DCABES 2009 Proceedings

第八届分布式计算及其应用国际学术研讨会论文集

The 8th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Wuhan, China
October 16~19, 2009

主　编：郭庆平　郭羽成

Editor in Chief: GUO Qingping　GUO Yucheng

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

DCABES 是一个在分布式计算及应用方面多学科汇集的国际学术研讨会，大会将有来自国内外的多名专家、学者及工程师参加，并将邀请知名计算机公司到会进行交流。会议的主题是分布式并行处理及其在工程、科学和电子商务领域的应用。主要包括：并行/分布式技术的应用、并行/分布式算法研究、系统结构、网络计算与基于网络的计算、基于分布式系统的多媒体应用、分布式并行软件工具及平台环境、并行/分布系统的其他方面等内容。

图书在版编目（CIP）数据

第八届分布式计算及其应用国际学术研讨会论文集：英文 / 郭庆平，郭羽成主编.—北京：电子工业出版社，2009.10
ISBN 978-7-121-09595-5

I. 第…　Ⅱ. ①郭…②郭… Ⅲ. 分布式计算机－计算机应用－国际学术会议－文集－英文 Ⅳ.TP338.8-53

中国版本图书馆 CIP 数据核字（2009）第 172318 号

责任编辑：董亚峰
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　　　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：880×1 230　1/16　印张：32.25　字数：1050 千字
印　　次：2009 年 10 月第 1 次印刷
定　　价：118.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。
质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。
服务热线：（010）88258888。

PREFACE

The DCABES is a community working in the area of Distributed Computing and its Applications in Business, Engineering, and Sciences, and is responsible for organizing meetings and symposia related to the fields. The DCABES 2009 is the Eighth International Conference on Distributed Computing and Applications to Business, Engineering and Sciences held on 16-19 September 2009 in Wuhan, Hubei, China. It is the fourth time for the DCABES international conference to be organized by School of Computer Science and Technology, Wuhan University of Technology.

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

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

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

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

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

I am also grateful to Professor Craig C. Douglas（University of Wyoming and University of Yale，USA），Professor Stefan Vandewalle（Katholieke Universiteit, Belgium）， Prof Haixiang Lin（Delft University of

Technology，Netherland）and Dr Robert Lovas（Hungarian Academy of Sciences Mail，Hungary）for their contributions of keynote speeches in the conference.

Sincerely thanks should be forwarded to the China Ministry of Education (MOE), without its support the DCABES 2009 could not be held in China successfully. We would also like to thank the WUT (Wuhan University of Technology, China), the National Parallel Computing Society of China (NPCS), the ISTCA (International Science and Technology Cooperation of Hubei Province, China), and the CAA (Computer Academic Association of Hubei Province & Wuhan Metropolis, China) for their supports as local organizers of the conference.

Finally I should also thank my graduate students, Mr. Jiansheng Lin for his efforts in organizing activities，Mr. Hai Peng for the conference website design and implementation. It also should be mentioned that my graduate students, Ms. Dongling Bai, Mr. Mingming Yang，Chengwei Wu，Haohao Wang，Chuhua Tang，Shouming Xu，Ming Zhou，Zhaohuan Pan and Ms. Qing Wang of the grade 2007 and 2008，spent a lot of time and efforts typesetting the proceedings. Without their help the proceedings could not look so good.

Enjoy your stay in Wuhan, China. Hope to meet you again at the DCABES 2010.

Guo, Professor Qingping
Chair of the DCABES2009
Dept. of Computer Science
Wuhan University of Technology
Wuhan, China

COMMITTEES

Chair of Scientific Committee

Guo, Professor Q. P., Wuhan University of Technology

Co-Chair of Scientific Committee

Lai, Dr. Choi-Hong, University of Greenwich

Chair of Organizer Committee

Guo, Professor Q. P., Wuhan University of Technology

Steering Committee

Guo, Professor Q.P. (Co-Chair) Wuhan University of Technology, Wuhan, China

Lai, Professor C.-H. (Co-Chair) University of Greenwich, UK

Tsui, Thomas. Chinese University of Hong Kong, Hong Kong, China

Xu, Professor W.B. Southern Yangtze University, Wuxi, China

Scientific Committee (in alphabetical order)

Cai, Professor X.C. University of Colorado, Boulder, U.S.A.

Cao, Professor J.W. R＆D Centre for Parallel Algorithms and Software, Beijing, China

Chi, Professor X.B. Academia Sinica, Beijing, China

Craig Professor C. Douglas, University of Wyoming, Mathematics Department, Yale University Computer Science Department

Guo, Professor Q.P. Wuhan University of Technology, Wuhan, China

Ho, Dr. P. T. University of Hong Kong, Hong Kong, China

Jesshope, Professor C. University of Amsterdam, the Netherlands

Kang, Professor L.S. Wuhan University, China

Keyes, Professor D.E. Columbia University, USA

Lai, Professor C.-H. University of Greenwich, UK

Lee, Dr. John. Hong Kong Polytechnic, Hong Kong, China

Liddell, Professor H. M. Queen Mary, University of London, UK

Lin, Dr. H.X. Delft University of Technology, Delft, the Netherlands

Lin, Dr. P. National University of Singapore, Singapore

Loo, Dr. Alfred Hong Kong Lingnan University, Hong Kong, China

Ng, Professor. Michael Baptist University of Hong Kong, China

Sloot, Professor P.M.A. University of Amsterdam, Amsterdam the Netherlands

Sun, Professor J. Academia Sinica, Beijing, China

Tsui, Mr.Thomas Chinese University of Hong Kong, Hong Kong, China

Xu, Professor W. Southern Yangtze University, Wuxi, China

Zhang, Professor J. University of Kentucky, USA

Local Organizing Committee

Guo, Professor Q.P. (Chair) Wuhan University of Technology, Wuhan, China

Zhong, Professor L . (Co-Chair) Wuhan University of Technology, Wuhan, China

Liu, Professor Q. Wuhan University of Technology, Wuhan, China

Xu, Professor H.Z. Wuhan University of Technology, Wuhan, China

Chen, Professor H. Wuhan University of Technology, Wuhan, China

Xu, Professor N. Wuhan University of Technology, Wuhan, China

Zeng, Professor C.N. Wuhan University of Technology, Wuhan, China

Zhang, Proffersor H. M. Wuhan University of Technology, Wuhan, China

He, Proffersor Y. X. Wuhan University, Wuhan, China

King, Professor Hai, Hua Zhong University of Science and Technology, Wuhan, China

Tan, Professor L.S. Central China Normal University, Wuhan, China

Kang, Professor L.S. Wuhan University, Wuhan, China

Lu, Professor J.G. South Central China Nationality University

Secretaries

Mr. Yucheng Guo, Wuhan University of Technology, Wuhan, China

Mr. Jiansheng Lin, Wuhan University of Technology, Wuhan, China

Editorial Board

CONTENTS

HIGH PERFORMANCE COMPUTING AND ALGORITHM DESIGN

DISTRIBUTED/PARALLEL APPLICATIONS

SOFTWARE TOOLS AND ENVIRONMENTS

INTELLIGENT TRANSPORTATION

INFORMATION SECURITY

IMAGE, MEDIA PROCESSING

TRENDS AND CHALLENGES IN HIGH PERFORMANCE COMPUTING

Hai Xiang Lin
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology
Mekelweg 4, 2628 CD, The Netherlands
Email: H.X.Lin@TUDelft

ABSTRACT

In the past 3 decades, high performance computing has undergone a tremendous fast development. The achievement in high performance computing is obtained through the synergy in the development of integrated circuit technology, parallel computing systems and efficient algorithms. Performance of a single processor has been increased by a factor of 10^6 since the introduction of the first electronic computer in the late 1940s, the application of massively parallel processing adds an increase by another factor of 10^5, and moreover the development of fast and efficient algorithms further accelerate the solution speed by factor of 10^6 or more for many large and complex computer applications. The current trends in parallel architectures moving towards multi-cores, much larger number of processors or cores, and heterogeneity pose great challenges to software and tool developers, to algorithm and application programmers.

Keywords: Parallel and distributed computing, Multi-core, GPU, HPC systems and algorithms, technological development in HPC

1. INTRODUCTION

Although there is not an exact date which one can speak as the beginning of high performance computing (HPC), generally speaking we may regard the introduction of the first supercomputer[1] Cray 1 in the mid of 70s of the last century as the starting point of the HPC era. The demands for increasingly powerful and faster computer systems have always been the driving forces in the development of supercomputers. In the past 3 decades, high performance computing has undergone a tremendous fast development. Although the performance of the fastest computer has been increased by a factor of 10^9 or more, the hunger for more computer power by computational scientists and engineers has not been satisfied. As more and more disciplines learned how to achieve their research goals and see more deeply into nature by using larger and more complex (mathematical) models with more details and higher accuracy, the demands for high-end computing power has become increasingly insatiable. The trends in parallel architectures moving towards multi-cores, much larger number of processors or cores, and heterogeneity pose great challenges to software and tool developers, algorithm designers and application programmers. It is widely known that the use of parallel computers with a larger number of processors does not automatically result in a shorter solution time for your application. In order to obtain good parallel performance, algorithms with high degree of parallelism and parallel programs with high data locality must be carefully designed. Due to the memory hierarchy (multi-level caches, local and remote distributed memories, etc.), optimization to data movements to increase the data locality during the computation is critical to the performance. Moreover, programmability of large systems with tens of thousands up to a million processors or cores remains a big challenge. In contrast to the development in hardware systems, parallel compiler technology is still poorly developed. Compiling for optimal parallel execution turns out to be a very hard problem to tackle. After decades of active research in parallel languages and parallel compilers, it has been relatively quite in the last decade and only few groups are now still doing research in this area.

In this talk, we will review the trends in parallel systems and recent architectural development, challenges and some ongoing projects in attacking the problems of software and tools for HPC, and discuss the needs for new algorithms which can scale to tens of thousands or even millions of processors.

2. PARALLEL ARCHITECTURE

2.1 A Brief History of Supercomputing

The introduction of the vector computer systems (e.g., Cray 1, CDC) marked the beginning of modern supercomputing. These vector supercomputers offer a performance advantage of at least one order of magnitude over the conventional systems of that time, and they are high-end expensive systems with a price tag of more than one million US dollar. Raw or peak performance in MFlop/s was the main if not the only selling argument. Then in the first half of eighties, we have seen the integration of vector system (vector multi-processors) in conventional computing environments. Strohmaier et al [1] have made a summary on recent trends in the marketplace of HPC using the data of supercomputer systems on the list of TOP500 [2]. The lists of the world top 500 fastest computers through the years give a quite representative history of the evolution of supercomputer systems.

At the end of the eighties, massively parallel computing (MPP) with scalable systems using distributed memory has been widely introduced. The main goal of MPPs was to overcome the hardware scalability limitations of shared memory architectures. In contrast to the earlier HPC computer systems which use custom designed special and expensive processors, MPP systems use “off-the-shelf” standard microprocessors[2]. As a result of this, an MPP system costs only a fraction of the earlier vector supercomputers.

At the beginning of nineties the Symmetric Multi-Processors (SMP) systems started to gain popularity. The SMP systems produced by various workstation manufacturers as well as the IBM SP series targeted the lower and medium market segments, their price/performance ratio were better due to cost advantages of the larger production numbers. In the first two

[1] Supercomputers are typically computer systems with a performance of at least one order of magnitude higher than conventional large servers and mainframes at that time.

[2] For instance, Intel processor which was used only for PCs has entered the market of high-end HPC systems. This has remained and even now there is a large number of cluster supercomputers using AMD and Intel processors on the list of TOP500.

decades, HPC systems have been deployed only for large scale scientific and engineering applications. The introduction of MPP systems with the shift to the market of medium performance resulted in the acceptance of parallel and distributed computing by new commercial applications such as database applications. At the same time, we see that due to the low volume the traditional high-end supercomputer manufacturers can no longer compete in the market. Nowadays no single stand alone supercomputer vendor still exists, many of them are merged with other traditional computer companies or went bankrupt.

2.2 Current trends

The landscape of the supercomputers and other fast HPC systems in the first decade of the 21st century is one of clustered systems build from commodity parts: IBM Cell, AMD and Intel processors. The problems with the heat dissipation severely limit the possibility of making ever faster processor chips, microprocessor manufacturers started to make multi-cores in a single chip in order to increase the processing power. In fact the clock frequency on the micro processors has been a kind of frozen since 2004, instead we are seeing processors with an increasing number of cores on a single chip. Now already there are implementations like the Cell with 8 (actually 9) cores. On the other hand, GPUs with 32 or 64 cores are becoming common. The trends is that HPC systems with thousands simple cores will become commodity goods, e.g., a PC with a GPU accelerator with a theoretical peak performance of 1 TFlops/s or above is now affordable for everyone.

Both multi-core CPUs and GPUs are used in HPC systems. GPUs, as a new competitor, are fast and will get a lot faster in performing certain scientific computing operations. In terms of performance-per-dollar GPUs are cheap and use less power than CPUs when compared on a performance-per-watt basis. However, they are only good in performing certain type of number crunching tasks. Another problem with GPUs is their limited programmability as compared to CPUs. Therefore we foresee that in the nearest future scenario of HPC systems that many hardware architectures will likely be a combination of specialized CPU and GPU type cores.

Data volumes will explode, and data centers face a big challenge how to store the increasing huge amount of data in a secure, fault tolerant and energy efficient way. We expect to see an expansion of HPC systems to small- and midsized companies. U.S. remains the most important player in HPC market both as manufacturer and as consumer. Companies from the U.S. (IBM, Cray, SGI, SUN, etc.) has together a market share of more than 90%, and more than 50% of the Top500 supercomputers are installed in the U.S. Europe remains in the second place in terms of the number of installed Top500 supercomputers (with a share of slightly less than 30%), however, it plays no role as a producer of HPC systems. A trends which can be observed is that parallel with the economic growth in Asian countries like China and India, HPC systems are increasingly been deployed. Moreover, China as one of the few countries has her own HPC system manufacturers.

The largest supercomputer systems today contain $O(10^5)$ cores, but soon there will be systems having $O(10^6)$ cores. Moreover, there will be not only a few but a large number of these really massive parallel systems. In fact, the IBM Sequoia-System with 1.6 millions of Power-Processors and 1.6 Pbytes of main memory and 20 PFlops/s peak performance is planned to be installed at LLNL in 2011 and will start running in 2012. By 2009, there are two supercomputers in the TOP500 that has passed the Linpack test obtaining the speed of 1 PFlops/s. It is expected that in 10 years time (2018-20) the speed of 1 ExaFlops/s will have been reached.

The major challenges to processor and HPC systems manufacturers are: low cost, low power consumption, (architectural) support to easy programming with high efficiency. Because the power consumption of a large computer system is approaching the capacity of a (small) power plant, a new metric flops/watt is very likely to be added as standard for performance measures of HPC systems.

3. SOFTWARE AND TOOLS

As discussed in Section 2, clustered multi-core systems with core numbers in the range of one-hundred thousands to one million and more will be the common standard for HPC systems in the near future. These massively parallel systems will raise questions about parallel software development and especially to the fault tolerance and reliability.

Programming a computer system with hundreds of thousands or even millions cores will be a difficult task. Obviously in most cases it would be an impossible task for a programmer to directly manage those hundreds of thousands to millions threads. Even more challenging is how to efficiently execute a parallel application on such a massive number of processors/cores. We all know from practice that it is difficult to make our parallel applications scale to the tens of thousands processors, so far only few (optimized) parallel applications can use tens of thousands or more processors efficiently. Furthermore, apart from the performance issues, the effort in implementing and debugging parallel programs running on such a large number of processors/cores is enormous.

Data locality ratio, i.e., the ratio between the amount of computation and communication, has been the key factor in early distributed memory parallel computer systems. As only distributed memory system can scale to the number of tens of thousands and beyond, future HPC systems will remain to be distributed. In addition, the use of multi-cores, Cell processors and GPUs alike, introduces an extra dimension of heterogeneity in the hierarchy of memory. For example, cores within a GPU often share the same common memory. A fundamental problem which faces every multi-core system is the limited bandwidth for the memory access. For instance, the effect of memory latency and bandwidth has been illustrated in [3]. All cores on a multi-core chip are in competition for the memory access. Concerning the memory access, a good solution would be one Byte per Flop and core. If this solution would be available even one hundred cores per chip would not be a problem. With Nehalem from Intel an impressive performance of half a Byte per Flop and Core will be implemented. Yet, the data bandwidth will remain behind the processing speed, and the consequence is performance difference can be as large as 1 order of magnitude between optimized and poorly managed data movements. Optimizing data access is far from trivial, in the past optimizing cache usages and enforcing cache consistency is done by hardware. This hardware mechanism occupies an increasingly larger area on a chip, which is both inefficient and energy consuming. It has lead to the shift to abandon such hardware mechanism in multi-cores architectures. Programmers are now responsible for this task and subsequently have the opportunity to optimize the data flow between caches and memory, the bad news is however this burden adds to the programmer’s already heavily loaded shoulders.

In contrast to the development in hardware systems, parallel compiler technology is still poorly developed. Currently, there is no compiler one can use to compile an application for running on a large distributed computer system. Automatic parallelization using a compiler for OpenMP can only apply on a small SMP (shared memory) system with a few processors. For scalable distributed systems with a large number of processors, the programmer must code his application using MPI (by hand). Compiling for optimal parallel execution turns out to be a very hard problem to tackle. After decades of active research in parallel languages and parallel compilers, the research land of parallel compiler technology has been relatively quite during the last decade and only a few groups are now still doing research in this area. We believe that HPC applications using tens of thousands processors or more can only be successful and be widely accepted if the programmability on such a large system is going to be largely improved. To this end, (renewed) effort must be done on parallel compliers and parallel language constructs (not necessarily a new language) to facilitate automatic parallelization of applications efficiently running on a very large number of processors/cores. In the past, automatic vectorization/parallelization of data parallel type of applications has been successful, and this success may be repeated on systems with tens of thousands processors. However, data parallelism represents only a subset of HPC applications, and past success (i.e., efficient parallelization) obtained on systems with only hundreds or a few thousands processors may not be automatically repeated on a system with ten to hundredfold more processors/core. Just to think of the Amdahl's law, using more cores means the time of the execution of the parallel fraction is becoming much smaller whereas the execution time of the sequential part remains unchanged, this means the total execution time will only decrease a bit by using tens to hundredfold more cores.

A big challenge is to close the gap between the fast development in hardware systems and the related software infrastructure. In order to increase the programmability, new software and tools are needed to help programmers to cope with the use of the large number of cores and to increase data locality via optimizing the data movements. We quote a description from a DOE report on ExaScale supercomputing [4] (p. 103) to underline the urgency of this emerging situation below,

> "The shift from faster processors to multicore processors is as disruptive to software as the shift from vector to distributed memory supercomputers 15 years ago. That change required complete restructuring of scientific application codes, which took years of effort. The shift to multicore exascale systems will require applications to exploit million-way parallelism and significant reduction in the bandwidth and amount of memory available to millions of CPUs. This 'scalability challenge' affects all aspects of the use of HPC. It is critical that work begin today if the software ecosystem is to be ready for the arrival of exascale systems in the coming decade."

Fortunately, the international HPC research community has recognized this problem and a plan called the "International Exascale Software Project (IESP) is now being considered [5]. Of course, more actions as this are needed, and besides tools and software, the more fundamental problems such as programming models, model of parallel computation, etc. also need to be considered.

4. ALGORITHMS AND APPLICATIONS

Talking about future supercomputers with a computing power of ExaFlops/s and millions of processors, we have implicitly assumed that: 1. there are applications which require such a high computing speed; 2. there are applications with a very large degree of parallelism which can effectively utilize the millions of professors. Is this the reality?

The answer to the first question is illustrated in a recent report [6]. Although some of the early so-called grand challenge applications (of the nineties) have now been solved after supercomputers start to provide TFlops/s and PFlops/computing power. Many applications ask for much higher performance. Simulation of climate and atmospheric models is one such field requiring Exascale performance. Just to illustrate: increasing the horizontal mesh resolution by a factor of between 4 and 10 for more accurate weather predictions will require a hundred- to thousand fold increase in computing capability. Other science and engineering fields described in [4] and [6] are astrophysics, energy research (e.g., combustion and fusion), and biology.

The answer to the second question is more complex. If a program requires a huge amount of computations, but exhibits little parallelism then obviously using a HPC system with a large number of processors will not help. And many legacy programs do show such limitations. However, we may then ask is this lack of parallelism inherent in the nature of the application? The answer is often negative. The serialization is often caused by the way we modeled the problem and by the way we coding them. Typically, we have solved and coded the applications by dividing a large complex problem into small subproblems, and then solve them one by one, both for lowering the complexity and simplifying the reasoning process. Nature is not like that, most phenomena in the nature have many physical processes and interactions all happening simultaneously. Therefore, we need to reconsider the mathematical and computational models we use to model physical or social processes. For example, particle models and Monte-Carlo methods exhibit a large degree of parallelism (e.g., compared to models of PDEs). Some of these models are computationally less efficient in the past, but with HPC systems containing millions of processors, the story may be different.

