}{l^p(B\cup C,\varnothing)}=\frac{l^p(B,\varnothing)}{l^p(C,\varnothing)}$$

That is $E(A,C)\le\min\{E(A,B),E(B,C)\}$.

From (1)-(4), it is known that

$$E(A,B)=\deg ree(A=B)=\frac{l^p(A\cap B,\varnothing)}{l^p(A\cup B,\varnothing)}$$

is a similarity measure.

3. REFERENCES

[1]. P.Z.Wang, Theory of fuzzy sets and their Applications(Shanghai Science and Technology Publishing House,1982).

[2]. Zeng Wenyi and Li Hongxing, Research on the Relation between Degree of Fuzziness and Degree of Similarity. XiTong LiLun Yu ShiJian, 1995,14(6): 76-79.

[3]. Liu Wenbin and Fan Zong, Unity and Development of several elementary similarity degrees. Pure and applied Mathematics, 1990, 6(6): 85-87.

[4]. Liu Xuecheng, Entropy, distance measure and similarity measure of fuzzy sets and their relations, Fuzzy Sets and Systems, 1992, 52:305-318.

[5]. Lin Zhong-zhen, σ -Entropy, σ -similarity and σ -distance. Journal of Jinan University (Natural Science), 2002, 23(3):8-14.

[6]. B. Kosko, Fuzzy Entropy and Conditioning, Infor. Sci. 1986, 40:165-174.

[7]. CHUA-CHIN WANG and HON-SON DON, A Modified Measure for Fuzzy Subsethood, Infor. Sci. 1994, 79:223-232.

[8]. D.Bhandari and N.Pal, Some new Information Measure for Fuzzy Sets. Infor. Sci. 1993, 67:209-228.

[9]. Yan Deqin, A new approach for information measures on fuzzy sets. Pattern Recognition and Artificial Intelligence, 2001, 14(1): 23-26.

Wei Yiliang was born on October 1960, and earned Chemical Engineering Bachelor Degree in 1982, at the Department of Chemical Engineering, Zhejiang University, China, earned Chemical Engineering master degree in 1988. The major fields of study are Chemical Engineering theory and its application. Now, He is a associate professor at the Department of Biotechnology and Chemical Engineering, Wuhan Polytechnic University, Wuhan, Hubei Province, China.

Common Function Expression of Similarity Measure of Fuzzy Set at l^p – distance

Wei Yiliang
Wuhan Polytechnic University
Wuhan, Hubei, Postcode 430023, The People's Republic of China
Email: weiyiliang@yahoo.com.cn **Tel:** +86(0)2783955676

ABSTRACT

The common function expressions of similarity measure and σ –similarity measure of the fuzzy set in the condition l^p -distance are presented in this paper. The calculating formula of similarity measure and σ –similarity measure of the fuzzy set at l^p -distance are made more perspicuous, which is the underlying application to pattern recognition, fuzzy clustering and information and image processing.

Keywords: fuzzy set, σ–similarity measure, l^p -distance.

1. INTRODUCTION

The similarity measure of fuzzy set (mutual subset hood) is applied widely to pattern recognition, fuzzy clustering and information and image processing, so the study on the similarity measure of fuzzy set has never been disconnected. The similarity measure of fuzzy set was put forward by Wang P. Z. [1] from China for the first time, and then many researchers [2-6] made a lot of research work about the concept and calculation of similarity measure by two approaches: 1). Banding the l^p - distance and difference measure together by Kosko [3], Wang [4]; 2). Performing completely by l^p -distance, Liu [5] defined σ –similarity measure, which has a good σ –addivity and directly brought l^1 -distance into the definition of similarity measure based on reference [6], and then gave a series of properties of the similarity measure. However, what is the common function expression of similarity according to this definition? What is the other relationship about this similarity measure except the same monotony as l^1 -distance? And what is the common function expression of the similarity measure and σ –similarity measure at l^p - distance? With these questions, the common function expression is suggested in this paper and this expression shows completely the relationship between the similarity measures and l^p - distance and the limitation of this similarity measure.

2. THE FUNCTION EXPRESSION OF THE SIMILARITY MEASURE AT l^1

Definition 1. [1,2] Let $X=\{x_1,x_2,\Lambda\ ,x_n\}$, $A,B,C\in F(X)$, $F(X)$ is the class of all fuzzy sets of X , $m_A(x_i)$ is the subjection measure of element x_i about fuzzy set A, $P(X)$ is the class of all crisp sets of X . Prescribe the map:

$$E \quad : F(X)\times F(X)\rightarrow[0,1], \quad A,B\rightarrow E(A,B)$$

Satisfying：

(1) $E(A,B)=E(B,A)$;
(2) If $A(x)\in\{0,1\}$, then $E(A,A^c)=0$;
(3) $E(A,B)=1\Leftrightarrow A=B$;
(4) If $A\subseteq B\subseteq C$, then $E(A,C)\le\min\{E(A,B),E(B,C)\}$
Then calling $E(A,B)$ as the similarity measure of A,B .

Perfecting the above definition, reference[6] put forward the following definition：

Definition 2. [6] Define the map E as following.

$$E \quad : F(X)\times F(X)\rightarrow[0,1], \quad A,B\rightarrow E(A,B)$$

Satisfying:

(1) $E(A,A)=1$;
(2) If $A(x)\in\{0,1\}$, then $E(A,A^c)=0$;
(3) $\forall A,B,C,D\in F(X)$, If satisfying

$$\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|\ge\sum_{i=1}^{n}|m_C(x_i)-m_D(x_i)|$$

And then $E(A,B)\le E(C,D)$.
We define E as a similarity on $F(X)$.

Definition 3. [5] Let E be a similarity measure on F . We call a σ –similarity measure on F , if for any $\forall A,B\in F(X)$ and $D\in P(X)$, there holds

$$E(A,B)=E(A\cap D,B\cup D^c)+E(A\cap D^c,B\cup D)$$

In order to distinguish the above two definitions of the similarity measure, the similarity measure in definition 1 is marked as E_1 , the similarity measure in definition 2 is marked as E_2^1. A series of properties and examples about E_1 and E_2^1 is presented in Refs. [1,5,6]. The following is some commonly useful similarity measure：

$$E(A,B)=\frac{\sum C(A\cap B)}{\sum C(A\cup B)} \tag{1}$$

Where $C(A)=\sum m_A(x_i)$.

$$E(A,B)=\frac{1}{n}\sum C(\frac{A\cap B}{A\cup B}) \qquad (2)$$

$$E(A,B)=1-(\frac{1}{n}\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|^p)^{1/p} \qquad (3)$$

$$E(A,B)=\frac{2\sum C(A\cap B)}{\sum C(A)+\sum C(B)} \qquad (4)$$

For two fuzzy sets A,B , their l^p -distance are：

$$l^p(A,B)=(\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|^p)^{1/p}$$

From which the norm can be derived:

$$||(\sum_{i=1}^{n}(m_A(x_i))^p)^{1/p}||_p$$

Which is marked simply as $|A|$.

As to the function E_2^1 , we have the following theorems:

Theorem 1. the similarity measure of E_2^1 may be expressed as：

$$E_2^1(A,B)=f(|A-B|)$$

Note: From the definition (3) of E_2^1 , it is known that for arbitrary fuzzy sets A,B,C,D , if $|A-B|=|C-D|$, then $E_2^1(A,B)=E_2^1(C,D)$. If let

$$C=\{|A-B|,0,0,L\ ,0\},\ D=\varnothing,$$

we have：

$$\begin{aligned}&E_2^1(A,B)=E_2^1(C,D)\\&=E_2^1(\{|A-B|,0,0,L\ ,0\},\varnothing)\\&=f(|A-B|)\end{aligned}$$

Theorem 2. if the function $f(x)$, $x\in[0,n]$, meets $f(0)=1, f(n)=0$, and $f(x)$ is the monotonous decreasing function, then $\forall A,B\in F(X)$, $E_2^1(A,B)=f(|A-B|)$ is the similarity of definition 2.