Consider problem solving from top to down, after the model describing the problem at hand has been chosen, the next step is to select and/or design methods and algorithms to be used to solved the problem. Here again, algorithm with large degree of parallelism and good data locality is important for the performance on a large HPC system. Some efficient traditional algorithm may not exhibit sufficient parallelism, in such a case alternative or new algorithms must be designed. Beside the degree of parallelism, another at least as important issue is the data locality and latency tolerating property of a parallel algorithm/program. Latency exists in memory access and in transferring data from one location to another location, and in general this latency is equivalent to the time hundreds to tens of thousands (or more for grid/cloud computing) floating point operations. If the program executes in synchronized fashion, then lots of overhead will be induced due to the big latency delay by the data transfers. At large scale, we are going to need algorithms and programs organized in such a way that computation of a part A continues while the data transfer for part B takes place, in other words (partial) overlapping of data transfer (cache per-fetch, communication) with computation.

REFERENCES

[1] E. Strohmaier, J. Dongarra, H.W. Meuer and H.D. Simon, “Recent Trends in the Marketplace of High Performance Computing”, Parallel Compouting, Vol. 31, pp. 261-273, 2005.
[2] The TOP500 list: www.top500.org.
[3] R.C. Murphy, “On the Effects of Memory Latency and Bandwidth on Supercomputer Application Performance”, in Proc. IEEE 10th International Symposium on Workload Characterization (IISWC 2007), 2007, pp. 35-43.
[4] R. Stevens, T. Zacharia and H. Simon, “Modeling and Simulation at the Exascale for Energy and Environment”, Department of Energy Office of Advance Scientific Computing Research, Washington, DC, U.S.A.. Report on the Advanced Scientific Computing Research Town Hall Meetings on Simulation and Modelling at the Exascale for Energy, Ecological Sustainability and Global Security, 2008. http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/TownHall.pdf.
[5] The International Exascale Software Project: A call to cooperative action by the global high performance community. www.exascale.org.
[6] US National Research Council Committee on the Potential Impact of High-End Computing on Illustrative Fields of Science and Engineering, “The Potential Impact of High-End Capability Computing on Four Illustrative Fields of Science and Engineering”, Washington, DC, U.S.A., 2008.

Hai Xiang Lin is with the Delft Institute of Applied Mathematics (DIAM), Delft University of Technology. He graduated from Delft University of Technology in 1986, from 1986 to 1990 he was a researcher at the Netherlands Organization of Applied Research (TNO). In 1990 he joined the department of applied mathematics as an assistant professor, and he got his PhD in 1993. Currently he is an associate professor in the group of Large Scale Models and Simulation. He has published about 100 refereed journal- and international conference papers. He has been a co-chair of the international conferences Euro-Par 2009, and a general chair of the IEEE 12th international conference on Computational Science and Engineering (CSE-09).

PARALLEL ADI METHOD FOR PARABOLIC PROBLEMS ON GP-GPU

Hyoungseok Chu[1], Hyoseop Lee[1] and Dongwoo Sheen[1,2]
[1] Computational Science and Technology
[2] Department of Mathematics,
Seoul National University
Seoul, 151-747, Korea
[2] Email: sheen@snu.ac.kr

ABSTRACT

In this paper we report some results on implementation of an ADI scheme to solve a three-dimensional parabolic problem by using GP-GPU of NVIDIA. Speedup rates as well as convergence have been investigated in. Both single and double precisions have been implemented and analyzed.

Keywords: Alternating Direction Implicit ADI, GPU, GPGPU, CUDA, Parabolic PDE, Parallel.

1. INTRODUCTION

We consider a parabolic equation in three-dimensional space in the following form

$$u_t(x,t) - \sum_{i,j=1}^{3} \nu_{ij} u_{x_i x_j}(x,t) + \sum_{i=1}^{3} c_i u_{x_i}(x,t) + \gamma u(x,t) = 0, \qquad (1)$$

where $(x,t) \in \Omega \times (0,T]$ and

$$au(x,t) + b\frac{\partial u}{\partial \mathbf{n}}(x,t) = f(x,t), \quad (x,t) \in \partial\Omega \times (0,T],$$
$$u(x,t) = u_0(x), \quad (x,t) \in \Omega \times \{0\},$$

where $\Omega \subset \mathbb{R}^3$ is a rectangular domain, (0,T] is the time interval, and u_0 and f are the initial and boundary data. Depending on the choice of coefficients a and b, the boundary condition may have the Dirichlet, Neumann or Robin type. For Eq. (1) to be a well-posed problem, we restrict the symmetric matrix (ν_{ij}) to a positive definite one.

Due to the curse of dimensionality of the grid, even in the three dimensional case, the given problem has been a challenging one until now. As an effort to reduce computational time and resources, Douglas and Rachford proposed the ADI (alternating direction implicit) method in [4, 2]. There have been many successors of the original ADI method, among them Craig and Sneyd introduced a second order scheme in both time and space including cross and first order derivatives in [1].

One of the benefits of using an ADI method is a parallelizable nature. Since the resulting matrix produced by an ADI method is a set of tridiagonal systems, the parallelized ADI can be easily implemented and guaranteed to offer an almost ideal speedup if the parallel system has a shared memory between parallel processors. A recent progress in GP-GPU provides a cost-effective many-core system with a shared memory. Although the program language, which was limited in assembler level, had prevented from general use in scientific computing, recently the CUDA platform supported by NVIDIA provides a usable programming environment.(cf. [6])

In this paper we describe the parallel implementation of the Craig-Sneyd ADI method to solve general parabolic equations on a GP-GPU architecture. The Craig-Sneyd ADI method is introduced in Section 2, and a brief introduction of NVIDIA CUDA platform and parallel implementation of the method are described in Section 3. In Section 4, we confirm the efficiency of the proposed method numerically and analyze the performance of the method. We conclude with some concluding remarks in Section 5.

2. The CRAIG-SNEYD ADI METHOD

In spite of a variety of successors of the original ADI scheme, it had not been introduced a stable ADI scheme including mixed derivatives before [1]. It is because the cross derivative terms are difficult to handle implicitly using the ADI method as described in [3].

In [1] Craig and Sneyd tackled the complicated parabolic problems including mixed and first derivatives effectively with a sequence of ADI operations. In what follows we consider a uniform rectangular grid with $\Delta x_i = \Delta x$. Let $u^n_{i,j,k}$ denote the value of u at time $n\Delta t$ and the i-th, j-th and k-th position on x_1, x_2 and x_3 axises, respectively, and set $r = \Delta t / \Delta x^2$. Discrete operations are defined as the conventional central difference, namely,

$$\delta_{x_1} u_{i,j,k} = u_{i+1,j,k} - u_{i-1,j,k}$$
$$\delta^2_{x_1} u_{i,j,k} = u_{i+1,j,k} - 2u_{i,j,k} + u_{i-1,j,k}$$
$$\delta_{x_1 x_2} u_{i,j,k} = u_{i+1,j+1,k} - u_{i+1,j-1,k} - u_{i-1,j+1,k} + u_{i-1,j-1,k}$$

Then the Craig-Sneyd ADI method is described as follows:
Algorithm (Craig-Sneyd ADI Method I)
SET NT to be the number of time step
DO n=0 to NT-1

$$(1-\theta\nu_{11} r\delta^2_{x_1}) u^{n+1(1)} = Au^n$$
$$(1-\theta\nu_{22} r\delta^2_{x_2}) u^{n+1(2)} = u^{n+1(1)} - \theta\nu_{22} r\delta^2_{x_2} u^n$$
$$(1-\theta\nu_{22} r\delta^2_{x_3}) u^{n+1} = u^{n+1(2)} - \theta\nu_{33} r\delta^2_{x_3} u^n$$

END DO

To achieve a second order convergence in time discretization, central differences in time of the mixed derivatives are included as a second step,
Algorithm (Craig-Sneyd ADI Method II)
SET NT to be the number of time step
DO n=0 to NT-1

$$(1-\theta\nu_{11} r\delta^2_{x_1}) u^{n+1(1)}_{(1)} = Au^n$$
$$(1-\theta\nu_{22} r\delta^2_{x_2}) u^{n+1(2)}_{(1)} = u^{n+1(1)} - \theta\nu_{22} r\delta^2_{x_2} u^n$$
$$(1-\theta\nu_{22} r\delta^2_{x_3}) u^{n+1}_{(1)} = u^{n+1(2)} - \theta\nu_{33} r\delta^2_{x_3} u^n$$
$$(1-\theta\nu_{11} r\delta^2_{x_1}) u^{n+1(1)} = Au^n + \lambda B(u^{n+1}_{(1)} - u^n)$$
$$(1-\theta\nu_{22} r\delta^2_{x_2}) u^{n+1(2)} = u^{n+1(1)} - \theta\nu_{22} r\delta^2_{x_2} u^n$$
$$(1-\theta\nu_{22} r\delta^2_{x_3}) u^{n+1} = u^{n+1(2)} - \theta\nu_{33} r\delta^2_{x_3} u^n$$

END DO

The discrete operators A and B in the above algorithms are defined as

$$A = 1 + r(1-\theta)\nu_{11}\delta_{x_1}^2 + r\sum_{i=2}^{3}\nu_{ii}\delta_{x_i}^2 + \frac{1}{2}r\sum_{i=2}^{3}\sum_{j=1}^{i-1}\nu_{ij}\delta_{x_i x_j} + \frac{1}{2}r\Delta x\sum_{i=1}^{3}c_i\delta_{x_i} + \gamma\Delta t,$$

and

$$B = \frac{1}{2}r\sum_{i=2}^{3}\sum_{j=1}^{i-1}\nu_{ij}\delta_{x_i x_j} + \frac{1}{2}r\Delta x\sum_{i=1}^{3}c_i\delta_{x_i} + \gamma\Delta t.$$

The choice of coefficients $\lambda = \theta = 0.5$ yields a second-order convergence in both time and space variables, and the stability depending on the coefficients is analyzed in [1].

3. CUDA AND PARALLEL IMPLEMENTATION

GPU (Graphic Processor Unit) has evolved into a highly parallel, multi-threaded, many-core processor with tremendous computational power and high memory bandwidth. Recent common GPUs consist of over 30 multiprocessors, a set of several arithmetic cores. For example, in NVIDIA's GTX 285, there are 30 multiprocessors and each multiprocessor has 8 single-precision scalar processor and 1 double-precision scalar processor.

CUDA, released by NVIDIA, is a general purpose parallel computing architecture that enables developers to use GPU for solving many complex computational problems by providing a C-like programming language. In what follows, we describe the parallel algorithm for ADI method and its implementation on the CUDA platform.

3.1 Parallel Implementation

To implement the parallel algorithm for the alternating direction implicit method on GPU, we focus on solving the tridiagonal matrix which is the discretization matrix of elliptic operator in each direction. In general, the tridiagonal matrix solver is practically a sequential algorithm, thus it is hard to parallelize.

Consider the tridiagonal matrix form produced by ADI method shown in Figure 1 with the grid size N for each direction. Since the dimension of the tridiagonal matrix is N^2, there are N^2 number of independent tridiagonal matrices of size $N \times N$.

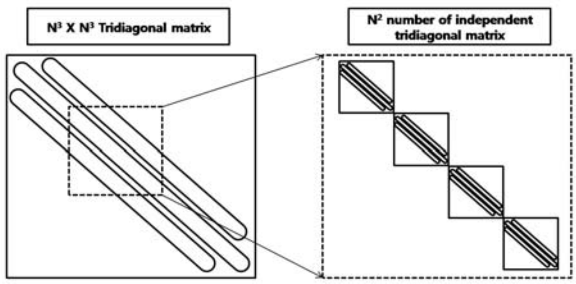


Figure 1. The composition of the matrix

Instead of parallelizing the tridiagonal solver, we choose the parallelization in which each thread is assigned to solve certain number of the tridiagonal systems like Figure 2.

The number of independent tridiagonal matrices, which is responsible for each thread, is determined by the available thread number of GPU device. Suppose that the number of threads per block is T and the block size is $(m \times n)$. Then the total number of threads will be

$$T_{total} = T \times m \times n\ .$$

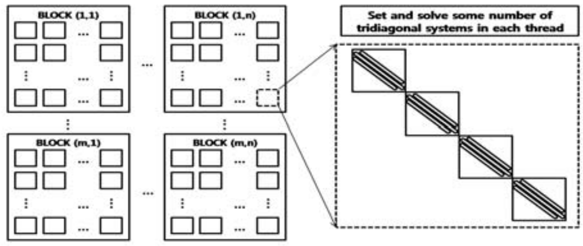


Figure 2. Tridiagonal Systems Assigned to Each Thread

Then the integer part of N^2 / T_{total} determines the number of the independent tridiagonal matrices and the modulus $M = N^2 (\bmod T_{total})$ is added to the M number of threads so that every tridiagonal matrix is assigned to a certain thread. During the computation, all the independent tridiagonal matrices will be solved in parallel. For the three-dimensional case, the proposed algorithm will be applied three times in x, y, and z direction.

4. NUMERICAL RESULT

We used the proposed parallel method on the graphic card, NVIDIA GTX285 with Intel Dual Core E2150 1.6GHz CPU. In the followings, the reduction rate and the speedup are defined by

$$\text{reduction rate} = \log_2 \frac{\| u_{\Delta \mathbf{x},\Delta t} - u_{\text{exact}} \|_{L^2(\Omega)}}{\| u_{\frac{\Delta \mathbf{x}}{2},\frac{\Delta t}{2}} - u_{\text{exact}} \|_{L^2(\Omega)}},$$

and

$$\text{speedup} = \frac{\text{time consumption on CPU}}{\text{time consumption on GPU}},$$

where $u_{\Delta x,\Delta t}$ denotes the numerical solution with the spatial mesh size $\Delta x_i = \Delta x$ for $i = 1,2,3$ and time mesh size Δt.

Example We consider the parabolic problem of the form Eq. (1) with the coefficients $\gamma = 0,\ c_1 = -0.4,\ c_2 = -0.7,\ c_3 = -0.5$ and

$$(\nu_{ij}) = \begin{pmatrix} 0.5 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}$$

on the domain $(0,2)^3 \times (0,1]$. The initial data and Dirichlet bundary data are given such that the exact solution is given by

$$u(x_1,x_2,x_3,t) = \frac{\exp\left(-\frac{(x_1 - c_1 - 0.5)^2}{\nu_{11}(1+4t)} - \frac{(x_2 - c_2 - 0.5)^2}{\nu_{22}(1+4t)} - \frac{(x_3 - c_3 - 0.5)^2}{\nu_{33}(1+4t)} \right)}{(1+4t)^{3/2}}.$$

The initial data is a three dimensional Gaussian pulse, and it is an extension of the example in [5], while the original problem is considered in two dimensional space.

The following tables confirm convergence of the proposed method, depending on the choice of λ as describes in [1]. If $\lambda = 0.5$, even though it doubles the total number of iteration compared with the case $\lambda = 0.0$ as described in Section 2, since it gives second order convergence in time, the strictly small number of time step is required for the same level of error required in the case $\lambda = 0.0$.

Table 1. Errors and Time Consumption for Example 1 with $\theta = 0.5$ and $\lambda = 0.0$

nt,nx	Device	Time(sec)	L2-Error	Reduction	Speedup
20,10	CPU	0.03	2.81E-03		
	GPU	0.18	2.81E-03		0.17x
40,20	CPU	0.24	1.17E-03	1.26	
	GPU	0.25	1.17E-03	1.26	0.96x
80,40	CPU	3.32	5.73E-04	1.03	
	GPU	0.56	5.74E-04	1.03	5.93x
160,80	CPU	49.99	2.88E-04	0.99	
	GPU	3.81	2.91E-04	0.98	13.12x

Table 2. Errors and Time Consumption for Example 1 with $\theta = 0.5$ and $\lambda = 0.5$

nt,nx	Device	Time(sec)	L2-Error	Reduction	Speedup
20,10	CPU	0.05	2.03E-03		
	GPU	0.20	2.03E-03		0.25x
40,20	CPU	0.51	5.02E-04	2.02	
	GPU	0.35	5.02E-04	2.02	1.46x
80,40	CPU	7.18	1.25E-04	2.01	
	GPU	0.95	1.25E-04	2.00	7.56x
160,80	CPU	118.24	3.10E-05	2.01	
	GPU	8.49	3.20E-05	1.97	13.97x

A comparison of the time consumptions between GPU and CPU is also available. CUDA supports the debug mode which executes the same code on the CPU. However, since this prevents the optimized execution on the CPU, we rewrite the code only for CPU and execute this to be a fare comparison. Even though at the largest space mesh size, the speedup is less than 1, the finer space mesh size provides, the better speedup is observed. It is because at the largest space mesh grid, the main bottleneck comes from the memory copy from the host machine to the GPU device.

In spite of the importance of the double-precision arithmetic in scientific computation, the newest version of CUDA started to support the double-precision. We also compare the speedup in the double-precision. The lower speedup of the double-precision is observed due to the larger requirement of the memory size and the smaller number of double-precision scalar processor than the single-precision computation as described in Section 3.

Table 3. Errors and Time Consumption for Example 1 with $\theta = 0.5$ and $\lambda = 0.5$ Based on Double-precision Computation

nt,nx	Device	Time(sec)	L2-Error	Reduction	Speedup
20,10	CPU	0.06	2.03E-03		
	GPU	0.42	2.03E-03		0.14x
40,20	CPU	0.52	5.02E-04	2.02	
	GPU	0.96	5.02E-04	2.02	0.54x
80,40	CPU	7.77	1.25E-04	2.01	
	GPU	3.07	1.25E-04	2.01	2.53x
160,80	CPU	126.49	3.12E-05	2.00	
	GPU	15.61	3.12E-05	2.00	8.10x
320,160	CPU	2141.38	7.78E-06	2.00	
	GPU	258.19	7.78E-06	2.00	8.29x

5. CONCLUSIONS

Parallelized alternating direction implicit (ADI) method for general parabolic equations on GPU architecture has been implemented and analyzed. Its computational efficiency has been confirmed showing the speedup 13 times in single precision and 8 times in double precision. Even though GPU has limitations in maximum memory size and communication speed between CPU and itself, for the computations which need tremendous arithmetic with a shared memory it could be a good choice replacing the traditional CPUs. However, since the CUDA or other programming architectures supporting GPU cannot optimize a user program code fully yet, the optimization issues related to a memory allocation and handling on GPU should be considered by programmers themselves.

REFERENCES

[1] I. J. Craig and A. D. Sneyd. An alternating-direction implicit schemefor parabolic equations with mixed derivatives. Comput. Math. Applic., 16(4):341–350, 1988.

[2] J. Douglas. Alternating direction methods for three space variables. Num. Math., 4:41–63, 1962.

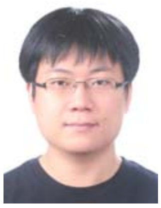
[3] J. Douglas and J. E. Gunn. A general formulation of alternation directionmethods. Num. Math., 6:428–453, 1964.

[4] J. Douglas and H. H. Rachford. On the numerical solutionof heat conductionproblems in two and three space variables. Trans. Am. Math. Soc., 82:421–439, 1956.

[5] S. Karaa and J. Zhang. High order adi method for solving unsteadyconvection–diffusion problems. J. Comp. Phys., 198:1–9. 2004.

[6] NVIDIA. Cuda 2.2 programming guide. CUDA Zone Website, 2009. http://www.nvidia.com/object/cuda_home.html.

Hyoungseok Chu is a graduate student pursuing Master's degree in Interdisciplinary Program in Computational Science & Technology, Seoul National University. He graduated from Ajou University in 2009 majoring in Computer Science. His research interests are in numerical analysis and parallel computing especially in the many-core system.



Hyoseop Lee is a doctoral student in Interdisciplinary Program in Computational Science & Technology, Seoul National University. He graduated with a Bachelor degree in Mathematical Science in 2005; a Master degree in Program in Computational Science & Technology in 2007 from Seoul National University with specialty in scientific computation. He was the first prize winner of 5th EASIAM student paper competition. His research interests are in parallel algorithm, scientific computation and financial engineering.

Dongwoo Sheen is a Full Professor in Mathematics Department and Chair of Interdisciplinary Program in Computational Science & Technology, Seoul National University. He graduated from Seoul National University in 1981; from Purdue University in 1991 with specialty of numerical analysis and scientific computation. He was a postdoctoral fellow at the Istituto di Analisi Numerica (1991~1992), and Purdue University (1992~1993), an ARC International Fellow at the University of New South Wales(1999) and Visiting Professor at Purdue University (2006~2007). Currently he is the President of East Asia Section of SIAM (EASIAM). His research interests are in numerical analysis, frequency domain algorithm, mathematical biology modelling and parallel computing.

UTILIZING CUDA FOR PRECONDITIONED GMRES SOLVERS

Shiming Xu [1], Hai Xiang Lin [1], Wei Xue [2], Ke Wang [2]
[1] Delft Institute of Applied Mathematics, TU Delft
Mekelweg 4, Delft 2628 CD, The Netherlands
[2] Dept. of Comp. Sci. & Tech., Tsinghua University
Qinghuayuan 1, Beijing 100084, P.R. China
Email: auhgnist@gmail.com, haix_lin@yahoo.com, {xuewei,wangke}@tsinghua.edu.cn

ABSTRACT

Krylov subspace based iterative methods is one of the most fundamental solving technique for large sparse linear systems. In this paper we discuss the parallelization of Krylov subspace methods by CUDA, using GMRES as an example. We compare various algorithm options for orthogonalization and preconditioner choices. Experiments show that the speed-up for the computation kernels associated with GMRES is higher than 4, which is comparable to the speed-up achieved by dense linear algebra methods. The best speed-up for the GMRES iteration is over 5, while the averaged speed-up is about 3.2.

Keywords: Krylov Subspace Methods, GMRES, CUDA, AINV Preconditioner

1. INTRODUCTION

Krylov subspace based iterative methods are one of the fundamental solving techniques for large linear systems [2]. Especially, preconditioning techniques play an important role in the overall behavior of the iterative solvers. Due to the ever growing demand for solving larger scientific problems with finer precision, there has been much work on parallel algorithms and implementations for Krylov subspace methods.

In recent years high performance computing has undergone a shift in paradigm. Throughput-oriented systems are emerging as accelerators to traditional computing platforms. Most notably is the adopting of programmable graphics processors (GPU) for applications such as scientific computing [12]. Instead of only a few, out-of-order processor cores, GPU integrates many simple cores to conduct computation in a data-parallel way. In the most recent NVidia GPUs, 240 cores are divided into 30 clusters, and cores in each cluster execute in an SIMT manner. Also the memory subsystem for GPU is stream-optimized, with much smaller cache and much higher theoretical throughput than CPU.

Programming GPU explicitly is mandatory to exploit the computational power of GPU, mainly due to the lack of compiler for legacy code. Currently CUDA [6] is the most widely used platform for programming GPU. In CUDA, user should specify how parallel threads are organized and each thread can have its own identity, and execute code/manipulate data accordingly, similar to parallel MPI processes, but with a shared memory space. Figure 1 shows a sample configuration of threads organized into blocks and grids, in a hierarchical way.

There has been some recent works on using CUDA to the solving/factorization of dense systems [7], showing an over 4x speed-up in performance. Also in [10] the factorization of sparse matrix is exploited by super-node approach. All these work involve using dense matrix operation to enhance performance which is not generally applicable to sparse matrices. In this article we explore the potential of adopting CUDA and GPU for Krylov subspace methods based on sparse structure; we also implement preconditioner which are suited for utilizing parallelism provided by GPU. Experiments show that GMRES Computing kernels show nearly 4x speed-up, while preconditioner construction show 2x speed-up for matrices from real-world applications

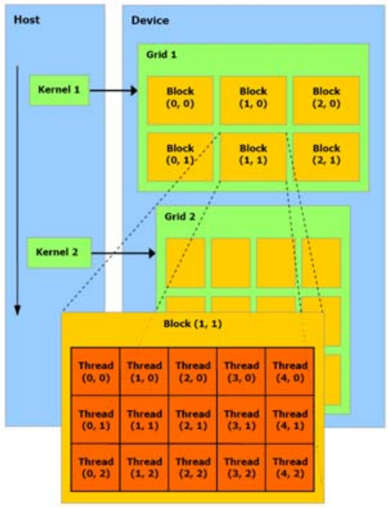


Figure 1. A Sample CUDA Thread Organization.

In the next section we introduce the basic knowledge of Krylov subspace solvers and preconditioning techniques. In Section 3 we discuss the parallelization schemes for different computation kernels in Krylov subspace methods. We carry out experiments of these kernels in Section 4, on matrices from real world applications and give detailed analysis to the performance behavior and tuning possibilities. Section 5 concludes the article.

2. BASICS OF ITERATIVE SOLVERS AND PRECONDITIONERS

2.1 Krylov Space Based Iterative Solvers

Here we discuss the problem of solving a linear system:

$$Ax = b$$

Here A is a square, non-singular sparse matrix of size $n \times n$. b and x are non-zero vectors of length n. b is given and x is the solution to the equation. In practice n can be very large and it incurs prohibitive computational and memory cost to solve the system directly, by factorizing A into triangular matrices as L and U [1]. In practice, one usually resorts to solving the system iteratively and seek a solution which is precise enough for practical usage.

The most popular iterative solvers are those based on Krylov subspace, including CG, GMRES, BiCGStab, etc. A Krylov space is a subspace in R^n defined in the form of $\{r, Ar, A^2r, ..., A^{m-1}r\}$, where r is the initial residue vector. Krylov subspace methods mainly involve two kinds of operations: (1) sparse matrix-dense vector products (SpMV), and (2) vector orthogonalization. Although different solvers has different definitions over the Krylov space and there are specific solvers for special types of matrices, such as Symmetric Positive-Definite matrices, these solvers all feature the three kinds of operations above as the fundamental operations. As an example, the algorithm of GMRES [2] is listed below.

```
   Input: Sparse matrix A
          Right hand side b
          Initial guess x_0
   Output: Solution x
1  r_0 ← b − A x_0
2  v_1 ← normalized r_0
3  for i ← 1 to m do
4  |  w_i ← A v_i
5  |  Orthogonalize w_i against v_j for 1 ≤ j ≤ i
6  |  if ||w_i||_2 < ε then
7  |  |  break
8  |  end
9  |  Update Hessenberg matrix H̄
10 |  v_{i+1} ← normalized w_i
11 end
12 y_m ← arg min_y ||βe_1 − H̄y||_2
13 x ← x_0 + Σ_{i=1}^{m} y_i v_i
14 return x
```

The main computation time are consumed by the following part: (1) SpMV on line 4; (2) BLAS [3] based operations on line 5 (orthogonalization), line 2 and line 13, mainly including AXPY, DOT and NORM operations. Since n is very large, usually we set m to be $m << n$ to avoid too much memory overhead. If convergence test (line 3) never succeeds and i reaches m, the whole GMRES process can be restarted with the currently retrieved solution x as x_0. Since in practice m does not exceeds several hundred, and the computation on line 12 can be implemented easily with Givens rotation and it does not incur much computation.

In this article we choose GMRES as the example for implementation using CUDA, because it is a most popular iterative solver which is widely used, and its application is not restricted to symmetric matrices.

2.2 Preconditioning Techniques

If A is ill-conditioned, it may takes too much iterations to reach convergence. In this case, preconditioners are introduced to speed up the convergence process [2]. Generally, preconditioners fall into two categories [9]: (1) incomplete factorization-based ones, and (2) approximate inverse-based ones.

The factorization like Incomplete LU (ILU), involves factorizing A inexactly into L and U, satisfying:

$$A \approx L \times U$$

L and U are lower-triangular and upper-triangular matrices, respectively, which should be easy to compute and sparse by dropping values or constraining sparsity patterns. With L and U, we are solving an equivalent system with form of $U^{-1}L^{-1}Ax = U^{-1}L^{-1}b$. The matrix of $U^{-1}L^{-1}A$ needs not be formed explicitly; instead at each SpMV operation, forward elimination (F-E) and backward substitution (B-S) are performed after SpMV is performed. Since F-E and B-S are inherently sequential, they cannot be implemented in parallel in an easy way.

Approximate inverse-based preconditioners (AINV) try to find a good approximation to the inverse of A:

$$M \approx A^{-1}$$

Like L and U in ILU, M should be easy to compute and sparse but retain effective information of A^{-1}. With M we solve an equivalent system of form $MA^{-1}x = Mb$. Instead of the inherently sequential application of L and U, each preconditioning operation with M only involves SpMV operations, which can be easily parallelized. Hence the process of applying the approximate preconditioner can be more easily implemented using CUDA.

Here we take AINV [8] preconditioner based on *A*-biconjugate algorithm as an example (D is a diagonal matrix):

```
   Input: Matrix A
   Output: Matrix W, Z and D
1  w_i ← e_i for 1 ≤ i ≤ n
2  z_i ← e_i for 1 ≤ i ≤ n
3  for i ← 1 to n do
4  |  for j ← i to n do
5  |  |  p_j ← ⟨A(i,:), z_j⟩
6  |  |  q_j ← ⟨A(:,i), w_j⟩
7  |  end
8  |  for j ← i+1 to n do
9  |  |  z_j ← z_j − (p_j / p_i) z_i
10 |  |  w_j ← w_j − (q_j / q_i) w_i
11 |  end
12 |  W ← [w_1, ..., w_n]
13 |  Z ← [z_1, ..., z_n]
14 |  D_{i,i} ← p_i for 1 ≤ i ≤ n
15 end
```

After the computation we have $M = ZD^{-1}W^T = A^{-1}$. Z and W are upper triangular matrices. If these two matrices are generated by keeping only non-zero elements on certain positions, an approximate to A^{-1} is computed which can be used for preconditioning.

Note that similar to ILU, the construction of AINV preconditioner feature a row-wise parallelism in construction: line 4 to 7 and line 8 to 11 are both parallelizable on the iteration level.

Preconditioner construction incurs a constant overhead to the whole solving process; the application of preconditioner incurs a constant overhead for each GMRES iteration, so the overall overhead for preconditioner application is linearly proportional to the iteration count.

3. ANALYSIS OF PARALLELIZATION USING CUDA

In this section we give a thorough analysis over the potential of utilizing GPU to speed up the process of GMRES algorithm.

We mainly focus on three aspects: SpMV, orthogonalization, and preconditioner-related operations.

3.1 Sparse Matrix-Vector Multiplication

Sparse matrix and dense vector multiplication (SpMV) is an essential operation in iterative methods, due to that the generation of the bases of the Krylov subspace depends on SpMV operations. In each SpMV operation, the nonzero elements of A will be traversed exactly once. Given a matrix A with $nnz(A)$ non-zero elements, the total count of floating-point operations is $2 \times nnz(A)$, and the total amount of data transferred is $nnz(A) + 2n$. Given the low ratio of operation count/data amount, the SpMV operation is mainly bandwidth-bound. On optimizing SpMV for GPU, utilizing the limited bandwidth is the key to high performance. Several recent works [4, 5] focus on optimizing the SpMV kernel for GPU. The memory layout of A plays an important role to bandwidth utilization, and ELLPACK format shows best overall performance due to its effect of memory access coalescing among consecutive threads, by packing the actual non-zero elements in a `column' format, as implied in [4].

In this paper we will use ELLPACK and HYB formats and SpMV kernel as the basis for the generation of Krylov subspace bases.

3.2 Orthogonalization

In iterative methods, the bases of the Krylov subspace are usually orthogonalized to guarantee certain properties. For long-recurrence methods such as GMRES, orthogonalization is carried out between the newly generated basis and each of the orthogonal bases of the subspace which are already generated (shown in the algorithm above); in short-recurrence methods such as BiCG, such orthogonalization is only carried out between current basis and the most recent basis. Here we take GMRES as an example to show the basic rules and the potential of utilizing GPU in this process.

There are several algorithms for orthogonalizing a vector against a set of orthogonal vectors. Among them, algorithms based on Gram-Schmidt method and those based on Householder transformation are most common. The modified G-S algorithm (M-G-S) is shown below.

Input: Orthogonal vector set V
Vector to be orthogonalized u
Output: Orthogonalized vector u
1 **for** $i \leftarrow 1$ **to** $|V|$ **do**
2 $\quad \beta \leftarrow \langle v_i, u \rangle$
3 $\quad u \leftarrow u - \beta v_i$
4 **end**
5 **return** u

Since there is carried dependence in the loop from line 1 to 4, many works resort to the less stable version of M-G-S, namely Classical G-S and use re-orthogonalization (G-S-Re) for more parallelism on the outer loop level [13]. But re-orthogonalization consumes similar operation counts as Househoulder version. While incurring over 100% more computation, G-S-Re may introduce limited parallelism for CUDA to exploit, especially when the length of the vector/basis is large. So for orthogonalization we implement MKL-based MGS algorithm and CUBLAS based MGS algorithm for comparison. Both of them are automatically parallelized. While MKL only uses parallelized implementation when the vector is long enough to avoid overhead, CUBLAS is always parallelized and have additional calling overhead.

3.3 Preconditioning Techniques

In Section 2 we surveyed preconditioner options for GMRES algorithm. AINV preconditioner has the same root as the widely used ILU counterparts but features good parallelizability in both construction and application. So for parallelizing preconditioner construction we mainly focus on AINV.

Since it is hard to allocate various lengths of memory dynamically, especially when concurrent threads are running, generating all non-zero values and dropping small ones accordingly is not feasible for GPU implementation. Here we adopt a static non-zero pattern for W and Z matrix, and only non-zero fill-in's that occur inside the prescribed pattern are included. In this way memory and certain data structures can be initialized in advance. As a most simple form for sparsity pattern, we use the sparsity pattern of the original matrix A: W^T will has the same non-zero structure as the lower part of A, while Z will have the structure of the upper part of A.

The corresponding lines (line 9 and 10) in the AINV algorithm should be modified to accommodate fill-in's which only occur at the prescribed positions in w_j and z_j.

The parallelization of AINV construction follows these rules:

(1) For each i, the inner product on line 5 and 6, and the sparse axpy operation on line 9 and 10 are carried out in parallel. Each iteration (indexed by j) is assigned to a CUDA thread.

(2) The outer loop indexed by i contains carried dependency, hence it's performed sequentially.

After generation, Z and W^T are organized into HYB format for the application of preconditioners, which is performed by SpMV operations.

4. EXPERIMENTS

In this section we carry out experiments of the computing kernels of GMRES algorithm on GPU. We mainly compare the parallel algorithm on CPU and GPU. For the sake of impartial comparison, the CPU version of the algorithm is also implemented using processor-specific optimization and state-of-the-art libraries. All computations are in double precision floating point.

4.1 Experiment Settings

In experiments we mainly test the speed-up's achieved by parallelized version of the computing kernels using CUDA, compared with the CPU-based version.

The platform we are using is listed below. Note that the GPU we use support double floating point (DP) which complies with IEEE.754 standards. Although the peak floating point for DP operations is much lower than that of SP, this does not appear to be a constraint, since most kernels involved are bandwidth bound.

CPU	Intel i7-920 (4-core, 2.66GHz)
Memory	12GB (DDR-III, 1066MHz)
GPU	NVidia Tesla C1060
GPU Memory	4GB
CUDA Version	2.0

GMRES algorithm mainly involves three parts of computation: SpMV, vector orthogonalization, and preconditioner construction and application. For SpMV we compare the matrix multiplication on GPU using HYB format and that on CPU using MKL and CSR format. Since there is built-in automatic parallelization in MKL library for compressed sparse row (CSR) format based SpMV, this serves as a good reference since it represents an ideal performance baseline for SpMV on the highly-tuned library on a state-of-the-art processor. We compare the achieved FLOPS for GPU and CPU.

For orthogonalization methods, we compare: (1) M-G-S method implemented in MKL using standard BLAS operations, (2) M-G-S method implemented in CUBLAS. We mainly compare the FLOPS associated with orthogonalizing a new vector against a set of orthogonal bases. Due to the difference between the cache structure of GPU and that of CPU, trade-offs are possible for the configuration of vector length.

4.2 SpMV Performance Comparison

We tested the performance on a set of matrices, including 5 included in [4] and 5 from our scientific applications. These matrices are summed in Table 1.

Table 1. Sparse matrices for SpMV test.

	Protein	Cant	WindTunnel	Epidem	Circuit
Size	36K	62K	218K	526K	171K
NNZ	4.3M	4.0M	11.6M	2.1M	959K

	Petro	OPF	TDS	Cubes	Parabolic
Size	132K	2.1M	25K	101K	526K
NNZ	4.2M	8.1M	160K	874K	2.1M

We all use HYB format for the benchmarking of SpMV. Figure 2 shows the speedup of GPU over CPU.

The last bar on the right shows the geometric mean of the speed-up's of the matrix set. Matrices with the following characteristics show good speed-up, including Cant, Parabolic, WindTunnel and Epidem: (1) even distribution of nonzero elements in rows, (2) good locality in accessing X vector (i.e., the bandwidth of the matrix is small), and (3) sufficiently large amount of non-zero elements (NNZ) to amortize calling overhead. The 2 matrices with the lowest speedup are from Power System applications, which feature randomly sparse structures inherited from power system topologies.

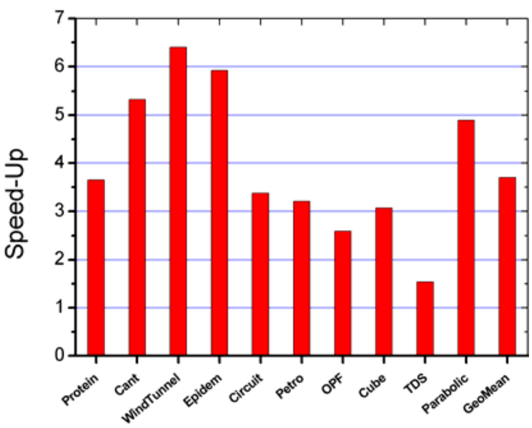


Figure 2. Speedup of SpMV on GPU over CPU.

4.3 Orthogonalization Performance Comparison

Figure 3 shows the performance for M-G-S on CPU and GPU. This graph corresponds with the orthogonalization of 1 vector against 64 orthogonal bases. We first generate 64 vectors and orthogonalize them; after that we benchmark the orthogonalization of a new vector against these bases. This implies that before benchmarking, some data of the bases may reside on CPU cache system. We see a clear benefit of M-G-S on CPU compared with M-G-S on GPU when the length of the vector is small, i.e., under 50'000.

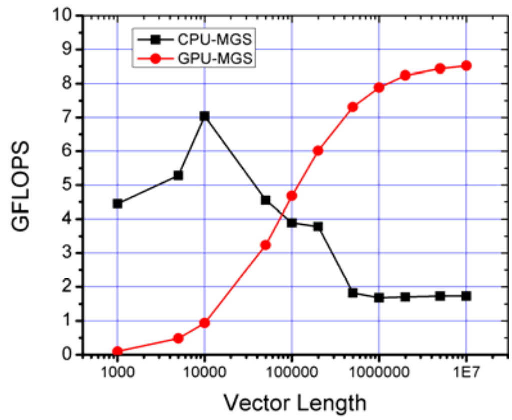


Figure 3. Performance of Orthogonalization on CPU and GPU.

The change in the relative performance is caused by the drastic difference between CPU and GPU. The low ratio of computation amount and data amount for orthogonalization has several implications:

(1) When the vector is very large in size (which cannot be effectively contained in CPU cache), the comparison between CPU and GPU mainly fall into the comparison between the available bandwidth of the two. Since given good access patterns, GPU has much higher bandwidth potential compared with CPU (140GB/s v.s. 25.6GB/s in our case).

(2) If the vector is small in length and there is good temporal locality as featured by our testing methodology, the whole set of vectors can be effectively contained by the CPU cache. Hence the effective bandwidth of CPU should be evaluated as that between the cache system and the computing core of the CPU, which is over 100GB/s. In this case CPU will have higher overall performance.

When used for Krylov solvers, the performance of CPU will be compromised compared with the results in Figure.3. Since orthogonalization happens when a new vector is generated by SpMV, so the CPU cache will be contaminated by the sparse matrix (which most probably will be much larger than the available cache size of the CPU). An orthogonalization with such a *cold*-start will have lower performance since all the bases must be fetched from memory.

4.4 AINV-0 Preconditioner Construction

Figure 4 shows the relative performance for the construction of AINV preconditioner of GPU and CPU. The geometric mean of the speed-up's is over a factor of 2.

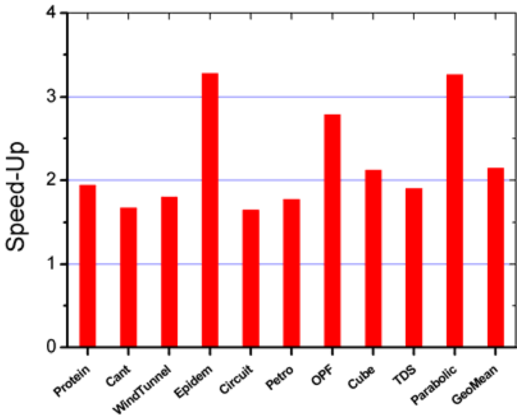


Figure 4. Speed-up for AINV Preconditioner Construction

4.5 Overall GMRES Performance

Figure 5 shows the performance ratio between GPU and CPU for a single GMRES iteration. GMRES is restarted after 64 iterations. This ratio is decided by several factors. For small matrices with a low non-zero element count such as TDS, GPU hardly outperforms CPU due to the effective utilization of CPU cache (no cold start happens). For very large matrix such as OPF, this ratio also tends to be more decided by orthogonalization performance ratio. For other matrices, it is a mixed case between the speed-up's achieved by various kernels in GMRES.

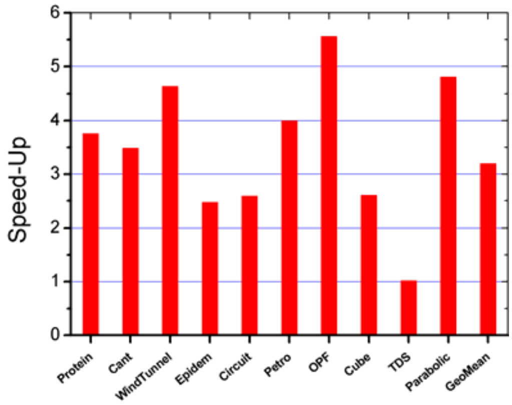


Figure 5. Performance Ratio between GPU and CPU

5. CONCLUSIONS

In this article we explored the potential of utilizing GPU hardware to implement and optimize Krylov subspace based iterative linear solvers. Specifically we use GMRES as an example to demonstrate the specific technique in applying CUDA to the computation kernels in GMRES, including orthogonalization algorithm preference, preconditioner preference, and parallelization.