Note: Validating directly by definition 2 ：

(1) $E_2^1(A,A)=f(|A-A|)=f(0)=1$ ；
(2) If $m_A(x)\in\{0,1\}$, then

$$E_2^1(A,A^c)=f(|A-A^c|)=f(n)=1 \qquad (3)$$

$\forall A,B,C,D\in F(X)$ and

$$\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|\ge\sum_{i=1}^{n}|m_C(x_i)-m_D(x_i)|,$$

we have

$$E_2^1(A,B)=f(|A-B|)\le f(|C-D|)=E_2^1(C,D)$$

So the conclusion is proved.

If the condition (3) of the definition 2 is modified as: for some $p\ge1$, if satisfying

$$\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|^p\ge\sum_{i=1}^{n}|m_C(x_i)-m_D(x_i)|^p$$

then $E(A,B)\le E(C,D)$, and the similarity measure at this l^p is simply marked as E_2^p , then we may obtain similarly the following theorem:

Theorem 3. The similarity measure of E_2^p are all expressed as: $E_2^p(A,B)=f(||A-B||_p^p)$; Reversely, if function $f(x)$, $x\in[0,n]$, meets $f(0)=1, f(n)=0$, and $f(x)$ is a monotonous decreasing function , then for $\forall A,B\in F(X)$, $E_2^p(A,B)=f(||A-B||_p^p)$ is the similarity measure of E_2^p .

From theorem 1，it is not difficult to find that at $p\ne1$ (1), (2) and (3) do not all meet definition 2，so it is not the similarity measure by definition 2. The following is the properties of the similarity measure of E_2^1 by Ref. [6]:

Proposition.

(1) $\forall A,B\in F(X),\ E_2^1(A,B)=E_2^1(A^c,B^c)$;
(2) $\forall A,B\in P(X),\ E_2^1(A,B)=E_2^1(A^c,B^c)$;
(3) $\forall A\in F(X),B\in P(X),\ E_2^1(A,F)\ge E_2^1(B,F)$;
(4) $\forall A\in F(X),\ E_2^1(A,F)=E_2^1(A^c,F)$;
(5) $\forall A,B\in F(X),\ E_2^1(A\cup B,A\cap B)=E_2^1(A,B)$;

The above properties about E_2^1 by theorems 1, 2 are clear, for $\forall p\ge1$, the properties of E_2^p -based similarity measure are similar completely with those of E_2^1 , for $\forall p\ge1, E_2^p$ -based σ –similarity measure has the following conclusion：

Theorem 4. $\forall p\ge1$, the E_2^p -based similarity measure is σ –similarity measure, if

$$E_2^p(A,B)=f(||A-B||_p^p)=1-\frac{1}{n}\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|^p$$

Proof. Let $C=\{x^{1/p}/x_1,y^{1/p}/x_2,0/x_3,L\ ,0/x_n\}$ and $D=\{1/x_1,0/x_2,L\ ,0/x_n\}$. According to the definition of

σ –similarity measure, we have

$$E_2^p(C,\varnothing)=E_2^p(C\cap D,\varnothing\cup D^c)+E_2^p(C\cap D^c,\varnothing\cup D)$$

From Theorem 3, there exist a function f, which satisfy

$$E_2^p(A,B)=f(\|A-B\|_p^p)$$

Thus, we can get

$$f(x+y)=f(n-1+x)+f(1+y)$$

And differentiate above formula for x and y separately, there is

$$f'(x+y)=f'(n-1+x)=f'(1+y)$$

So $f'(x)=c$ and from

$$f(x+y)=f(n-1+x)+f(1+y),$$

we have

$$c(x+y)+b=c(n-1)+cx+b+c+cy+b$$

Then $f(x)=1-\frac{x}{n}$. From

$$E_2^p(A,B)=f(\|A-B\|_p^p),$$

we obtain

$$\begin{aligned}E_2^p(A,B)&=f(\|A-B\|_p^p)\\&=1-\frac{1}{n}\sum_{i=1}^{n}|m_A(x_i)-m_B(x_i)|^p\end{aligned}.$$

Finally, comparing definition 1 to definition 2, we may conclude:

(1) Definition 2 is the special case of definition1；

(2) Definition 2 modified the (4) in definition 1 as the (3) in definition 2，so the similarity measure is defined completely by distance. As described by theorem 1, it is well known that distance satisfies the immutability of parallel-shift. Therefore, This limits the relationship between the similarity measure and fuzzy set to a large extent.

3. REFERENCES

[1]. P.Z.Wang, Theory of fuzzy sets and their Applications(Shanghai Science and Technology Publishing House,1982).

[2]. Zeng Wenyi and Li Hongxing, Research on the Relation between Degree of Fuzziness and Degree of Similarity. XiTong LiLun Yu ShiJian,1995,14(6) :76-79.

[3]. B. Kosko, Fuzzy entropy and conditioning, Infor. Sci. 1986, 40:165-174.

[4]. CHUA-CHIN WANG and HON-SON DON, A Modified Measure for Fuzzy Subsethood, Infor. Sci. 1994, 79:223-232.

[5]. Liu Xuecheng, Entropy, distance measure and similarity measure of fuzzy sets and their relations, Fuzzy Sets and Systems, 1992, 52:305-318.

[6]. Lin Zhong-zhen, σ -entropy, σ -similarity and σ -distance, journal of Jinan University, 2002, 3:8-14.

Wei Yiliang was born on October 1960, and earned Chemical Engineering Bachelor Degree in 1982, at the Department of Chemical Engineering, Zhejiang University, China, earned Chemical Engineering master degree in 1988. The major fields of study are Chemical Engineering theory and its application. Now, He is a associate professor at the Department of Biotechnology and Chemical Engineering, Wuhan Polytechnic University, Wuhan, Hubei Province, China.

Equality of Vector and Similarity Measure of Fuzzy Sets *

Tong Xiaojun[1], Xu Xiaozeng[2], Li Zhijun[3]
[1]Department of Mathematics and Physics Wuhan Polytechnic University
Wuhan, Hubei, Postcode 430074, the People's Republic of China
[2] School of Information, Xi'an University of Finance and Economics, Xian, Sannxi, Postcode 710061
[1,2,3]Department of Control Science and Engineering, Huazhong University of Science and Technology
Wuhan, Hubei, Postcode 430074
Email: tongxiaojun1998@yahoo.com.cn Tel.: +86(0)2762097718(+862783937411)

ABSTRACT

Almost present similarity measure of fuzzy set is rooted in the relation of equivalence of the equality of sets. A fuzzy set can be considered a vector, and make use of the equality of two vectors, we propose a similarity measure of two fuzzy sets, which do not meet the relation of equivalence of the equality of sets.

Keywords: fuzzy set, similarity measure of fuzzy set, vector.

1. Introduction

If tracking the method of similarity measure of fuzzy set, one can find that almost similarity measure of two fuzzy sets A and B meet $S(A,B)=S(A\cup B, A\cap B)$ or $S(A^c,B^c)=S(A,B)$, or meet them two. For example, Kosko [1], Wang [2] have defined the mutual subsethood make use of factor of proportionality, which satisfy $S(A,B)=S(A\cup B, A\cap B)$, and all similarities measure of fuzzy set that are proposed according as l^p -distance [3,4] of two vectors (points) or fuzzy entropy [5] meet $S(A^c,B^c)=S(A,B)$. Tong [6] has put forward some similarity measure of fuzzy set make use of fuzzy entropy, which meet $S(A,B)=S(A\cup B, A\cap B)$, but not $S(A^c,B^c)=S(A,B)$. Is there some similarity measure of fuzzy set which do not meet $S(A,B)=S(A\cup B, A\cap B)$ and $S(A^c,B^c)=S(A,B)$? Let $X=\{x_1,x_2,\cdots,x_n\}$ be a set. Kosko [1] represented fuzzy subsets of X as fit vectors or fuzzy message. "Fit" seems appropriate because it contracts "fuzzy unit" in the way that "bit" contracts "binary unit" and because a fit value measures the degree to which an element x fits in or belongs to a subset A. In a word, a fuzzy set of X can be represent by vector (we call it fuzzy vector): $\{\mu_A(x_1),\mu_A(x_2),\cdots,\mu_A(x_n)\}$, where membership value $\mu_A(x_n)$ is element x_n belong to the fuzzy set A. So we can use the equality of vector represent similarity measure of fuzzy set. Through this idea, we propose some similarity measure of fuzzy set, which do not meet