The computation kernels associated with GMRES and other Krylov subspace methods are SpMV, orthogonalization, and preconditioner construction feature different parallelizability on GPU. The preference between CPU and GPU version of the orthogonalization kernel may be subjected to characteristics of the workload. Overall speaking, CUDA achieved good speed-up for GMRES performance: 3.2 times speed-up in overall iteration performance and 2 times speedup in AINV preconditioner construction.

Thanks: This work was supported in part by EU Asia Link programme under grant CN/ASIA-LINK/020, Microsoft Corp., the National Natural Science Foundation of China (Grant No.60703055), the Science and Technology Supporting Program of China (Grant No. 2006BAA02A17), and Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-Discipline Foundation.

REFERENCES

[1] Timothy A. Davis, Direct Methods for Sparse Linear Systems, SIAM, 2006.

[2] Yousef Saad, Iterative Methods for Sparse Linear Systems, 2nd Ed., SIAM, 2003.

[3] BLAS – Basic Linear Algebra Subprograms, http://www.netlib.org/blas/.

[4] Nathan Bell and Michael Garland, Implementing Sparse Matrix-Vector Multiplication on Throughput-Oriented Processors, SC'09.

[5] Muthu Manikandan Baskaran and Rajesh Bordawekar, Optimizing Sparse Matrix-Vector Multiplication on GPUs, Technical Report 2009.

[6] CUDA Zone, http://www.nvidia.com/cuda.

[7] Vasily Volkov and James W. Demmel, Benchmarking GPUs to Tune Dense Linear Algebra, SC'08.

[8] M. Benzi and M. Tuma, A Sparse Approximate Inverse Preconditioner for Nonsymmetric Linear Systems, SIAM J. Sci. Comput. 19 (1998).

[9] Michele Benzi, Preconditioning Techniques for Large Linear Systems: A Survey, Journal of Computational Physics 182 (2002) pp.418-477.

[10] Matthias Christen, Olaf Schenk and Helmar Burkhart, General-Purpose Sparse Matrix Building Blocks using the NVIDIA CUDA Technology Platform, Workshop on GPGPU, 2007.

[11] L. Giarud, J. Langou and M. Rozloznik, The Loss of Orthogonality in the Gram-Schmidt Orthogonalization Process, Intl. J. Computers & Math. with Applications, 50 (2005), pp.1069-1075.

[12] GPGPU, http://www.gpgpu.org.

[13] W. Hoffmann, Iterative Algorithms for Gram-Schmidt Orthogonalization, Computing 41 (1989), 335-348.

XU Shiming is a Ph.D. student at Delft Institute of Applied Mathematics of TU Delft. He graduated from Tsinghua University in 2003 with Bachelor's Degree; from Dept. of Computer Sci. & Tech of Tsinghua University in 2006 with Master's Degree in computer science. He entered TUDelft in 2006 to pursuit Ph.D. degree in Mathematical Physics group of Delft Institute of Applied Mathematics. His research interest includes power system simulation, iterative methods and preconditioners of linear systems, GPGPU, etc.

MULTIRESOLUTION FAST METHODS FOR A PERIODIC 3-D NAVIER-STOKES SOLVER

Jun Jia, Judith Hill, George Fann and Robert J. Harrison
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA
Email: jiaj@ornl.gov

ABSTRACT

We describe the construction of an approximation of the Green's function for the Poisson kernel and the Helmholtz kernel and its application to solving the Navier-Stokes equations in a 3-D periodic domain using a multiresolution analysis approach. A low-separation approximation of the Poisson kernel and the Helmholtz kernel are constructed based on Beylkin-Mohlenkamp's low-separation rank approach and a fast lattice sum. The Poisson kernel is applied to solve the Poisson equation for the pressure, and the Helmholtz operator is applied to solve the backward Euler implicit time-stepping problem.

Keywords: Multiwavelets, Multiresolution Analysis, Low-separation rank approximation, High order discretization, Navier Stokes equation.

1. INTRODUCTION

We describe the low-separation rank approximation [3] of the free space Poisson kernel 1/r and the bound state Helmholtz kernel e^{-kr}/r and their application to solving the Navier-Stokes equations in a periodic domain. The approximation of the free space Poisson kernel and the Helmholz kernel is to arbitrary-order and to arbitrary, but finite, accuracy. It is used to solve two of the most non-scalable parts of many Navier-Stokes solvers on a massively parallel computer: the Poisson equation for pressure and the Krylov iterative solver for time-stepping.

To our knowledge, this is the first application of multiwavelets and the low-separation rank approximation to solve the Navier-Stokes equations in a 3-D periodic domain using an adaptive pseudo-spectral approach. The method that we are developing is a fast real-analysis $O(N \log ò)$ approach where N is the number of unknowns and ε is the precision, similar in spirit to the Fast Multipole Method (FMM). FMM and its cousins, the kernel-free FMM method [8, 9, 14, 15], have solved the Poisson and Helmholtz equations to high accuracy and have been applied to solving the Poisson equation for the Navier-Stokes equations.

In our pseudo-spectral approach, all of the functions are approximated using a tensor-product of the 1-D Alpert multiwavelets [1]. The approximation and the construction of the Green's function and its fast application for the solution of the Poisson equation and for the bound state Helmholtz equation are performed using Beylkin-Mohlenkamp's low-separation rank approximation [3]. The compatibility of the wavelet approximation and that of the low-separation rank approximation for convolution operators is described in [4]. In the past, these approaches have been successively applied to solving for all of the wave-functions of the 3-D Schrodinger's equation and the Kohn-Sham equation, without assumption on symmetry of the potential, to between twelve and fourteen digits of accuracy in computational chemistry applications [10].

Our initial prototype codes for solving the Navier-Stokes equation and for problems in computational chemistry are realized in our software package Multiresolution ADaptive Numerical Environment for Scientific Simulation (MADNESS) [4,10].

2. MULTIRESOLUTION REPRESENTATIONS AND LOW-SEPARATION RANK REPRESEN-TATIONS OF FUNCTIONS AND OPERATORS

Functions and variables in MADNESS are represented in a multiresolution manner based on Alpert's multiwavelets [1]. Alpert's multiwavelets are formed using scaling functions based on normalized Legendre's polynomials rescaled to [0,1]. The multiwavelets have many properties similar to wavelets, such as vanishing moments, orthonormality, and compact support. The basis functions have two-scale relations that produce different levels of refinement based on the support of the basis function. The basis functions do not overlap on a given scale and are organized into small groups of several functions which share the same support. Some of the basis functions are singular and discontinuous, similar to the Haar basis yet different from the wavelets with regularity. Because of the vanishing moment property, many operators and functions can be represented in a nearly sparse fashion. These properties enable the multiwavelet basis to be used in a computationally efficient adaptive fashion with a sparse representation of a large class of weakly singular operators and a fast algorithm with guaranteed precision for many common operations such as inversion, multiplication and addition [2,4].

However, operators and functions in the singular and the oscillatory regions are represented by low-separation rank approximations. The low-separated rank representation (LSR) [3] of a multivariable function f is defined by an approximation of the form,

$$\left\| f(x_1, x_2, \cdots, x_n) - \sum_{l-1}^{r} s_l \varphi_1^l(x_1)\varphi_2^l(x_2)\cdots\varphi_n^l(x_n) \right\| \le \varepsilon$$

with a user desired accuracy ε . The functions $\{\varphi_i^l(x_i)\}$ and coefficients $\{x_i\}$ are adjusted to achieve this accuracy with a near minimal separation rank r . The set of functions $\{\varphi_i^l(x_i)\}$ is not fixed and depends on the function f.

We approximate the application of the Poisson kernel and the Helmholtz kernel in free space using the LSR approximation. In the next section, we construct fast lattice sums for the 3-D periodic boundary conditions. Multiwavelets approximation of functions can be efficiently combined with an efficient application of the kernels using low separation approximations with Gaussian functions.

3. LATTICE SUM AND THE GREEN'S FUNCTION FOR THE POISSON EQUATION AND THE HELMHOLTZ EQUATION WITH PERIODIC BOUNDARY CONDITION

In this section we describe our construction and application of the periodic Green's function from the free space Green's function for the Poisson and the Helmholtz kernels. From the free space kernels $1/r$ and e^{-kr}/r, an integral convolution operator for a periodic domain is constructed using lattice sums.

We summarize our approach in this section. Assume that a free space kernel f is given and is valid in a segment $(0,p)$, a periodic function K with periodic boundary condition is constructed by an infinite sum

$$K(r) = \sum_{n=-\infty}^{\infty} f(r+np).$$

If the function f has a low-separation rank approximation in the interval $(0,p)$, $\| f(r) - \sum s_m e^{-a_m r^2} \| < \varepsilon$, we can substituted it into the approximation for K. Assuming that the sum exists we switch the finite and the infinite sums. We seek to approximate an infinite sum of Gaussians of the form, $\sum e^{-a_m (r+np)^2}$, for each m, by a finite sum.

For simplicity assume $p=1$. Fixing r, the sum $\sum e^{-a_m (r+n)^2}$ is less in absolute value than $\int e^{-a_m (r+n)^2} dn = \sqrt{\pi / a_m}$. Thus, the function $K(r)$ is defined.

In order to construct a fast and accurate summation we truncate the infinite sum to a finite sum and estimate the error for the application of the periodic convolution operator $K(r-s)$, and truncate an infinite to a finite sum with $r,s \in [0,1]^3$.

An estimate of the finite sum of Gaussians to estimate the infinite sum $\sum e^{-a_m (r+n)^2}$ can be obtained for each a_m in a number of different ways. For example, one can look up a probability density table estimating the tails of the probability density and modulus of precision. We use the computational approach, which is akin to Ewald sums. Since each term of the sum is positive and decreases as a function of n, we perform adaptive 1-D quadrature to compute the sum to the number of digits of precision that is required for the computation. If the precision, the error tolerance, and the number of multiwavelets being used are known ahead of time, the expansions can be precomputed. The number of coefficients for the application of the Green's function depends is $O(N \log \grave{o})$ where N is the number of term coefficients in the pseudo-spectral representation of a scale variable (i.e., the pressure p).

4. APPLICATION TO NAVIER-STOKES EQUATION

We now demonstrate the Poisson and Helmholtz solvers in the context of the Navier-Stokes equations. Consider the equations for the incompressible fluid case

$$u_t + \nabla p = -(u \cdot \nabla) u + \mathrm{Re}^{-1} \nabla^2 u + f$$
$$\nabla \cdot u = 0,$$

where **u** is the fluid velocity, p is the fluid pressure, and Re is the Reynolds number. We solve these coupled equations with a projection method that explicitly treats the nonlinear convective term. Taking the divergence of the momentum equation and applying divergence-free condition yields the Poisson problem for pressure

$$\Delta p = \nabla \cdot f - \nabla \cdot (u \cdot \nabla) u.$$

Based on the above computed pressure, the time derivative is discretized with a backward Euler method and the non-stiff advection and gradient terms are treated explicitly. This yields the Helmholtz problem

$$\left(\frac{\mathrm{Re}}{\delta t} - \Delta\right) u^{t+1} = \mathrm{Re}\left(f - (u^t \cdot \nabla) u^t - \nabla p + \frac{1}{\delta t} u^t\right).$$

The Poisson equation is solved using the Green's function for the Poisson operator. The Helmholtz operator is formally $(\mathrm{Re}/\delta t - \Delta)^{-1}$. The operator $(\mathrm{Re}/\delta t - \Delta)$ is usually discretized and inverted numerically, typically using a Krylov iterative solver or block cyclic reduction techniques. In our test, a conjugate gradient solver required 15 iterations to converge. After the Helmholtz operator has been constructed and stored as a part of initialization, each subsequent application require one matrix-vector multiply to obtain an updated velocity to the same precision.

In summary, the algorithm is

1. initialize the computations
   - the velocities,
   - the time step,
   - distance of the low-separation rank from 0 (singularity for the Green's functions)
   - the precision from the user specified tolerance,
2. compute the pressure using Poisson operator,
3. compute the updated velocity,
4. if the divergence is too large, refine and go back to 2; otherwise advance the time step,
5. goto 2.

Although this type of Chorin projection algorithm [5,6,13] is not new the use of Green's functions based on low-separation rank approximation and the use of multiwavelets as a basis for the pseudo-spectral method for representing functions and vector fields in 3-D is new. Wavelets have been applied to analysis of turbulence or as a basis for discretization of the differential form of fluid equations. See [7, 11, 12] as well as their citations for many recent examples. However, the application of wavelet methods for solving many problems effectively requires the construction of wavelets with special properties. Discretizing the differential form of the Navier-Stokes equation results in linear systems which require iterative linear solvers. The accuracy of these solvers depends on the condition number of the system. The difficulty of solving these equations to high precision is due to the high condition numbers of the resulting linear system with the unbounded Laplacian and the derivative operators. This is analogous to the situation with the application of adaptive finite difference and finite element techniques. Further details and analysis are in forthcoming papers.

5. NUMERICAL EXAMPLE

Our method of solution for the time-dependent Navier-Stokes equations is demonstrated for two test examples. In the first, the accuracy and stability of this method is evaluated. In the second, an interacting vortices problem is solved. All of the computations are performed on a 2.4 GHz AMD Opteron machine.

A forcing function f was applied to a periodic domain of size $[0,\pi]^3$ such that the exact solution is

$$
\begin{aligned}
u_{ex} &= \cos(t)\sin^2(x)(\sin(2y)\sin^2(z) - \sin^2(y)\sin(2z))\\
v_{ex} &= \cos(t)\sin^2(y)(\sin(2z)\sin^2(x) - \sin^2(z)\sin(2x))\\
w_{ex} &= \cos(t)\sin^2(z)(\sin(2x)\sin^2(y) - \sin^2(x)\sin(2y))\\
p_{ex} &= \cos(t)\cos(x)\sin(y)\cos(z).
\end{aligned}
$$

The streamlines of the velocity field, colored by the pressure, at time t = 0.04 are shown in Figure 1. In this example 7 multiwavelets were used, and the Reynolds number is 200. Figure 2 illustrates that the error in the velocity computation of this algorithm remains constant for each time step, as expected.

In the second example, we apply a time-dependent body force throughout a periodic domain of size $[0, 2\pi]^3$ yielding a flow field with two counter-rotating vortices, as illustrated in Figure 3 for Reynolds number 100. The forcing function applied is

$$
\begin{aligned}
f_x &= f_z = 0\\
f_y &= 18e^{-\frac{(x-x_0)^2+(y-y_0)^2}{4}}.
\end{aligned}
$$

6. CONCLUSIONS

We have demonstrated that real analysis type methods with approximating Green's functions, to arbitrary but finite precision, can be used to solve the Navier-Stokes equations by using a multiresolution, low-separation rank, adaptive pseudo-spectral method. Accuracy and precision are maintained with the fast *O(N)*-nature of these methods.

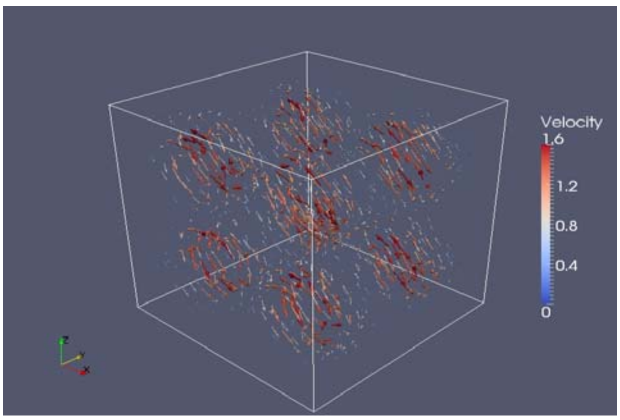


Figure 1. Illustration of the velocity streamlines for an analytical problem at time t=0.04. In this example, the domain Reynolds number is approximately 200. This computation was done with six multiwavelets with a precision of 10^{-4}.

Figure 2. The L2-norm of the error of the computed solution compared with the analytical solution in time using 10 multiwavelets.

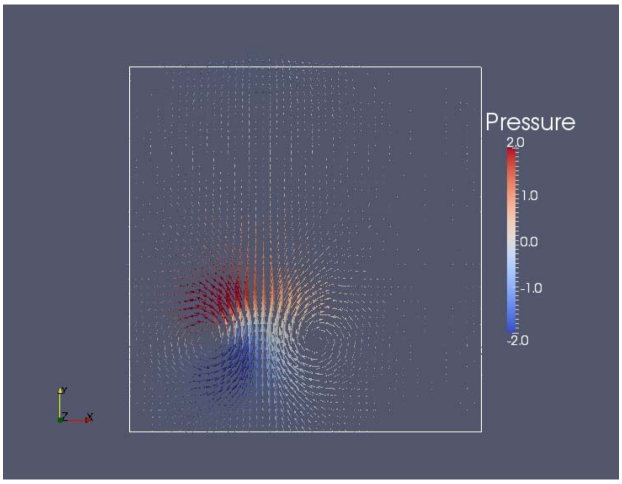


Figure 3. The velocity streamlines of two counter-rotating vortices at time t=0.25 is displayed. The domain Reynolds number is approximately 100. This computation was performed using seven multiwavelets with a precision of 10^{-5}.

ACKNOWLEDGEMENTS

This work was partially supported by the US DOE Office of Science, Office of Advanced Scientific Computing Research program in applied mathematics, the Scientific Discovery through Advanced Computing (SciDAC) SAP program and the division of Basic Energy Science, Office of Science, under contract number DE-AC05-00OR22725 with Oak Ridge National Laboratory.

REFERENCES

[1] B. Alpert. A class of bases in L2 for the sparse representation of integral operators. SIAM J. Math. Anal, 24(1):246-262, 1993.

[2] B. Alpert, G. Beylkin, D. Gines, and L. Vozovoi. Adaptive solution of partial differential equations in multiwavelet bases. J. Comput. Phys., 182(1):149-190, 2002.

[3] G. Beylkin and M. J. Mohlenkamp, "Numerical operator calculus in higher dimensions," *Proceedings of the National Academy of Sciences*, v. 99, 16, pp. 10246-10251, 2002.

[4] G. Beylkin, R. Cramer, G. Fann and R. J. Harrison, "Multiresolution separated representations of singular and weakly singular operators," *Applied and*

Computational Harmonic Analysis, 23 (2007) 235-253
[5] A. J. Chorin. “Numerical solution of the Navier-Stokes equations,” Math.Comp. 22:745-762, 1968.
[6] A. J. Chorin, “On the convergence of discrete approximations of the Navier-Stokes equations. Math.Comp. 23:341-353, 1969.
[7] M. Farge, N. Kevlahan, V. Perrier, K. Schneider, “Turbulence analysis, modelling and computing using wavelets,” Chapter 4 in Wavelets in Physics, Ed. J. C. van den Berg, Cambridge 1999.
[8] L. Greengard and J. Huang, *“A New Version of the Fast Multipole Method for Screened Coulomb Interactions in Three Dimensions,”* J. Comput. Phys. 180, 642 (2002).
[9] L. Greengard and V. Rokhlin, *“A Fast Algorithm for Particle Simulations,”* J. Comput. Phys. 73, 325 (1987).
[10] R. J. Harrison, G. Fann, T. Yanai, Z. Gan and G. Beylkin, "Multiresolution Quantum Chemistry: basic theory and initial applications,” *J. Chem. Phys*. 121 (23) (2004) 11587-11598.
[11] L. Jameson, T Adachi, O. Ukai, A. Yuasa, “Wavelet-based Numerical Methods,” Int. J. of Comp. Fluid Dynamics, 1029-0257, Vol. 10, 4, 1998, pp. 267-280.
[12] C. Meneveau, “Analysis of turbulence in the orthonormal wavelet representation,” J. Fluid Mech., 323, p. 469, 1991.
[13] R. Temam. Sur l'approximation de la solution des equations de Navier-Stokes par la methode des pas fractionaires II. Arch.Rati.Mech.Anal. 33:377-385, 1969.
[14] L. Ying, G. Biros, D. Zorin, A Kernel-independent Adaptive Fast Multipole Method in Two and Three Dimensions, Journal of Computational Physics, 196(2), 591-626, 2004.
[15] G. Biros, L. Ying, D. Zorin, “An Embedded Boundary Integral Equation Solver for the Unsteady Incompressible Navier-Stokes Equations,” submitted, 2004.

Jun Jia is a post-doctoral fellow with the Computer Science and Mathematics Division (CSMD), Oak Ridge National Laboratory (ORNL), USA. He graduated from Tsinghua University with a B.S. degree in 2000 and with a M.S. degree in computer science and a Ph.D. degree in mathematics from the University of North Carolina at Chapel Hill in 2004 and 2006. He is a member of Sigma Xi, the International Honor Society of Science and Engineering and a member of Alpha Epsilon Lambda, the National Academic Excellence and Leadership Honor Society. Dr. Jia’s research interests are fast algorithms, high order methods, integral equation methods, potential theory, and their applications in electromagnetic, solid and fluid dynamics, quantum chemistry and molecular mechanics.

Judith Hill is a technical staff member with CSMD at ORNL. She received her Ph.D. in computational science from Carnegie Mellon University in 2004. In 2005, she was named the Frederick Howes Scholar in Computational Science. Prior to joining ORNL, she was a member of the Computation, Computers, Information and Mathematics Center at Sandia National Laboratories from 2005 until 2008. Dr. Hill specializes in the development, implementation, and application of numerical methods for massively parallel computers to a variety of applications in computational fluid dynamics.

George Fann is a senior staff member with CSMD at ORNL. He received an A.B. degree from the University of California at Berkeley, and M.S. and Ph.D. degrees from MIT, all in mathematics. He received an R&D 100 Prize in 1999 for the parallel eigensolver PeIGS used in the computational chemistry software NWChem and other computational chemistry software. Dr. Fann's interests include computational harmonic analysis, fast computational methods and parallel computing. He is the author of more than 50 publications.

Robert J. Harrison is a Laboratory Fellow at ORNL. He received a Ph.D. degree from Cambridge University, England, in 1984. He currently works at the Oak Ridge National Laboratory, with a joint faculty appointment at the University of Tennessee. His major research interests are the electronic structure theory of molecules with an emphasis on high precision, and high-performance computing. Dr. Harrison was the main architect of NWChem, a large computational chemistry package for parallel computers for which he received an R&D 100 award in 1999 and the IEEE Sydney Fernbach Award in 2002. He is the author of more than 100 publications.

THE PARALLEL MODELS OF CORONAL POLARIZATION BRIGHTNESS CALCULATION*

Jiang Wenqian[1], Zhong Dingkun[2], Xie Zhangpeng[1], Wang Qun[2], Feng Xueshang[1]
[1] Information Engineering Institute, China University of Geosciences (Beijing), Beijing 100083
[2]Key Laboratory of Space Weather, Center for Space Science and Applied Research,
Chinese Academy of Sciences, Beijing 100080
Email: jwq406@126.com

ABSTRACT

The three-dimensional numerical Magnetohydrodynamics (MHD) simulation is one of the methods to study the corona and solar wind. And comparing the simulation results of the polarization Brightness (*pB*) in the low corona with the observation is important to validate the MHD models. Because of the massive data and the complexity of the *pB* model, the computation will cost too much time to visualize the *pB* in nearly real time while using a single CPU (or core). In order to reduce the time cost, two parallel models to convert MHD simulation density to *pB* are presented in the paper, which are separately based on the MPI and the CUDA programming technology. Then the paper analyses and compares the efficiency of the two models, and reaches the conclusion that the CUDA-based parallel model is more suitable for the conversion.

Keywords: Coronal Polarization Brightness, MPI, CUDA

1. INTRODUCTION

Space weather forecast needs an accurate solar wind model for the solar atmosphere and the interplanetary space. The global model of corona and heliosphere is the basis of numerical space weather forecast, and the observation basis of explaining various relevant relations[1]. Meanwhile, three-dimensional numerical Magnetohydrodynamics (MHD) simulation is one of the most common numerical methods to study corona and solar wind. Besides, calculating and converting the generated coronal electron density to the coronal polarization brightness is the key method of comparing with observation results, and is important to validate the MHD models.

pB calculation in the paper uses electron density generated by MHD models as input parameters, and then computes *pB* values of the specific cross section according to the *pB* calculation formula. Due to the massive data and the complexity of the *pB* model, the computation will cost too much time to visualize the *pB* data in nearly real time while using a single CPU (or core). According to the characteristic of CPU/GPU computing environment, this paper analyzes the *pB* conversion algorithm, implements two parallel models of *pB* calculation with MPI and CUDA, and compares the two models' efficiency. And the comparison reveals that the CUDA-based parallel model is more suitable for *pB* calculation, and provides a better parallel solution for the numerical calculation of MHD.

2. SERIAL MODEL OF PB CALCULATION

2.1 The pB Calculation Formula

pB is derived from electron-scattered photosphere radiation. It can be used in the inversion of coronal electron density and to validate numerical models. Taking limb darkening into account, *pB* calculation formula of a small coronal volume element is as followed [2]:

$$I_t - I_r = I_0 \frac{N_e \pi\sigma}{2} \sin^2 \chi [(1-\mu) A + \mu B] \tag{1}$$

$$A = \cos\Omega \sin^2 \Omega \tag{2}$$

$$B = -\frac{1}{8}[1 - 3\sin^2\Omega - \frac{\cos^2\Omega}{\sin\Omega}(1 + 3\sin^2\Omega)\ln\frac{1+\sin\Omega}{\cos\Omega}] \tag{3}$$

Where I_t and I_r are the tangential and radial component of the scattered radiation, I_0 is the brightness of solar center, N_e is an integral electron density over solar limb, σ is known as the "Thomson cross section", Ω is the angle between *OP* and a tangent from *P* to the photosphere (as shown in Figure 1), *A* and *B* are functions of Ω only, χ is the angle between *OP* and *BP*, *μ* is an empirical function of wavelength with the value of 0.63 in the paper.

As shown in Figure 1, the polarization brightness image for comparing with the observation of coronagraph can be generated through integrating the electron density along the line of sight.

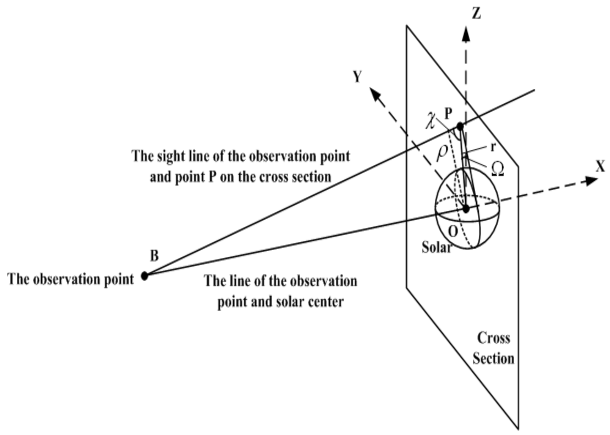


Figure 1. Density integral Process of *pB* Calculation (Note: *B* is the observation point of imager, *P* is an arbitrary point on the cross section, *O* is solar center, *ρ* is the distance between solar center and the line of sight)

The paper uses 42×42×82 (*r, θ, φ*) electron density data[3,4] as experimental data (Here, 1*Rs* <= *r* < 23*Rs*), it will generate 321×321 (-4*Rs* <= *y* <= 4*Rs*, -4*Rs* <= *z* <= 4*Rs*) *pB* values of the cross section through calculation. As shown in Figure 1, we set solar center *O* to be the original point, the connected direction of observation (*B*) and solar center (*O*) to be *x*-axis, the cross section to be *yoz* plane, then construct a 960×320×320(-12*Rs* <= *x* <= 12*Rs*,-4Rs <= *y* <= 4*Rs*,-4*Rs* <= *z* <= 4*Rs*)uniform grid in cartesian coordinate system.

*Sponsored by State Key Laboratory of Special Fund projects.

2.2 Serial Calculation Process

The steps of the serial model of *pB* calculation on CPU with the experimental data in the paper are shown in Figure 2:

(1) Read in the electron density data generated by MHD numerical simulation to store in the memory;

(2) Initialize the grid information of electron density in spherical coordinate system, specify the cross section according to the imager position in the coordinate system, and construct a 960×320×320 uniform grid in cartesian coordinate system;

(3) According to the *pB* calculation formula, each *pB* value of grid node on the cross section can be computed through integrating the polarization brightness values along the line of sight. During the process, we should determine the position of each cartesian coordinate grid point in original spherical coordinate system; afterwards, the electron density of each point along the line of sight will be interpolated;

(4) Draw coronal polarization brightness image with the calculation results.

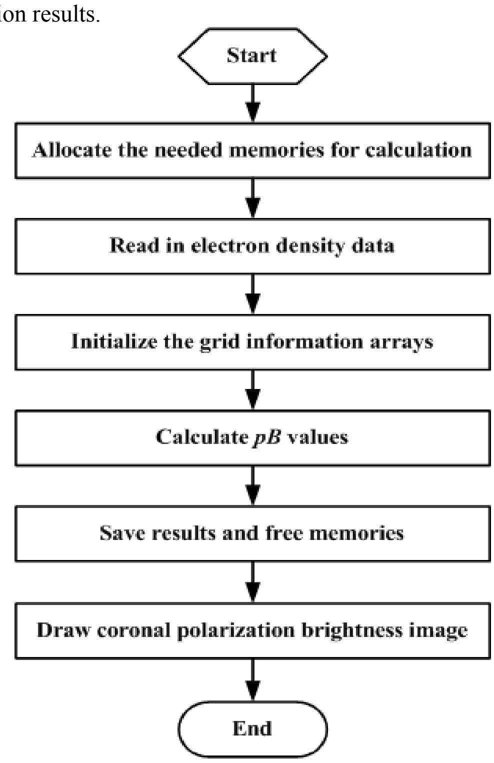


Figure 2. The serial process of *pB* calculation

According to the serial process of *pB* calculation above, we implement it under the environment of G95 on Linux and Visual Studio 2005 on Windows XP respectively. And With being measured the time cost of each step, it is found that the most time-consuming part of the whole program is lay in the calculation of *pB* values, accounting for 98.05% and 99.05% of the total time cost respectively. Therefore, in order to improve the performance to meet the command of getting coronal polarization brightness in nearly real-time, we should optimize the calculation part of *pB* values. Because the density integration of each point over solar limb along the line of sight is independent, the parallel computation method is very suitable for *pB* calculation. Currently, parallelized MHD numerical calculation is mainly based on MPI. With the development of high performance computation, using GPU architecture to solve intensive computation shows obvious advantages. Based on this situation, it will be an efficient parallel solution to implement the parallel MHD numerical calculation using GPU. In the paper, we present two parallel models based on MPI and CUDA respectively, and compare them to find a better solution for MHD numerical calculation.

3. PARALLEL COMPUTATION

3.1 MPI

MPI (Message Passing Interface) is a standard specification of message passing library, used for developing the parallel program based on message passing. It is one of most popular parallel program environments. It provides a parallel library with the support of C and Fortran languages.

MPI has a characteristic of easily used and ported, and has an advantage of complete asynchronous communication. Hence, MPI has obvious advantage on massive numerical calculation and data process. MPI has become a programming paradigm of scalable parallel computers, workstation network and clusters with distributed storage. It has been applied into many fields, such as oil exploration, earthquake prediction and forecasting, satellite image process etc. Presently, using MPI to solve massive numerical calculation problems of complex models is an important means of improving the efficiency of space weather numerical simulation and forecast. For example, SWMF(Space Weather Modeling Framework)[5,6] and NiCT (National Institute of Information and communications Technology) Space View[7,8] are both using MPI as the solution for massive numerical computation.

3.2 CUDA

With the development of graphics card, GPU graphics computing power is becoming more and more powerful, and has gain more computing capacity than the general-purpose CPU[9]. GPU has distinct advantages while handling intensive data and parallel data computation[10]. So in high performance computing, using the GPU to do the general-purpose calculation can provide a more efficient solution than that using CPU.

CUDA (Compute Unified Device Architecture) technology is based on the GPU parallel computing architecture. It uses GPU as a device for parallel computing to allocate and manage the computation[11,12]. CUDA SDK is based on the standard C language, and it does not need to apply the computing to the graphics API as traditional GPU architecture. With the characteristic, it can help the users developing their own CUDA application quickly, as long as the uses are familiar with C language.

At present, CUDA technology is in its infancy. However, it has achieved achievements in many fields, such as Video Transcoding[13], Compute Network Security[14], Analysis of Financial Stocks[15], Oil/Gas Exploration[16], Medical Diagnosis[17,18,19], Weather Forecast[20], Scientific Research[21,22,23] etc. Currently the G80 represents the mainstream NVIDIA graphics card supported CUDA, which owns 128 independent computing units. It is suitable for parallel computing, and its computing speed is much faster than CPU.

4. THE MPI-BASED AND CUDA-BASED PARALLELIZED MODEL OF PB CALCULATION

4.1 Experiment Environment

In the experiments, 42×42×82(r, θ, φ)density data(*den*), 321×321×481(x , y, z) cartesian coordinate grid are used, and 321×321 *pB* values will be generated.

The hardware environment is Intel(R) Xeon(R) CPU, E5405 @ 2.00GHz(8 CPUs), 1GB memory and NVIDIA Quadro FX

4600 GPU, 760MB Global Memory GDDR3 SDRAM graphics card (It owns G80 kernel architecture, 12 MPs and 128 SPs). And the compiling environment of CUDA-based parallel model is Visual Studio 2005 on Windows XP with CUDA 1.1 SDK, and that of MPI-based parallel model is G95 on Linux with MPICH2.

4.2 MPI-based Parallelized Implementation

In the experiment, given the density data of the whole calculating region in the spherical coordinate system, taken YOZ plane in cartesian coordinate system as solar section, the final result is a 321×321 matrix. Because every *pB* value calculation is independent, it is able to use the parallel algorithm of decomposing original computational domain to implement the *pB* parallelized calculation, that is, the whole computational domain is decomposed to multiple sub-domains and each sub-domain is considered as a computing task allocated to a node to process data. The utilization ratio of processors and the communication between processors has an effect on the total performance of the program directly. Therefore, balancing the utilization ratio of processors and the communication between processors can help reaching a better parallel efficiency. When solving a larger scale data, increasing the number of sub-domains can reduce the proportion of calculation time accounted for the total time cost. While solving a smaller scale data, decreasing the number of sub-domains can reduce the proportion of communication time accounted for the total time cost.

In the MPI environment, how the experiment decomposes computing domain into sub-domains is shown as Figure 3: the range of y-axis for each node is the ratio of the cell number along y-axis and the node number, and the range of x-axis and z-axis for each node is equal to the cell number along x-axis and y-axis respectively. The nodes are classified as a main node and child nodes. The main node is in charge of initializing, allocating, sending and collecting data, calculating sub-domains and drawing images. Meanwhile, child nodes are only responsible for calculating sub-domains, retrieving and sending data after receiving subtasks.

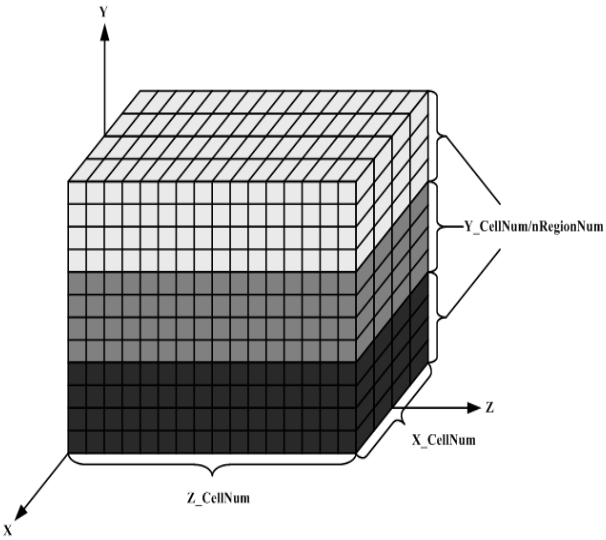


Figure 3. The division of sub-domains in MPI-based parallel model

The MPI-based parallel model of *pB* calculation is shown as Figure 4:

(1) Compute domain decomposition on the main node before calculating, then call MPI_BCAST to broadcast the needed data to each child node. The needed data in the calculation are included density data *den*, and grid information arrays *r, θ, φ, x, y, z, d*;

(2) Pass the task processed by each sub-domain to each computing node through MPI messages. Since each sub-domain process is independent, it only communicates with main node once when returning *pB* values after the calculations of all the child nodes complete;

(3) After main node completes its calculation, it will call MPI_BARRIER to wait for receiving the returned *pB* values from child nodes. While receiving *pB* values, it will merge the received values into the final 321×321 matrix simultaneously;

(4) Free memories allocated dynamically and draw coronal polarization brightness image.

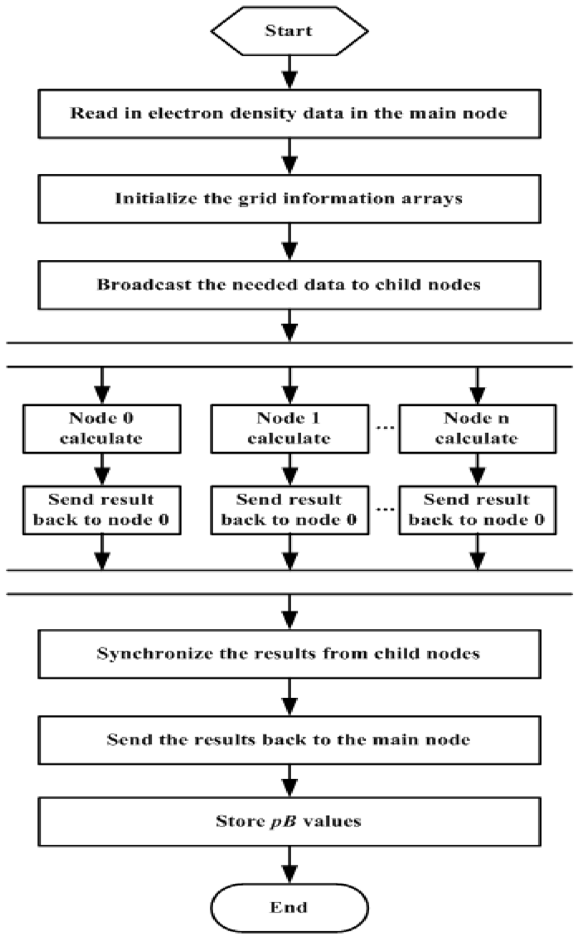


Figure 4. The parallel process of *pB* calculation in the MPI-based model

The final result shows that MPI-based parallel model reaches a speedup of 5.8. As the experiment is implemented under the platform with 8 CPU cores, the speed-up ratio of the result is closed to its theoretical value. Meanwhile, it is revealed that the MPI-based parallel solution for the experiment has balanced the utilization ratio of processors and the communication between processors.

4.3 CUDA-based Parallelized Implementation

According to *pB* serial calculation process and the CUDA architecture, we should put the calculation part into the Kernel function to implement the parallel program. Since the calculation of density interpolation and the cumulative sum involved in every *pB* value are independent, we can use multi-threads to process the *pB* value calculation in the CUDA, and each thread calculates one *pB* value. However, the *pB* values to be calculated is much larger than the available thread number of GPU, so each thread should calculate multiple *pB* values. According to experimental conditions, the thread number is setting to 256 for each block so as to maximize the use of computing resources. And the block number depends on the ratio of *pB* number and thread number. In addition, since the access time of global memory is large, we can put some independent data to the shared memory to reduce data access time.

In the CUDA-based parallel computing model, the arrays that can be put into shared memory include the grid information under two coordinate systems: *r, θ, φ, dr, x, y, z*. It generates high time cost to transmit data between CPU and GPU. Hence, we package the data into the array *Sdata* to reduce the transmission time. In order to accurately take out every part of *Sdata* in the program, an array that records the length of each part is needed. Thus, it should be indexed from original data and packaged data, and can determine the size of shared memory to be used before calling the kernel function. The size of data put into shared memory is about 7KB, less than 16KB provided by GPU, so the parallel solution is feasible. The CUDA-based parallel calculation process is shown as Figure 5. Moreover, the data-length array is read-only and its using frequency is very high, so the optimized strategy that the data-length array is migrated from shared memory into constant memory is adopted to further improve its access efficiency.

The final result shows that the CUDA-based parallel model under Visual Studio 2005 can speed up 24.65 times than the serial model based on CPU.

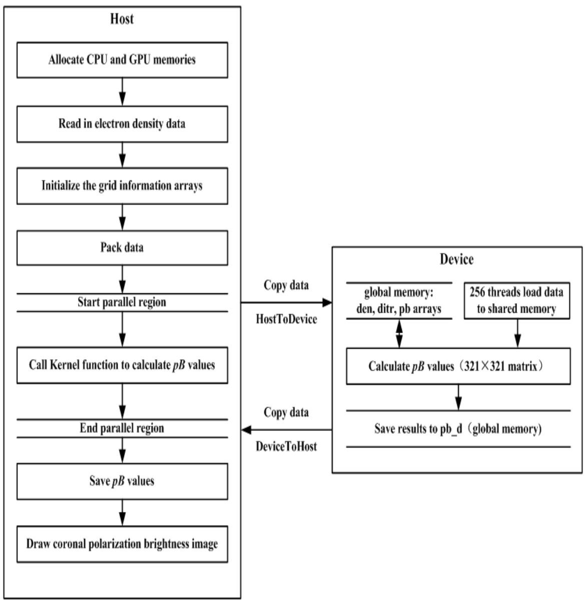


Figure 5. The parallel process of *pB* calculation in the CUDA-based model

4.4 The Comparison of Experiment Results

In the paper, we have implemented two parallel calculation models of *pB* in the MPI and CUDA environment. And these two models need data transmission before calculating. In the MPI environment, it will call MPI_BCAST to broadcast the prepared data on the main node to other child nodes. While in the CUDA environment, it will need cudaMemcpy to transmit data between CPU and GPU. Similarly, two parallel models will exchange data again to collect results after calculating. However, the data-transmission mechanisms of two models are different. The communication of MPI model is data transmission between kernels or PCs which is implemented by calling MPI functions and needs to take network transportation, protocol conversion etc. into consideration. The data transmission of CUDA model is between the North and South Bridges on the chip. And its transmission speed is based on PCI-Express Standard; it can reach 10GB/s under the current highest version. Therefore, compared to MPI model, CUDA model has obvious advantages on data transfer rate.

The *pB* calculation time of two models is shown in Table 1. With the comparison, we find that the CUDA-based implementation is better than MPI-based implementation. Because that each *pB* calculation process is same, that is, calculating the cumulative sum and interpolating density values according to known electron density data. As CUDA architecture is targeted to such model which has high computing density and the data with the same calculating process, it is more suitable for *pB* calculation. While during the MPI process, it only reduces computing domain of each node. If the data scale increases, the computation amount will increase correspondingly and its efficiency will be worse when the node number is fixed.