$$S(A,B)=S(A\cup B, A\cap B)$$

And

$$S(A^c,B^c)=S(A,B).$$

2. Preliminary

For the two fuzzy sets A,B on $X=\{x_1,x_1,\cdots,x_n\}$, their l^p -distance is：

$$l^p(A,B)=(\sum_{i=1}^{n}|\mu_B(x_i)-\mu_A(x_i)|^p)^{1/p}$$

$F(X)$ is the class of all fuzzy sets of X; $\wp(X)$ is the class of all crisp sets of X; $\left[\frac{1}{2}\right]_X$ is the fuzzy set of X for which $m_{[\frac{1}{2}]_X}(x)=\frac{1}{2}$, $\forall x\in X$. F is a sub-class of $F(X)$ with (1) $\wp(X)\subset F$, (2) $\left[\frac{1}{2}\right]_X\in F$, (3) $A,B\in F\Rightarrow A\cup B\in F, A^c\in F$, where A^c is the complement of $A\in F$, i.e. $\mu_{A^c}(x)=1-\mu_A(x)$, $\forall x\in X$. The number n is the member of the element of set X.

Definition 2.1. [3,5] A real function S: $F^2\rightarrow R^+$ is called a similarity measure of fuzzy set on F if S satisfies the following properties:

(D1) $S(A,B)=S(B,A)$, $\forall A,B\in F$;

(D2) If $A(x)\in\{0,1\}$, then $S(A,A^c)=0$;

(D3) $S(A,A)=1$;

(D4) If $A\subseteq B\subseteq C$, then

$$S(A,C)\le\min\{S(A,B),S(B,C)\}.$$

3. Analysis of the present similarity measure of fuzzy sets

The following is some commonly useful similarity measure：

$$S(A,B)=\frac{\sum C(A\cap B)}{\sum C(A\cup B)} \quad (1)$$

$$S(A,B)=\frac{1}{n}\sum C(\frac{A\cap B}{A\cup B}) \quad (2)$$

$$S(A,B)=1-(\frac{1}{n}\sum_{i=1}^{n}|\mu_A(x_i)-\mu_B(x_i)|^p)^{1/p} \quad (3)$$

$$S(A,B)=\frac{2\sum C(A\cap B)}{\sum C(A)+\sum C(B)} \quad (4)$$

For the similarity measure of fuzzy set is defined by distance, following propositions are obvious:

＊ This work is supported by National Natural Science Foundation under Grant 79970025 of China, Foundation of Hubei Provincial Department of Education under Grant 2003X130 and 2004D005, and Scientific Research Team of Wuhan Polytechnic University under Grant 03-T-06.

Proposition 3.1. If

$$S(A,B)=1-(\frac{1}{n}\sum_{i=1}^{n}|\mu_A(x_i)-\mu_B(x_i)|^p)^{1/p},$$

then

$$S(A,B)=S(A\cup B,A\cap B),\quad S(A^c,B^c)=S(A,B)$$

Proposition 3.2. If $S(A,B)$ is defined as (1) or (2) or (4), then

$$S(A,B)=S(A\cup B,A\cap B),\quad S(A^c,B^c)\neq S(A,B)$$

Liu [5] defined following similarity measure of fuzzy set using fuzzy entropy:

$$S(A,B)=\frac{1}{2}(e(R_1)-e(R_2)+e(R_3)+e(R_4)$$
$$+e(R_5)+e(R_6)-e(R_7)+e(R_8))-1 \quad (5)$$

Where $e(A)$ is the fuzzy entropy of fuzzy set A, and the meaning of R_i please see [4], and he has proved following conclusion:

Proposition 3.3. If $S(A,B)$ is defined as (5), then

$$S(A,B)=S(A\cup B,A\cap B)$$

For this similarity measure of fuzzy set, Tong [6] has proved:

Proposition 3.4. If $S(A,B)$ is defined as (5), then

$$S(A^c,B^c)=S(A,B)$$

And in [6], Tong proposed a similarity measure of fuzzy set is:

$$S(A,B)=1-\left[e(\frac{A\cup B}{2})-e(\frac{A\cap B}{2})\right] \quad (6)$$

Where the membership function of the fuzzy set $\frac{A}{2}$ is $\mu_{\frac{A}{2}}(x)=\frac{\mu_A(x)}{2}$. Tong [6] has gotten following proposition:

Proposition 3.5. If $S(A,B)$ is defined as (6), then

$$S(A,B)=S(A\cup B,A\cap B),\quad S(A^c,B^c)\neq S(A,B)$$

4. The definition of new similarity measure of fuzzy set

Let us first review the equality of two vectors $T_1=\{x_1,x_2,\text{L}\ ,x_n\}$ and $T_2=\{y_1,y_2,\text{L}\ ,y_n\}$, the requirement of the equality of two vectors are in two following aspects:

1)

$$\cos(T_1,T_2)=\frac{T_1\cdot T_2}{|T_1|\cdot|T_2|}$$
$$=\frac{\sum x_iy_i}{(\sum x_i^2\sum y_i^2)^{1/2}};$$
$$=1$$

2)

$$|T_1|=|T_2|\Rightarrow\frac{\sum x_i^2}{\sum y_i^2}=1\cdot$$

According to the front 1) and 2), we can define similarity measure of two fuzzy sets A and B as following:

$$S(A,B)=\frac{\sum\mu_A(x_i)\mu_B(x_i)}{(\sum\mu_A{}^2(x_i)\sum\mu_B{}^2(x_i))^{1/2}}$$
$$\times\min\{\frac{\sum\mu_A{}^2(x_i)}{\sum\mu_B{}^2(x_i)},\frac{\sum\mu_B{}^2(x_i)}{\sum\mu_A{}^2(x_i)}\} \quad (7)$$

And when $\sum\mu_A{}^2(x_i)=0$ or $\sum\mu_B{}^2(x_i)=0$, $S(A,B)=0$.

Theorem 4.1. If $S(A,B)$ is defined as (7), then $S(A,B)$ is a similarity measure of two fuzzy sets A and B.

Proof. 1) Obviously we can get $S(A,B)=S(B,A)$;

2) When $D\in\wp(X)$, $\sum\mu_D(x_i)\mu_{D^c}(x_i)=0$, so $S(D,D^c)=0$;

3) Obviously, we have $S(A,A)=1$;

4) $\forall A,B,C\in F$, if $A\subseteq B\subseteq C$, Let $\sum$ and $\sum_m$ stand for $\sum_{i=1}^{n}$ and $\sum_{i=1}^{m}$ respectively, we have

$$S(A,B)=\frac{\sum\mu_A(x_i)\mu_B(x_i)}{(\sum\mu_A{}^2(x_i)\sum\mu_B{}^2(x_i))^{1/2}}$$
$$\times\min\{\frac{\sum\mu_A{}^2(x_i)}{\sum\mu_B{}^2(x_i)},\frac{\sum\mu_B{}^2(x_i)}{\sum\mu_A{}^2(x_i)}\}$$
$$=\frac{\sum\mu_A(x_i)\mu_B(x_i)}{(\sum\mu_A{}^2(x_i)\sum\mu_B{}^2(x_i))^{1/2}}\frac{\sum\mu_A{}^2(x_i)}{\sum\mu_B{}^2(x_i)},$$

$$S(A,C)=\frac{\sum\mu_A(x_i)\mu_C(x_i)}{(\sum\mu_A{}^2(x_i)\sum\mu_C{}^2(x_i))^{1/2}}$$
$$\times\frac{\sum\mu_A{}^2(x_i)}{\sum\mu_C{}^2(x_i)}$$

And

$$\mu_A(x_i)\le\mu_B(x_i)\le\mu_C(x_i),(i=1,2,\text{L}\ ,n)$$

For showing $S(A,B)\ge S(A,C)$, we can only prove

$$\frac{\sum\mu_A(x_i)\mu_B(x_i)}{(\sum\mu_B{}^2(x_i))^{3/2}}\ge\frac{\sum\mu_A(x_i)\mu_C(x_i)}{(\sum\mu_C{}^2(x_i))^{3/2}}$$

Define

Where $p_i\ge 0$, and $p_i\le t_i\le 1,(i=1,2,\text{L}\ ,n)$, since

$$\frac{\partial\ln f}{\partial t_k}=\frac{p_k}{\sum p_it_i}-\frac{3t_k}{\sum t_i^2},(k=1,2,\text{L}\ ,n)$$