Table 1. The *pB* calculation time of serial models and parallel models and their speed-up ratio

	***MPI*（G95）**	**CUDA（Visual Studio 2005）**
***pB* calculation time of serial models（s）**	32.403	48.938
***pB* calculation time of parallel models（s）**	5.053	1.536
Speed-up ratio	6.41	31.86

The total performance of two models is as shown in Table 2. It is illustrated that the serial model of Fortran language is superior to that of C language. This is because that Fortran is more suitable for scientific computation. Meanwhile, the cost time of MPI-based model is 2.84 times than that of CUDA-based model as a result of data transmission time of CUDA-based model reduces and its calculation belongs to SIMD. So the CUDA-based parallel model is superior to the MPI-based parallel model.

Table 2. The total running-time of two parallel models and the speed-up ratios compared with their serial models

	***MPI*（G95）（s）**	***CUDA*（*Visual Studio 2005*）（s）**	**The speed-up ratio of running-time**
Serial models	33.05	49.406	0.67
Parallel models	5.70	2.004	2.84

Finally, we draw the coronal polarization brightness image as shown in Figure 6.with using calculated data.

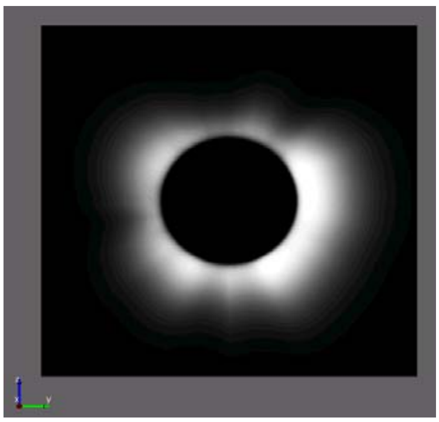


Figure 6. The coronal polarization brightness image calculated by experimental data

5. CONCLUSIONS

In the paper, we design and implement two parallel models of *pB* calculation using MPI and CUDA separately, and further improve the performance according to their respective characteristics. Under the same environment, *pB* calculation time of MPI-based parallel model costs 5.053 seconds while the serial model costs 32.403 seconds. The model's speedup is 6.41. The *pB* calculation time of CUDA-MPI-based parallel model costs 1.536 seconds while the serial model costs 48.936 seconds. The model's speedup is 31.86. And the total running-time of CUDA-based model is 2.84 times than that of MPI-based model.

With the deeper research of numerical space weather model and the continued progress of computational conditions (computing speed and storage space), the amount of data generated by MHD numerical calculation continues to shrink. Meanwhile, the real-time demand of post-process and visualization for these data is also increased. Using the original serial method can not meet the demand. In order to solve this problem, the paper implements two parallel models to meet the real-time demand of the conversion from electron density to *pB* and the visualization. And it finds that the CUDA-based parallel model is more suitable for *pB* calculation, and it provides a better solution for post-processing and visualizing the MHD numerical calculation results.

REFERENCES

[1] Xiang Changqing, Feng Xueshang, Fan Quanlin, Yao Jiusheng. An observation-based model of solar wind background. *Chin. J. Space Sci.,* 2006, 26 (3): 161~166.

[2] Billings, D. E., A Guide to the Solar Corona. Academic, San Diego, California, 1966

[3] Feng X.S., Y.F. Zhou, and S.T. Wu. A Novel Numerical Implementation for Solar Wind Modeling by the Modified Conservation Element/Solution Element Method Astrophys. *J*, 2007, 655: 1110~1126

[4] Hu, Y. Q., X. S. Feng, S. T. Wu, and W. B. Song. Three-dimensional MHD modeling of the global corona throughout solar cycle 23. *J. Geophys. Res.*, (2008) 113, A03106

[5] T´oth, G., Sokolov, I. V., Gombosi, T., et al. Space Weather Modeling Framework: A new tool for the space science community[J]. Geophys. Res., 2005, Vol. 110.

[6] DEN Mitsue, KUWABARA Takuhito, OGAWA Tomoya, et al. A 3D Visualization System for Real-Time Space Weather Simulator with a Glassless Stereoscope[J]. Journal of the National Institute of Information and Communications Technology, 2006, Vol.53 No.1.

[7] Mitsue Den, Tomoya Ogawa, Takuhio Kuwabara, et al. A Three-Dimensional Visualization System for Real-Time Earth's Magnetosphere Simulator[J]. Proceedings of ISSS-7, 2005:26-31.

[8] M. Den, T. Tanaka, S. Fujita, et al. Real-time Earth magnetosphere simulator with three-dimensional magnetohydrodynamic code[J]. SPACE WEATHER, 2006, Vol 4.

[9] Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., and Purcell, T. J. 2007. A survey of general-purpose computation on graphics hardware [J]. Computer Graphics forum, 26(1), 2007, 80-113.

[10] Buck I, Hanrahan P. Data parallel computation on graphics hardware [EB]. Stanford University Computer Science Department, 2003.

[11] NVIDIA CUDA Compute Unified Device Architecture: Programming Guide, Version 1.1. 2007

[12] Ryoo, S., Rodrigues, C. I., Baghsorkhi, S. S., Stone, S. S., Kirk, D. B., and Hwu, W. 2008. Optimization principles and application performance evaluation of a multithreaded GPU using CUDA [J]. In Proceedings of ACM PPoPP'08, Feb. 2008.

[13] Wladimir J. van der Laan , Andrei C. Jalba and Jos B.T.M. Roerdink. GPU-Accelerated Dirac Video Codec. http://www.cs.rug.nl/~wladimir/sc-cuda/, 2007

[14] Manavski, S. A. 2007. Cuda compatible GPU as an efficient hardware accelerator for AES cryptography. In: Proceedings of IEEE International Conference on Signal Processing and Communication, 2007, 11: 65~68

[15] Claudio Albanese and Alicia Vidler. Dynamic conditioning and credit correlation baskets.http://conference.mathfinance.com/2008/papers/albanese/CD03.pdf., 2008

[16] Laurent Billy. 3D visualization for oil and gas evolves.http://3dviz.mc.com/products/docs/Mercury_pr_%203DOilGasEvolves.pdf., 2006.

[17] Won-Ki Jeong, P. Thomas Fletcher, Ran Tao, and Ross T. Whitaker. Interactive visualization of volumetric White Matter connectivity in DT-MRI using a parallel-hardware Hamilton-Jacobi solver. IEEE Conference on *Visualization*., 2007

[18] Holger Scherl, Benjamin Keck, Markus Kowarschik, and Joachim Hornegger. Fast GPU-based CT reconstruction using the Common Unified Device Architecture (CUDA). Nuclear Science Symposium Conference Record., 2007

[19] Maud Marchal, Jeremie Allard, Christian Duriez, and Stephane Cotin. Towards a framework for assessing deformable models in medical simulation. In: Proceedings of the International Symposium on Computational Models for Biomedical Simulation (ISBMS), 2008

[20] Michalakes, J. and M. Vachharajani. GPU acceleration of numerical weather prediction. Parallel Processing , World Scientific., 2008, 18(4): 531~548

[21] Andrew Corrigan, John Wallin, and Matej Vesenjak. Visualization of Meshless Simulations Using Fourier Volume Rendering .Progress on Meshless Methods, Computational Methods in Applied Sciences Series., 2009, 11(291~305).

[22] Tsuyoshi Hamada, Keigo Nitadori, Tomonari Masada and Makoto Taiji. 50.5 Mflops/dollar and 8.5 Tflops Cosmological N-body Simulation on a GPU Cluster. ACM/IEEE conference., 2008

[23] Michael C. Schatz and Cole Trapnell. Fast exact string matching on the GPU. CMSC 740 Computer Graphics., 2007

THE FEATURE WEIGHTED FCM ALGORITHM WITH SEMI-SUPERVISION

Xiao-Jun Tong, Qin Jiang, Nong Sang, Hai-Tao Gan, Shan Zeng
Wuhan Polytechnic University
Wuhan ,Hubei,430074, China
Email:tongxiaojun1998@yahoo.com.cn

ABSTRACT

The traditional FCM algorithm is an unsupervised fuzzy clustering, but in real life, there are a lot of known knowledge, and a large number of samples have known information, how to take full advantage of these known information of cluster become a hotspot of research. If the known knowledge is added to the FCM algorithm in the optimal problem, we will get an optimal problem with some constraint conditions. Throng appropriate variable substitution and the thought of HPR (Hestenes-Powell-Rockafellar) algorithm, we obtain the feature weighting FCM algorithm with semi-supervision. Because the algorithm is compared with the original FCM algorithm, the number of variables is not increased, thus there is little effect on its speed of operation. IRIS data experiment shows that the algorithm not only deepens the scope of discussion of the semi-supervised FCM algorithm, but also makes the computation complexity little. Compared with the existing semi-supervised FCM algorithm, the new algorithm has greater improvement, and provides a way of thinking for the FCM algorithm with supervision, we will discuss in another text.

Keywords: FCM, The feature weighted, Semi-supervision

1. INTRODUCTION

Pattern Recognition is also called pattern classification. From the nature of dealing problem and the methods of solving problem, pattern recognition or pattern classification can be divided into supervised classification and unsupervised classification.

Cluster is a process of distinguish and classification by the similarity of samples, in the process, if there is no teacher to guide, it is unsupervised clustering; if teachers have the guidance, then it is a supervised clustering. Cluster analysis is to use mathematical methods to study and deal with the classification of some given objects.

Traditional cluster analysis is a hardware division, which put each object to identify to a strict division of classes, with one or the other nature, so the category boundaries are clear. But in fact, the majority objects did not the strict property, their nature and behavior are neutral, and they are suitable for soft partition. Fuzzy set theory proposed by Zadeh provides a powerful analytical tool for this soft partition, people began to use fuzzy clustering approach to deal with clustering questions, and call it fuzzy clustering analysis. Because Fuzzy clustering can show the uncertainty of the degree of the samples belong to various categories and the intermediary of the samples' class property, that is setting up the uncertainty description of the samples' class, and more objectively reflecting the real world, thus becoming the mainstream of cluster Analysis.

In practice, the most popular fuzzy clustering method is put forward by Ruspin [1] based on the objective function. This method is designed simply and can solve problem broadly, furthermore, it can be translated into optimization problem and use the non-linear programming theory to solve, so be achieved easily on the computer. But the real effective algorithm-fuzzy c-means algorithm (FCM) was given by Dunn [2]. Bezdek [3, 4] made it further deep, and established the fuzzy clustering theory. Since then, such a fuzzy clustering method is developed vigorously, and has a huge system currently. Thus, with the use and development of computer, the fuzzy clustering algorithm based on the objective function has become a new hotspot.

The tradition FCM algorithm is an unsupervised fuzzy clustering, and gain wide application in pattern recognition, data mining and other fields. Although FCM is an unsupervised classification algorithm, but before the clustering analysis, two parameters must be given an appropriate assignment, that are, fuzzy weight index m and the cluster number c, otherwise, they will affect the clustering result of FCM, and will directly affect the reasonable explanation of the clustering analysis [5].

People's most attention in the real life is that there are a lot of known knowledge, and a large number of samples have known information, such as data structure and statistical information, as well as the experience knowledge about object, how to take full advantage of these knowledge has become a problem. Making the known knowledge and unknown samples together to reveal the structure of the whole data set, which is semi-supervised clustering, also called partial supervised clustering. Now the more sophisticated study about semi-supervised clustering are: seed model, probabilistic model, the clustering model based on the objective function, the clustering model based on genetic algorithm, support vector machines and the clustering model based on graph theory [6]. Because of the wide application of FCM algorithm, the FCM algorithm with semi-supervision has been brought up, and gradually become a research hotspot.

At present, most of dealing known knowledge is based on the clustering process or the transformation of the objective function. Pezdracy [7] in 1985 proposed a semi-supervised FCM algorithm firstly. This algorithm introduced two improved ways:1) a new summation term, corresponding to training patterns, was added to J_m; 2) symmetric positive matrix A became the covariance matrix used here, the matrix A controls the shape of the clusters and since it uses training patterns only, it is crucial that the training set capture the shapes of clusters. Because of the calculation of the method is very complicated, and the number of variables is increasing, thus research about the FCM algorithm with known knowledge is still a hot topic of people's attention.

In the traditional FCM algorithm, each feature of the samples plays a uniform contribution for clustering. But in fact, due to the feature selection are not perfect, and their scalarization are eyeless, each feature of the feature vector is not uniform for clustering contribution, so we have to take into account the different effect of each feature. In this paper, based on the feature contribution balance principle and the most separate degree principle of intra-cluster, we construct a method of calculating the feature weight [8], and the known knowledge is added to the clustering process, finally, get the feature weighted FCM algorithm with semi-supervision. By the IRIS example, we find that the method can not only reduce number of misclassification, but also make the clustering result closer with the actual clustering.

2. FCM ALGORITHM

Suppose the samples set which should be classed is: $X=\{X_1, X_2, \cdots, X_n\} \subset R^{n\times q}$, and the number of the samples is n, q is the feature space dimension. If we want to divide the sample set of X into class of c, then the membership degree of the samples belong to the respective class is recorded as $U=\left[u_{ij}\right]_{c\times n}$ (fuzzy partition matrix): in where, $u_{ij}(1\le i\le c, 1\le j\le n)$ shows the membership degree of sample of X_j belong to class of i, and u_{ij} should be meet the following two conditions:

$$\sum_{i=1}^{c} u_{ij}=1, 1\le j\le n \tag{1}$$

$$0\le u_{ij}\le 1, 1\le i\le c, 1\le j\le n \tag{2}$$

Bezdek defined the general description of the fuzzy c-means:

$$\min J_m(X,U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \tag{3}$$

In formula (3), $V=(v_1, v_2, \cdots, v_c)$ is a matrix of $s\times c$, $v_j(j=1,2,\cdots,c)\in R^s$ is the center of class of j, $(d_{ij})^2=\left\|x_j-p_i\right\|_A=(x_j-p_i)^T A(x_j-p_i)$ shows the general distance of data point x_j and v_i. If A is identify matrix, then the corresponding distance is Euclidean distance. Euclidean distance criteria are suitable for the ultra-spherical distribution. Using the different distances definition can make the clustering algorithm suitable for different types of data clustering problem; m is fuzzy index which control fuzzy degree of matrix U.

3. THE FEATURE WEIGHTED FCM ALGORITHM

In the traditional FCM algorithm, each feature of the samples plays a uniform contribution for clustering. But in fact, due to the feature selection are not perfect, and their scalarization have some eyeless, each feature of the feature vector is not uniform for clustering contribution, so we have to take into account the different effect of each feature.

Suppose the clustering center of cursory classification be: $p_1, p_2, \ldots, p_c$, $p_i=\left\{p_{i1}, p_{i2}, \ldots, p_{is}\right\}$ .Next we will gain the feature weight by the two principles:

The principle of feature contribution balance: for ordinary classification methods, each feature contribution for clustering is important coequally; but when a feature's contribution is great than the others, we have to process the original data and change the imbalance in order to make the feature balance.

Contribution balance principle, namely each feature contribution for clustering is important coequally. Suppose the balance coefficient be r_j, we write it down as follows:

$$r_j=\frac{\max\left\{\sum_{i=1}^{c} p_{il}, l=1,2,\cdots,s\right\}}{\sum_{i=1}^{c} p_{ij}} \quad j=1,2,\cdots,s \tag{4}$$

The principle of most intra-cluster separate degree: the size of the separate degree shows that each feature has otherness, namely: the contribution of the separate degree is great, the feature weight is larger, so the new separate degree of all sorts is larger, furthermore, the separate degree is stronger.

As said by the above, we define: the weight caused by the separate degree of different sorts is equal to the separate degree. We also know that standard deviation expresses how the data points concentrate and how they separate, so we can use the clustering prototype's standard deviation to scale the separate degree of different sorts. The corresponding expression is expressed as:

$$d_j=\sqrt{\sum_{i=1}^{c}\left(p_{ij}-\bar{p}_j\right)^2} \quad j=1,2,\cdots,s \tag{5}$$

As the above principle, we get the following process: in order to process the clustering center, firstly we can use the traditional FCM algorithm to calculate the original clustering prototype, then, we normalize each feature of every clustering prototype. Thus we can get the balance coefficient of r, and the normalized separate degree of

$$d_j=\sqrt{\sum_{i=1}^{c}(r_j p_{ij}-r_j\bar{p}_j)^2} \quad j=1,2,\cdots,s \tag{6}$$

lastly, we can get the feature weight w.

From the above analysis, the feature weight w is expressed by the following form:

$$w_j=d_j\times r_j \quad j=1,2,\cdots,s \tag{7}$$

The purpose of feature weight is that the feature with larger otherness should work more contribution for classification, i.e. the feature with more separability should make its weight larger. On the contrary, if a feature's separability is smaller, we can ignore it. Sum up, we may give the separate degree of d_j a power of m, thus we have the formula (4):

$$w_j=(d_j)^{m'}\times r_j=\left(\sqrt{\sum_{i=1}^{c}(p_{ij}-\bar{p}_j)^2}\right)^{m'}\times r_j^{1+m'} \quad j=1,2,\cdots,s \tag{8}$$

From the above formula, we know: there is a relationship between the separate degree of the original clustering prototype and the balance coefficient, i.e.: the power of the balance coefficient =the power of the original clustering prototype's separate degree add 1.

Following the above analysis, we know the feature weight can be divided into two parts: the balance coefficient and the normalized separate degree. Next, we try to divide the latter into two parts: the balance and the separate degree of the original prototype, namely:

$$d_j = \sqrt{\sum_{i=1}^{c}(p_{ij} - \overline{p}_j)^2 \times r_j} \quad j = 1, 2, \cdots, s \tag{9}$$

and

$$w_j = \sqrt{\sum_{i=1}^{c}(p_{ij} - \overline{p}_j)^2 \times r_j^2} \quad j = 1, 2, \cdots, s \tag{10}$$

Now if we don't take into account the relationship between the separate degree of the original clustering prototype and the balance coefficient, we can give two powers (m_1 , m_2) to the parts, thus the feature weight w becomes:

$$w_j = \left(\sqrt{\sum_{i=1}^{c}(p_{ij} - \overline{p}_j)^2}\right)^{m_2} \times r_j^{m_1} \quad j = 1, 2, \cdots, s \tag{11}$$

4. THE FEATURE WEIGHTED FCM ALGORITHM WITH SEMI-SUPERVISION

Suppose the number of clustering samples is $n_s + n$, n_s is the number of the samples with known knowledge, and n is the number of the samples without known information. Where the sample from the first to n_1 belong to the first class, the sample from $n_1 + 1$ to $n_2 - n_1$ belong to the second class, $\cdots\cdots$,and the samples from $n_{s-1} + 1$ to $n_s - n_{s-1}$ belong to class s. now we suppose the number of class is c , and the membership matrix is $U' = (U^*, U)$,that is:

$$U' = \left[\begin{matrix} \mu_{11}^* & \mu_{12}^* & \cdots & \mu_{1n_1}^* & \mu_{1n_1+1}^* & \mu_{1n_1+2}^* & \cdots & \mu_{1n_2}^* \cdots \mu_{1n_{s-1}+1}^* & \mu_{1n_{s-1}+2}^* & \cdots & \mu_{1n_s}^* \\ \mu_{22}^* & \mu_{22}^* & \cdots & \mu_{2n_1}^* & \mu_{2n_1+1}^* & \mu_{2n_1+2}^* & \cdots & \mu_{2n_2}^* \cdots \mu_{2n_{s-1}+1}^* & \mu_{2n_{s-1}+2}^* & \cdots & \mu_{2n_s}^* \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \cdots \cdots & \cdots & \cdots & \cdots \\ \mu_{c2}^* & \mu_{c2}^* & \cdots & \mu_{cn_1}^* & \mu_{cn_1+1}^* & \mu_{cn_1+2}^* & \cdots & \mu_{cn_2}^* \cdots \mu_{cn_{s-1}+1}^* & \mu_{cn_{s-1}+2}^* & \cdots & \mu_{cn_s}^* \end{matrix}\right.$$

$$\left|\begin{matrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{c1} & \mu_{c2} & \cdots & \mu_{cn} \end{matrix}\right]$$

Thus the general formula of the feature weighted FCM algorithm with semi-supervision has the following form:

$$J_m(U', P) = \sum_{i=1}^{c}\sum_{k=1}^{n_s}(\mu_{ik}^*)^m (d_{ik}^*)^2 + \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m (d_{ik})^2, m \in [1, +\infty) \tag{12}$$

And meet the following conditions:

1) $\sum_{i=1}^{c}\mu_{ik}^* = 1, k = 1, 2, \cdots, n_s$;

2) $\sum_{i=1}^{c}\mu_{ik} = 1, k = 1, 2, \cdots, n$;

$\mu_{jk}^* \geq \mu_{ik}^*, i = 1, 2, \cdots, c;$

3) $k = n_{j-1} + 1, n_{j-1} + 2, \cdots, n_j$,

$(let n_0 = 0); j = 1, 2, \cdots s$

In the above objective function, the distance between the sample and the cluster prototype is defined as:

$$(d_{ik}^*)^2 = \left\|X_k^* - P_i\right\| = (X_k^* - P_i)^T W (X_k^* - P_i)$$

$$(d_{ik})^2 = \left\|X_k - P_i\right\| = (X_k - P_i)^T W (X_k - P_i)$$

Where, W is the feature weighted matrix.