We have $\sum\frac{\partial\ln f}{\partial t_k}t_k=-2<0$, thus there is not any point of the function $f(t_1,t_2,\text{L}\ ,t_n)$. For boundary points, we can define

$$f(t_1,t_2,\text{L}\ ,t_m)=\frac{c+\sum_m p_it_i}{(c+\sum_m t_i^2)^{3/2}}$$

Similarly, we have

$$\sum\frac{\partial\ln f}{\partial t_k}t_k=\frac{3c}{c+\sum_m t_i^2}-\frac{c}{c+\sum_m p_it_i}-2$$
$$<0$$

And

$$S(A,A)=1$$

Therefore

$$S(A,B) \geq S(A,C)$$

The proof of the conclusion that $S(A,B) \geq S(A,C)$ is similar.

Theorem 4.2. If $S(A,B)$ is defined as (7), then

1) $S(A^c, B^c) \neq S(A,B)$;

2) $S(A,B) \neq S(A \cup B, A \cap B)$.

Proof. We can take fuzzy sets A and B as

$$\mu_A(x_1) = \frac{1}{2}, \mu_A(x_2) = \frac{1}{4},$$
$$\mu_A(x_k) = 0, (3 \leq k \leq n)$$
$$\mu_B(x_1) = \frac{1}{3}, \mu_A(x_2) = 1,$$
$$\mu_B(x_k) = 0, (3 \leq k \leq n)$$

1) We have

$$S(A,B) = \frac{9\sqrt{2}}{64} \approx 0.198874$$

And

$$S(A^c, B^C) = \frac{4\sqrt{13}}{507} \approx 0.02846$$

Therefore

$$S(A^c, B^c) \neq S(A,B)$$

2) We can get

$$\mu_{A\cup B}(x_1) = \frac{1}{2},$$
$$\mu_{A\cup B}(x_2) = 1,$$
$$\mu_{A\cup B}(x_k) = 0, (3 \leq k \leq n)$$
$$\mu_{A\cap B}(x_1) = \frac{1}{3},$$
$$\mu_{A\cap B}(x_2) = \frac{1}{4},$$
$$\mu_{A\cap B}(x_k) = 0, (3 \leq k \leq n)$$

Hence

$$S(A,B) = \frac{9\sqrt{2}}{64} \approx 0.198874$$

And

$$S(A \cup B, A \cap B) = \frac{\sqrt{5}}{18} \approx 0.124226$$

That is

$$S(A,B) \neq S(A \cup B, A \cap B).$$

5. REFERENCES

[1] B. Kosko, Fuzzy Entropy and Conditioning, Infor. Sci. 1986, 40:165-174.

[2] CHUA-CHIN WANG and HON-SON DON, A Modified Measure for Fuzzy Subsethood, Infor. Sci. 1994, 79:223-232.

[3] P.Z.Wang, Theory of fuzzy sets and their Applications(Shanghai Science and Technology Publishing House,1982).

[4] Liu Wenbin and Fan Zong, Unity and Development of several elementary similarity degrees. Pure and applied Mathematics, 1990, 6(6): 85-87.

[5] Liu Xuecheng, Entropy, distance measure and similarity measure of fuzzy sets and their relations, Fuzzy Sets and Systems, 1992, 52:305-318.

[6] Tong Xiaojun, Xu Xiaozeng, Li Zhijun, Definition and Research on a new similarity measure of fuzzy sets, to press.

[7] Zeng Wenyi and Li Hongxing, Research on the Relation between Degree of Fuzziness and Degree of Similarity. XiTong LiLun Yu ShiJian, 1995,14(6): 76-79.

Tong Xiaojun was born in Shanxi Province, China, on November 12, 1967, and earned Engineering Doctor Degree in Control Theory and Control Engineering in 2002, at the Department of Control Science and Engineering, Huazhong University of Science and Technology, China; the major fields of study are fuzzy theory and its application. He occupied Lecturer and vice professor at the Department of Mathematics, University of Petroleum between 1993 and 2002. Now, He is a full professor at the Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan, Hubei Province, China. His research interests are in fuzzy theory and its application, and research Biomolecular Compute. He dealt with the research of Partial Differential Equation.