The clustering criteria are to let $J_m(U', P)$ minimal:

$$\min\{J_m(U', P)\}$$

Now we add the post-relaxation factor ε_{ik} into inequality 3), so get an equation:

$$\mu_{ik}^* = \mu_{jk}^* - \varepsilon_{ik}$$

The above equation is substituted into the following Lagrange function:

$$\begin{aligned} F = &\sum_{i=1}^{c}\sum_{k=1}^{n_s}(\mu_{ik}^*)^m (d_{ik}^*)^2 + \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m (d_{ik})^2 \\ &+ \sum_{k=1}^{n_s}\lambda_k^*(1 - \sum_{i=1}^{c}\mu_{ik}^*) + \sum_{k=1}^{n}\lambda_k(1 - \sum_{i=1}^{c}\mu_{ik}) \end{aligned} \tag{13}$$

By Lagrange multiplier method, we gain the following formulas:

Initialization: given the number of cluster c , $2 \leq c \leq n$,the number of all data is n , set the iteration of stopping threshold ε , and set the prototype model initialization $P^{(0)}$, set iteration counter $b = 0$;

Step one: using the following formula to compute or update the division matrix: $U' = (U^{*(b)}, U^{(b)})$:

(1)For the column of k from $U^{*(b)}$, its max membership degree is:

$$\mu_{jk}^{*\,(b)} = \frac{1}{\sum_{l=1}^{c}\left(\frac{d_{jk}^{*\,(b)}}{d_{lk}^{*\,(b)}}\right)^{\frac{2}{m-1}}} \tag{14}$$

($j = 1, 2, \ldots\ldots, s$; $k = n_{j-1} + 1, n_{j-1} + 2, \ldots\ldots\ldots n_j$)

The other elements：

$$\mu_{ik}^{*\,(b)} = \begin{cases} \dfrac{1 - \mu_{jk}^{*\,(b)}}{\sum_{\substack{l=1 \\ l \neq j}}^{c}\left(\frac{d_{ik}^{*\,(b)}}{d_{lk}^{*\,(b)}}\right)^{\frac{2}{m-1}}}, \text{当}\ \varepsilon_{ik} > 0 \\ \mu_{jk}^{*\,(b)}, \text{当}\ \varepsilon_{ik} < 0 \end{cases} \tag{15}$$

($j = 1, 2, \ldots\ldots, s$, $i = 1, 2, \cdots, c, i \neq j$, $k = n_{j-1} + 1, n_{j-1} + 2, \cdots, n_j$)

(2)The membership degree matrix $U^{(b)}$ of the samples without supervision is:

$$\mu_{ik}^{(b)} = \frac{1}{\sum_{l=1}^{c}\left(\frac{d_{ik}^{(b)}}{d_{lk}^{(b)}}\right)^{\frac{2}{m-1}}} \tag{16}$$

($k = 1, 2, \ldots n; i = 1, 2, \ldots c$)

Step two: using the following formula to update the center of cluster $P^{(b+1)}$:

$$P^{(b+1)}{}_i = \frac{\sum_{k=1}^{n_s} (\mu_{ik}^{*\,(b+1)})^m X_k^* + \sum_{k=1}^{n} (\mu_{ik}{}^{(b+1)})^m X_k}{\sum_{k=1}^{n_s} (\mu_{ik}^{*\,(b+1)})^m + \sum_{k=1}^{n} (\mu_{ik}{}^{(b+1)})^m} \tag{17}$$

($i = 1, 2, \cdots, c$)

Step three: if $\left\| p^{(b)} - p^{(b+1)} \right\| < \varepsilon$, stop the algorithm and output the division matrix $U' = (U^*, U)$ and the cluster prototype P ,or, let $b = b + 1$,and do step one.

Note: if Note: If ε_{ik} is changed by ε_{ik}^2 ,it is theoretically feasible, but from the computer program, it is easy to give error, which is due to ε_{ik} is plural.

For the row of k , the maximum membership degree is:

$$\mu_{jk}^* = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{jk}^*}{d_{lk}^*} \right)^{\frac{2}{m-1}}} \tag{18}$$

($j = 1, 2, \ldots\ldots, s$; $k = n_{j-1} + 1, n_{j-1} + 2, \ldots\ldots\ldots n_j$)

The other element:

$$\mu_{ik}^* = \begin{cases} \dfrac{1 - \mu_{jk}^*}{\sum_{\substack{l=1 \\ l \neq j}}^{c} \left(\dfrac{d_{ik}^*}{d_{lk}^*} \right)^{\frac{2}{m-1}}}, when \varepsilon_{ik} > 0 \\ \mu_{jk}^*, 当 \varepsilon_{ik} < 0 \end{cases} \tag{19}$$

($j = 1, 2, \ldots\ldots, s$; $i = 1, 2, \cdots, c, i \neq j$; $k = n_{j-1} + 1, n_{j-1} + 2, \cdots, n_j$)

The unsupervised samples' membership degree is expressed as follows:

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{ik}}{d_{lk}} \right)^{\frac{2}{m-1}}} \tag{20}$$

($k = 1, 2, \ldots n; i = 1, 2, \ldots c$)

The center of clustering has the form:

$$P_i = \frac{\sum_{k=1}^{n_s} (\mu_{ik}^*)^m X_k^* + \sum_{k=1}^{n} (\mu_{ik})^m X_k}{\sum_{k=1}^{n_s} (\mu_{ik}^*)^m + \sum_{k=1}^{n} (\mu_{ik})^m} \tag{21}$$

($i = 1, 2, \cdots, c$)

Note: if ε_{ik}^2 instead of ε_{ik}, from theory, it is feasible, but from the computer program, there will be bring more errors, this is because of ε_{ik} may be plural.

5. EXPERIMENTAL ANALYSIS

In this part, we adopt the famous IRIS data to test our algorithm. The IRIS data is consisted of three clusters: Setosa, Vesicolor and Virginica, and every cluster contains fifty samples. Where Setosa is separate with Vesicolor and Virginica completely, but Vesicolor and Virginica hold common data. Now we use the traditional FCM algorithm、the feature weighted FCM algorithm and the feature weighted FCM algorithm with semi-supervision to process the IRIS data. According to their misclassification number, we can evaluate their capability.

Existing experiment tell us the optimal value of m is [1.5, 2.5] [9, 10], now we choose $m = 1.25$, $c = 3, \varepsilon = 10^{-5}$, for the latter two algorithms, let $G_{\max} = 1000$.

The experiment result shows that the method can not only reduce number of misclassification, but also make the clustering result closer with the actual clustering comparing with the traditional FCM algorithm, the feature weighted FCM algorithm.

Table 1 The clustering result of the traditional FCM algorithm, the feature weighted FCM algorithm and the feature weighted FCM algorithm with semi-supervision

Algorithm	The number of misclassified samples	The percent of misclassified samples	The vector of clustering prototype	The number and order of supervised samples	Error square sum
The traditional FCM	16	10.67%	$p_1 = (5.0062, 3.4242, 1.4684, 0.2492)$ $p_2 = (5.8946, 2.7460, 4.4154, 1.4273)$ $p_3 = (6.8484, 3.0750, 5.7283, 2.0741)$		0.1554
The feature weighted FCM	7	4.67%	$p_1 = (5.0060, 3.4280, 1.4621, 0.2460)$ $p_2 = (5.9361, 2.7441, 4.3149, 1.3289)$ $p_3 = (6.6284, 3.0159, 5.5694, 2.0668)$		0.0110
The traditional FCM with semi-supervision	11	7.3%	$p_1 = (5.0039, 3.4133, 1.4851, 0.25458)$ $p_2 = (5.9108, 2.7677, 4.3658, 1.3938)$ $p_3 = (6.7595, 3.0481, 5.6473, 2.0577)$	5 (51,53，78，102，107)	0.0671
The feature weighted FCM with semi-supervision	2	1.3%	$p_1 = (5.0061, 3.4211, 1.4699, 0.24587)$ $p_2 = (5.9752, 2.7745, 4.3410, 1.3436)$ $p_3 = (6.6549, 3.0191, 5.5966, 2.0932)$	5 (71,78,107,120,130)	0.0249
The feature weighted FCM with semi-supervision	0	0	$p_1 = (5.0064, 3.421, 1.4708, 0.2461)$ $p_2 = (5.9641, 2.774, 4.306, 1.3373)$ $p_3 = (6.646, 3.0131, 5.589, 2.0814)$	7 (71,78,107,120,130,134,135)	0.0153

6. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation under Grant 79970025, 60403002 and 30370356 of China, and the plan of Science and Technological Innovation Team of the Outstanding Young and Middle-aged Scholars of Hubei Provincial Department of Education, and Hubei Provincial Department of Education under Grant D20081802 and Hubei provincial Natural Science Foundation under Grant 2004ABA031, 2005ABA233and2007ABB030, and National Postdoctoral Science Foundation of china (Grant 2004036016), and Foundation of Hubei Provincial Department of Education Grant 2003X130 and Scientific Research of Wuhan Polytechnic University Grant 06Q15.

REFERENCES

[1] Ruspini HE. A new approach to clustering. Inf. Cont. 1969, 15: 22~32.

[2] Dunn JC. Well~ separated clusters and the optimal fuzzy partitions. J Cybernet, 1974, 4: 95~104.

[3] Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Plenum Pres, 1981.

[4] Bezdek JC. Hathaway R,et a1. Convergence and theory for fuzzy C- means clustering: counterexamples and repairs. IEEE Trans PAMI, 1987,17 (5): 873~877.

[5] Gao xinb．Fuzzy Cluster Analysis and its Applications. Xi an：the Publishing Company of Xian Electron Technology University. 2004.

[6] A. Bouchachia, W. Pedrycz. Data Clustering with Partial Supervision. Data Mining and Knowledge Discovery, 12, 47-78, 2006.

[7] W.Pedrycz.Algorithms of Fuzzy Clustering with Partial Supervision. Pattern recognition letters.No.3, pp.13-20, 1985.

[8] Xiaojun Tong, Qin Jiang, Haitao Gan, Shan Zeng, Kai Zhao. Research on the Calculation Method for Weight of the Feature Weight Fuzzy Clustering Algorithm. CDABES, July, 27-31, 2008.

[9] Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum Press, New York, 1981.

[10] Pal N R, Bezdek J C. On clustering for the fuzzy c-means model. IEEE Trans. FS, 1995, 3(3):370~379.

DISCUSSION OF FCM ALGORITHM WITH PARTIAL SUPERVISION

Hai-Tao Gan, Xiao-Jun Tong, Qin Jiang, Nong Sang, Xiao-Jiang Kong, Fang-Xiu Wang
Wuhan Polytechnic University
Wuhan, Hubei, 430023, China
Email: tongxiaojun1998@yahoo.com.cn

ABSTRACT

With the extensive application of fuzzy C-means(FCM) algorithm and a more profound understanding of objects, it has become a hot spot how to take full advantage of a priori knowledge and the combination of FCM algorithm. In this paper, FCM algorithm with partial supervision is preliminarily discussed and its recent applications of these methods are introduced. A new way to deal with a prior is proposed and the performance of FCM algorithm with partial supervision is illustrated by experimental results which indicate that the algorithm is better than the traditional FCM algorithm. Finally, some remarks on the further research and directions of the method are presented.

Keywords: FCM Algorithm, A Priori Knowledge, Partial Supervision.

1. INTRODUCTION

Pattern Recognition is also called pattern classification. From the nature of dealing with problem and the methods of solving problem, pattern recognition or pattern classification can be divided into supervised classification and unsupervised classification.

Cluster is a process of distinguish and classification by the similarity of samples, in this process, if there is no teacher to guide, it is unsupervised. If there are teachers to guide, it is supervised. Cluster analysis is to use mathematical methods to study and deal with the classification of some given objects.

Traditional cluster analysis is a hardware division, which put each object to identify to a strict division of classes, with one or the other nature, so the category boundaries are clear. In fact, the majority objects do not have the strict property. Their nature and behavior exist intermediary, and they are suitable for soft partition. Fuzzy set theory Zadeh proposed provides a powerful analytical tool for the soft partition; people began to use fuzzy clustering approach to deal with clustering questions, and call it fuzzy clustering analysis. Because Fuzzy clustering can show the uncertainty of the degree of the samples belong to various categories and the intermediary of the samples' class property, that is setting up the uncertainty description of the samples' class, and more objectively reflecting the real world, thus becoming the mainstream of cluster Analysis.

In practice, the most popular fuzzy clustering method is one that based on the objective function. This method is designed simply and can solve problem broadly, furthermore, it can be translated into optimization problem and we can use the non-linear programming theory to solve it. So it can be achieved easily on the computer. Thus, with the use and development of computer, the fuzzy clustering algorithm based on the objective function has become a new hot spot.

In the clustering methods based on the objective function, theory of FCM algorithm have the most perfect theory and the most widely application. It can be translated into optimization problem and we can use the non-linear programming theory to solve it. And then the strategy of alternate optimize is used for solving non-supervised classification problem.

The traditional FCM algorithm is an unsupervised fuzzy clustering, and has wide applications in pattern recognition, data mining and other fields. Although FCM is an unsupervised classification algorithm, but In the clustering analysis, two parameters must assign appropriate value, which are fuzzy weight index m and the classes c. Otherwise, they will affect the clustering performance of FCM, and will directly affect the reasonable explanation of the clustering analysis.

In real world, there is a lot of priori knowledge, and a large number of samples have priori information, such as data structure and statistical information, as well as the experience knowledge about an object. Supervised knowledge, also named labeled data, is used to guide the process of clustering and to enhance the accuracy of unsupervised clustering. However, unsupervised knowledge, which is unlabeled data, is used to reveal the hidden structure of the data and to improve the performance of clustering by the combination of labeled data. How to take full advantage of this knowledge has become a problem. Making the known knowledge and unknown samples together to discover the structure of the whole data set is called partial supervised clustering, which is also semi-supervised clustering.

This paper is organized as follows. Section 2 introduce the the traditional FCM algorithm. Section 3 discuss the improvement of FCM algorithm simply. Section 4 analysis the FCM algorithm with partial supervision in detail. Section 5 presents the recent applications of FCM algorithm with partial supervision. Section 6 shows the performance of FCM algorithm. Finally, Section 7 hignlights some further work.

2. THE TRADITIONAL FCM ALGORITHM

Assuming the sample set to be classified denotes as $X = \{X_1, X_2, \cdots, X_n\} \subset R^{n\times q}$, and the number of the samples is n, q is the feature space dimension. If we want to divide the sample set of X into class of C, then the membership degree of the samples belong to the respective class is recorded as $U = \left[u_{ij}\right]_{c\times n}$ (fuzzy partition matrix): in where $u_{ij}(1 \le i \le c, 1 \le j \le n)$ shows the membership degree of sample of X_j belong to class of i, and u_{ij} should be meet the following two conditions

$$\sum_{i=1}^{c} u_{ij} = 1, 1 \le j \le n \tag{1}$$

$$0 \le u_{ij} \le 1, 1 \le i \le c, 1 \le j \le n \tag{2}$$

Bezdek defined the general description of the fuzzy c-means:

$$\min J_m(X,U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \tag{3}$$

In formula (3), $V = (v_1, v_2, \cdots, v_c)$ is a matrx of $S \times C$,

$v_j (j = 1,2,\cdots,c) \in R^s$ is the center of class j, $(d_{ij})^2 = \|x_j - v_i\|_A = (x_j - v_i)^T A(x_j - v_i)$ shows the general distance of data point x_j and v_i. If A is identify matrix, then the corresponding distance is Euclidean distance. Euclidean distance criteria are suitable for the ultra-spherical distribution. Using the different distances definition can make the clustering algorithm suitable for different types of data clustering problem; m is fuzzy index which control fuzzy degree of matrix U [1].

3. DIRECTLY IMPROVEMENT OF FCM ALGORITHM

In the efforts of Ruspini [2] and Bezdek [3], fuzzy clustering algorithm has gradually become a hotspot of cluster analysis study. Since the eighties of the last century, study of fuzzy clustering have been divided into the following aspects: (1) new methods of fuzzy clustering; (2) approach of implementing fuzzy clustering algorithm; (3) studies of clustering validity; (4) the practical application of FCM algorithm.

Due to the efficiency and wide applications of FCM clustering algorithm, it is developed and deepened. And a number of types of FCM algorithm have been put forward. For this kind of research, there are mainly longitudinal studies, such as the evolution from the objective function; algorithm implementation and ways of measuring the effectiveness of methods. Researches of the objective function is mainly carried out on the fuzzy partition matrix U, the similarity criteria D (•), clustering prototype P, the weighted index m, a variety of data sets X clustering. Studies on approach of implementing fuzzy clustering algorithm mainly have the following aspects: based on alternating optimization, based on neural networks, evolutionary computation based on the implementation. Study of fuzzy clustering validity has the fuzzy partition of data sets, the geometric structure of data sets and statistical information of data sets. FCM algorithm has received a lot of successful applications in many fields, but the algorithm still has a lot of questions which need to be solved.

4. FCM ALGORITHM WITH PARTIAL SUPERVISION

Supervised clustering is also a self learning process. This process relies on the availability of knowledge about the data being analyzed. The process of assigning unlabeled data points to clusters using some similarity measure is known as clustering. This process is self-supervised. Ideally, two criteria have to be satisfied, namely intra-cluster similarity and inter-cluster dissimilarity. In fact, Acquiring knowledge of the data points is always an expensive and error-prone task that takes time and human effort. In many situations, the data is neither perfectly nor completely labeled. Thus, we may attempt to benefit from the available knowledge to cluster unlabeled data. This form of combining labeled and unlabeled data to generate the structure of the whole data set is known as semi-supervised(or partially) clustering (see Figure 1).

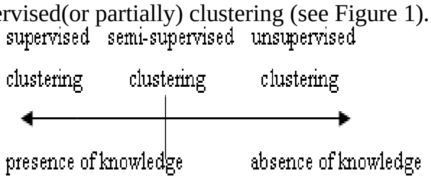


Figure 1. Clustering spectrum with respect to knowledge

Several semi-supervised algorithms have been proposed. The most known models are: the seeding model, the probabilistic model, the objective function optimization model, genetic algorithms, support vector machines, and graph-based model.

Basu [4] proposed two semi-supervised algorithms based on a seeding mechanism. These algorithms rely on the k-means algorithm. In the first algorithm, the seeds are used to initialize the partition centers and then updated during the clustering process. In the second algorithm, once initialized with the seeds, the centers are not updated. The idea here is that when seeds are noise-free, centers may be kept unchanged.

Nigam [5] investigated a probabilistic approach for text classification. The approach combines the Expectation-Maximization (EM) algorithm and a naive Bayes classifier. The algorithm trains the classifier using the labeled data only. Then, the labels of the unlabeled samples are iteratively estimated and the classifier is re-trained using all labeled data until convergence.

Support vector machines (SVM) have also been used to perform clustering with partial supervision. Klinkenberg [6] discussed the problem of information filtering as a task that requires the use of unlabeled documents to reduce the need for labeled documents known to be of interest to users. A SVM algorithm is first trained on a window of labeled data and then on a window of unlabeled data, taking care to choose the appropriate size of the window.

Graph-based methods depart from the idea that the data (labeled and unlabeled) is represented as a graph. The samples are represented as nodes which are interconnected by weighted edges. The weights indicate the similarity between samples. Basically, these approaches use a loss function and a regularization factor to propagate labels of the labeled samples to the unlabeled samples lying in the vicinity [7,8].

Pezdracy [9] firstly discussed FCM algorithm with partial supervision in 1985. In this paper, two modifications to the FCM algorithm were introduced. First, a new summation term, corresponding to training patterns, was added to J_m [equation (3)]. Second, A became the covariance matrix of the training data rather than the identity matrix used here. The matrix A controls the shape of the clusters and since it uses training patterns only, it is crucial that the training set capture the shapes of clusters.

Due to the widely application of FCM algorithm and people know more about the object, FCM algorithm with partial supervision has become a research hotspot. At present, two ways to deal with prior knowledge are that: (1) the transformation of the clustering process; (2) the transformation of the objective function. And we will propose a new way to deal with prior.

4.1 The transformation of the clustering process

Partially supervised clustering algorithm is discussed by Bensaid [10] in 1996. The main ideas in this paper are that:

(1) We denote the partially labeled data, which provide the property of partial supervision to the algorithm, as:

$$X = \left\{ \underbrace{x_1^1,\cdots,x_{n_1}^1}_{labeled\,1}, \underbrace{x_1^2,\cdots,x_{n_2}^2}_{labeled\,2}, \cdots, \underbrace{x_1^c,\cdots,x_{n_c}^c}_{labeled\,c}, \underbrace{x_1^u,\cdots,x_{n_u}^u}_{unlabeled} \right\} = X^d \cup X^u \tag{4}$$

Where $n_d = \sum_{i=1}^{c} n_i = |X^d|$, $n_u = |X^u|$, $n = |X| = n_d + n_u$;

(2) U is divided into two parts, as:

$$U = \left[\left[\underbrace{U^1_{(1)} \cdots U^1_{(n_1)}}_{e_1} \underbrace{U^2_{(1)} \cdots U^2_{(n_1)}}_{e_2} \cdots \underbrace{U^c_{(1)} \cdots U^c_{(n_c)}}_{e_c}\right] \left[\underbrace{U^u_{(1)} \cdots U^u_{(n_u)}}_{unlabeled}\right]\right]$$

$$= \left[\underbrace{U^d}_{c \times n_d} \middle| \underbrace{U^u}_{c \times n_u}\right] \quad (5)$$

The objective function is still unchanged. In the iterative process, partition matrix and clustering center is respectively updated by calculating (6) and (7):

$$u^u_{ik,t} = \left[\sum_{j=1}^{c}\left(\frac{x^u_k - v_{i,t-1}}{x^u_k - v_{j,t-1}}\right)^{2/(m-1)}\right]^{-1},$$

$$1 \le i \le c; 1 \le k \le n_u; t = 1, 2, \cdots T. \quad (6)$$

$$v_{i,t} = \left(\frac{\sum_{k=1}^{n_d}(u^d_{ik,t})^m x^d_k + \sum_{k=1}^{n_u}(u^u_{ik,t})^m x^u_k}{\sum_{k=1}^{n_d}(u^d_{ik,t})^m + \sum_{k=1}^{n_u}(u^u_{ik,t})^m}\right),$$

$$1 \le i \le c; t = 1, 2, \cdots T. \quad (7)$$

The result at [10] shows that this algorithm can classify the data set effectively. When there is less priori knowledge, the algorithm gives the labeled data a weight to overcome a tendency to recommend solutions that equalize cluster populations. But it has some shortcomings, such as: (1) it is inconsistent with the ambiguity of data that membership value of labeled data is assigned to 1 at initialization; (2) the algorithm gives the labeled data a weight, when there is less priori knowledge. The idea should not be given reasonable proof. So it does not have broad applicability.

4.2 The transformation of the objective process

Fuzzy clustering with partial supervision is discussed by Pedrycz and Waletzky [11] in 1997. It improve FCM algorithm. The objective function is:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}\mu^m_{ik}\left\|x_k - v_i\right\|^2 + \alpha\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik} - f_{ik}b_k)^m\left\|x_k - v_i\right\|^2 \quad (8)$$

Where N is the number of the samples, c is the number of classes, U is partition matrix, V is clustering center. Parameter α denotes a scaling factor whose role is to maintain a balance between the supervised and unsupervised component within the optimization mechanism. When $\alpha = 0$, the algorithm will be changed into the traditional FCM algorithm. To distinguish between labeled and unlabeled patterns, a two-valued (Boolean) indicator vector b is introduced, here $b = [b_k], k = 1, 2, \cdots, N$. When x_k is a labeled sample, $b_k = 1$ else $b_k = 0$. Similarly the membership values of the labeled patterns are arranged in a matrix form, say $F = [f_{ik}], i = 1, 2, \cdots, c; k = 1, 2, \cdots N$. The algorithm has its assumption that class is fixed in advance and reflected in the matrix F. The improvement of the objective function in the algorithm is by adding a penalty function. According to the optimization theory this method should not fundamentally solve the clustering problem with partial supervision; while the algorithm assumes that the cluster number and class number are equivalent. The former describes entire entire property of a data set, whereas the latter describes partial property. In other words, each class may contain several clusters.

Bouchachia and Pezdrycz in 2006 in the text [12] further improve the algorithm , this algorithm overcomes the limitation of the above-mentioned method. Its objective function denotes as follows:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}\mu^m_{ik}d^2_{ik} + \alpha\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik} - \tilde{\mu}_{ik})^m d^2_{ik} \quad (9)$$

Here $d^2_{ik} = \left\|x_k - v_i\right\|^2_2$, $\tilde{\mu}_{ik}$ denotes the membership degree of the labeled sample belongs to the respective class i, and it assumes that a class can be partitioned into several clusters. If H designates the number of class (labels), there is $c \ge H$.

The terms $\tilde{\mu}_{ik}$ are iteratively computed as follows:

$$\tilde{\mu}^{(r)}_{ik} = \mu^{(r-1)}_{ik} - \beta\frac{\partial Q(F,\tilde{U})}{\partial\tilde{\mu}_{ik}}$$

Here r is the iteration counter. β is a strictly positive parameter representing a learning rate, and $Q(F,\tilde{U}) = \sum_{h=1}^{H}\sum_{k=1}^{N}\delta_k(f_{ik} - \sum_{i=1}^{c}\mu_{ik})^2$. When $\delta_k = 0$, unlabeled data is handled by only the unsupervised component of the objective function. So, clustering process is mostly guided by the labeled data.

When $\alpha = 0$, the algorithm will be changed into the traditional FCM algorithm. As the scaling factor α increases, the number of misclassified points decreases until each of them is assigned to one of the clusters that fully belong to the class having the same label. This means that the number of misclassified points can be seen as a monotonically decreasing function of α.

This algorithm results show that the algorithm can solve the problem of data sets with the spatial distribution which the other simple algorithms can not solve, and can deal with the real world high-dimensional data clustering. It can make full use of the labeled data to guide the classification process and can effectively control the structure of cluster. And it has a good classification performance. However, the algorithm also has its shortcoming: (1) since the matrix F is a binary matrix, an element can only belong to a particular type, or do not belong to a particular type. When the value of the element in F is between 0 and 1, the algorithm could do nothing about it; (2) This algorithm applies to the data structure that it is spherical shape, when the data structure is hypersphere or other shape, it need further research and discussion; (3) This algorithm can only deal with the case that the cluster with a priori knowledge belongs to a class. Otherwise, it need further study.

Pedrycz and George Vukovich made another kind of transformation about the objective function in the text [13], the objective function with the following:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}\mu^m_{ik}\left\|x_k - v_i\right\|^2 + \beta\sum_{i=1}^{c}\sum_{k=1}^{N}u^p_{i_0(k)k}(f_{i_0(k)} - y_k)^2\left\|x_k - v_i\right\|^2 \quad (10)$$

Here, the parameter β strikes a balance between the structural properties of the data in the input space and the supervision component. The algorithm attempts to reconcile the data structure and the labels of the patterns forming such structure. But the algorithm is too complex, and it has cumbersome calculation process, an increase of variables, and greatly increasing computational complexity.

4.3 The transformation of the constraint

Two ways to deal with a prior are introduced above, but there are drawback respectively. In this section, we will present a new way that takes a prior as constraint. Suppose the number of clustering samples is $n_s + n$, n_s is the number of the samples with known knowledge, and n is the number of the samples without known information. Where the sample from the first to n_1 belong to the first class, the sample from n_1+1 to n_2-n_1 belong to the s econd class, $\cdots\ \cdots$,and the samples from $n_{s-1}+1$ to n_s-n_{s-1} belong to class s.

The general formula of FCM algorithm with semi-supervision has the following form:

$J_m(U',P)=\sum_{i=1}^{c}\sum_{k=1}^{n_s}(\mu_{ik}^*)^m(d_{ik}^*)^2+\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m(d_{ik})^2, m\in[1,+\infty)$ And satisfy the following conditions:

1) $\sum_{i=1}^{c}\mu_{ik}^*=1, k=1,2,\cdots,n_s$;

2) $\sum_{i=1}^{c}\mu_{ik}=1, k=1,2,\cdots,n$;

$\mu_{jk}^*\geq\mu_{ik}^*, i=1,2,\cdots,c;$

3) $k=n_{j-1}+1, n_{j-1}+2,\cdots,n_j,$

$(let\quad n_0=0); j=1,2,\cdots s$

This is an optimal problem. Throng appropriate variable substitution and Hestenes-Powell-Rockafellar (HPR) algorithm, we obtain the FCM algorithm with semi-supervision.

5. THE RECENT APPLICATIONS OF FCM ALGORITHM WITH PARTIAL SUPERVISION

In many areas, there is a lot of prior knowledge. These can guide the classification of object very well, so how to take full advantage of a priori knowledge is currently a hot topic. And FCM algorithm with partial supervision can make use of a priori knowledge to classify object. So FCM algorithm with partial supervision has a wide range of applications.

In the medical field, image segmentation [14] is a major problem. Image segmentation is to divide into several regions with different nature (such as gray level or texture), so that these regions do not cross each other. At present, there are many ways of image segmentation methods: the threshold, the edge, statistics, and so on. Every method of image processing has its own characteristics. However, because of the specificity of clinical images, there are some problems if we use the above method for medical image segmentation. So far, there is not yet a common approach. In clinic, we can get a lot of priori knowledge, such as rough structure of Organization, class number c; experienced doctors can provide some of the internal structure of organizations, so we can determine the correct classification of some pixels. How to take full advantage of priori knowledge has become a difficult problem. Semi-supervised clustering algorithm is able to deal with such problem. Text [10] proposed a semi-supervised FCM algorithm and applied to medical image segmentation. Experimental results show that the algorithm takes full advantage of a small amount of prior knowledge and segment magnetic resonance imaging (MRI) effectively.

In pattern recognition, text categorization is an important aspect. With the development of network technology, the quantity of electronic documents on network is also growing rapidly. How to help users search, filter and manage these huge amounts of data effectively has become more and more important. Therefore, Web text categorization came into being. Text classification is the automated assignment of text documents to predefined categories based on document contents. From the view of mathematics, the text categorization is a mapping process. It will map unlabeled text to labeled text. Mohammed Benkhalifa [15] discussed text categorization based on semi-supervised FCM algorithm. The algorithm exploits both class information contained in labeled data and structure information possessed by unlabeled data, and achieves a better result.

6. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the FCM algorithm with semi-supervision on the IRIS. The IRIS dataset records the physical property of there different classes of iris flowers. There are four attributes. The class Setosa is linearly separable from the other two classes, whereas the later two are not linearly separable from each other. Now we use the traditional FCM algorithm and the FCM algorithm with semi-supervision to cluster the IRIS dataset. According to their misclassification number, we can evaluate their capability.

The result shows that the FCM algorithm with semi-supervision can not only reduce misclassified numbers, but also make the cluster center closer with the actual center compared to the traditional FCM algorithm.

Table 1. Performance of Two Algorithms

Algorithm	misclassified number	misclassified rate	labeled number	labeled data	Error sum squares
The traditional FCM	16	10.67%			0.1554
The FCM with semi-supervision1	11	7.3%	5	51,53,78,102,107	0.0671
The FCM with semi-supervision2	2	1.3%	5	71,78,107,120,130	0.0249
The FCM with semi-supervision3	2	1.3%	4	71,120,130,134	0.0213

7. CONCLUSIONS

Although FCM algorithm with partial supervision can solve the classification problem with a priori knowledge, there are still a lot of problems. Existing algorithms basically improve through the clustering process and the objective function. Therefore, in this field there are some questions that need to be resolved following:

(1) An iterative method with a theoretical basis will be established through supervising the clustering process. Like the paper [8], the algorithm has descriptions only by giving example. But, it is feasible in theory that the algorithm gives the labeled data a weight to overcome a tendency to recommend solutions that equalize cluster populations

(2) The semi-supervised method based on the objective function is still rough and large calculation. It is related to 0-1 real programming, and increases the complexity of the algorithm. Looking for a semi-supervised algorithm with a small amount of calculation and low complexity still seems urgent;

(3) In the study and supervision, deviations exist in the training samples more or less. How to overcome these deviations and prevent hereditary of deviation, that is the restoration problem of training samples with deviation, will be a research hot spot;

(4) There is a succession problem of training samples in the FCM algorithm with partial supervision, In other words, by first studying and then increasing the number of training samples, how to improve learning efficiency and learning is also an important research direction.

ACKNOWLEDGEMENTS

This paper is supported by National Natural Science Foundation under Grant 79970025, 60403002 and 30370356 of China, and the plan of Science and Technological Innovation Team of the Outstanding Young and Middle-aged Scholars of Hubei Provincial Department of Education, and Hubei Provincial Department of Education under Grant D20081802 and Hubei provincial Natural Science Foundation under Grant 2004ABA031, 2005ABA233 and 2007ABB030, and National Postdoctoral Science Foundation of china (Grant 2004036016), and Foundation of Hubei Provincial Department of Education Grant 2003X130 and Scientific Research of Wuhan Polytechnic University Grant 06Q15.

REFERENCES

[1] Gao xinbo, Fuzzy Cluster Analysis and its Applications. The Publishing Company of Xian Electron Technology University, Xi’an, 2004.

[2] Ruspini HE. A new approach to clustering. Inf. Cont. 1969, 15: 22~32.

[3] Zadeh LA. Similarity relations and fuzzy orderings. Inf. Sci, 1971, 3: 177~ 191.

[4] Basu, S., Banerjee, A., and Mooney, R. Semi-supervised clustering by seeding, Proceedings of the Int, Conference on Machine Learning, pp. 19–26, 2002.

[5] Nigam, K., McCallum, A., Thrun, S., and Mitchell, T., Text classification from labeled and unlabeled documents using Expectation-Maximization, Machine Learning, 39(2/3):103–134, 2002

[6] Klinkenberg, R., Using labeled and unlabeled data to learn drifting concepts, Proceedings of the Workshop on Learning from Temporal and Spatial Data, pp. 16–24, 2001.

[7] Blum, A., Lafferty, J., Rwebangira, M., and Reddy, R., Cluster kernels for semi-supervised learning, Proceedings of the 21th International Conference on Machine Learning, pp. 92–100, 2004.

[8] Zhu, X., Kandola, J., Ghahramani, Z., and Lafferty, J., Nonparametric transforms of graph kernels for semi-supervised learning, Advances in Neural Information Processing Systems, 17:1641–1648, 2005.

[9] W.Pedrycz, Algorithms of Fuzzy Clustering with Partial Supervision. Pattern recognition letters, No.3, pp.13-20, 1985.

[10] Amine M.Bensaid, Lawrence O.Hall, James C.Bezdek And Laurence P.Clarke, Partially Supervised Clustering For Image Segmentation, Pattern Recognition, Vol.29, No5,pp.859-871,1996.

[11] W. Pedrycz and Waletzky, Fuzzy clustering with partial supervision, IEEE Transactions on Systems Man and Cybernetics, B27 (5):787–795, 1997.

[12] ABDELHAMID BOUCHACHIA, WITOLD PEDRYCZ, Data Clustering with Partial Supervision, Data Mining and Knowledge Discovery, 12, 47-78, 2006.

[13] Witold Pedrycz, George Vukovich, Fuzzy clustering with supervision. Pattern Recognition 37, pp.1339-1349, 2004.

[14] DOU Wei, HUANG Xin, YANG Wei-song, Discussion of MRI segmentation by using FCM, Chinese Journal of medical physics, Vol.17, No4, 2000.

[15] Mohammed Benkhalifa, Amine Bensaid, Abdelhak Mouradi, Text Categorization using the Semi-Supervised Fuzzy c-Means Algorithm, Fuzzy Information Processing Society. pp. 561 – 565, 1999.

HALF ANALYTIC HALF NUMERIC MULTIGRID VIRTUAL BOUNDARY PARALLEL NUMERICAL MODEL OF INVERSE HEAT CONDUCTION PROBLEMS

Jingjing Qi[1] , Shesheng Zhang[2] , Jiansheng Lin[1]
[1]School of Computer Science and Technology, Wuhan University of Technology
Wuhan 430063, China;
[2]School of Science, Wuhan University of Technology
Wuhan 430063, China
E-mail: ginjiq@126.com

ABSTRACT

The half analytic half numeric multigrid virtual boundary parallel numerical model is constructed by using invert heat transient equation. This numeric model has high speedup and high parallel efficient. In this paper, the invert heat transfer equation is discussed, and the formal exact solution is derived by using delta function. Meanwhile, the discrete multi-grid is set up according to the solution. At last, the virtual boundary parallel technique is also applied to this model.

Keywords: Half Analytic Half Numeric, Multigrid Virtual Boundary, Inverse Heat Conduction Problems, Parallel Computation.

1. INTRODUCTION

The inverse heat conduction problems can be found in most filed of engineering, and is a difficult problem for scientific research. The problem mainly incorporates two parts. One is the heat transfer factor k which is unknown that will need more calculation, another is the condition of determine solution. It is easy to get solution if we know temperature varied in space and time[1]. But it is difficult to provide this kind of temperature in practice. In fact, temperature may be given in some part sub-domain.

Tao[2] considered sideways invert heat equation, and derived stable numeric solution. Tao[2] also discussed inverse problem of one-dimensional parabolic equation by total variation regularization method. Cortes[3] took artificial neural networks to calculate inverse heat conduction problems, and found fast computer was useful to finish numerical calculation.

The main work of inverse heat conduction problems is to investigate the heat transfer factor k. It becomes more difficult if k is the function of temperature and coordination. Another is how to define the invert problem condition that is used to find solution.

2. INVERSE HEAT CONDUCTION PROBLEMS

The initial boundary inverse heat conduction problems may described as:

$$\begin{cases} u_t = \nabla(k\nabla u) + g \\ F_0(t=0, X, u, k) = 0 \\ F_x(X \in D, t, u, k) = 0 \\ F_k(X \in D_0, t \in T_0, u, k) = 0 \end{cases} \tag{1}$$

Where k=k(X,u) is heat transfer factor. D is boundary of domain D_T, D_0 is a subdomain that determine the solution k. That is, heat transfer k is a unknown function. T_0 is time subdomain. Condition F_k=0 is used to determine factor k. If factor k is constant, we only need one point temperature value. If k=k(X) is not dependent temperature u, we only need temperature on a time t0. If k=a+bu, two point u value are need. Above discuss show that the solution condition must given according to the factor k.

If $k=au^{m-1}$, Eq.(1) can be written as:

$$\begin{cases} u_t = \dfrac{a}{m}\Delta u^m + g \\ F_0(t=0, X, u, k) = 0 \\ F_x(X \in D, t, u, k) = 0 \\ F_k(X \in D_0, t \in T_0, u, k) = 0 \end{cases}$$

Or written as

$$\begin{cases} U_t = k\Delta U + kg \\ F_0(t=0, X, u, k) = 0 \\ F_x(X \in D, t, u, k) = 0 \\ F_k(X \in D_0, t \in T_0, u, k) = 0 \end{cases}$$

Above equation may consider as $U=u^m$ transfer equation. We can calculate $U=u^m$ first, and then calculate u. This ideal may extend to that k=k(u) is a polynomial of u,

$$k = \sum_{m=1}^{M} m a_m u^{m-1} = a_1 + 2a_2u^1 + a_3u^2 + \ldots + Ma_M u^{M-1}$$

We can easily get

$$ku_t = U_t$$
$$k\nabla u = \nabla U$$

where U is integration of k about u

$$U = \int k(u)du = \sum_{m=1}^{M} a_m u^m$$

In this case the Eq.(1) can be written as:

$$\begin{cases} U_t = k\Delta U + kg \\ F_0(t=0, X, u, k) = 0 \\ F_x(X \in D, t, u, k) = 0 \\ F_k(X \in D_0, t \in T_0, u, k) = 0 \end{cases}$$

From above equation, there only be partial derive about U. This shows us a way to calculate U first, and then calculate temperature u.

For a special case, temperature u is given, the Eq(1) may write as

$$b_x k_x + b_y k_y + b_z k_z = b_0 k + g_0$$

Here b_x, b_y, b_z, b_0, g_0 are (X,t) function calculated by using temperature u. Above is first order equation, easy calculate by using difference method.

For one dimension problem, above formula reduce to

$$b_x k_x = b_0 k + g_0$$

We can easily get its analysis solution

$$k = e^{P(x)} \int_0^x [e^{-P(\zeta)} g(\zeta) + c] d\zeta$$

Where,

$$P(x) = \int \frac{b_0(x)}{b_x(x)} dx$$

The formula will become simple if b_x=0, in this case, we have

$$k = \frac{g_0(x)}{b_0(x)}$$

We know, for unbounded Poison equation Δu=f(X), the solution is

$$U(X) = \int G(X, X') f(X') dX' \tag{2}$$

where G(X,X') is the green function.
For one dimension

$$G(X, X') = (X - X')$$

For two dimension problem

$$G(X, X') = -\frac{1}{2\pi} \ln \frac{1}{|X - X'|}$$

For three dimension problem

$$G(X, X') = -\frac{1}{4\pi |X - X'|}$$

Form Eq(1), we easy get

$$f(t) = (u_t - \nabla k \nabla u - g) / k \tag{3}$$

Eq(2) is form integral formula of Eq(1).

3. MULTIGRID AND PARALLEL CALCULATION [4]

By using multigrid method[5], consider coarse grid C_0, the numeric solution is u_0 on fine grid the error u_1=u-u_0 is also can be get by using above method[6]. The domain D_T is vivid to Np subdomain. Np is the process number. Data is exchange between the boundaries of subdomain. The virtual technique is used to reduce the times of data exchange[7]. The calculation steps are:

(1) give k value on the domain D_T;
(2) calculate function f from Eq(3);
(3) get new temperature u from Eq(2);
(4) calculate error from F_k in Eq(1) on fine grid
(5) if error is larger than the value we given, go to step (1); otherwise go to next step
(6) out put convergent result k.

In the Eq.(3), there are first derive of temperature about time t. We may calculate it as :

$$f(t) = ((u\ (t-\tau) - u\ (t-2\tau)) / \tau - \nabla k \nabla u - g) / k$$

That is, the derive u_t is calculate by using u(t-τ) and u(t-2τ). If k=constant, above formula can reduce to:

$$f(t) = \frac{1}{k\tau}[u(t-\tau) - u(t-2\tau) - g]$$

In the case k=k(u), f(t) may represented as:

$$f(t) = \frac{1}{k\tau}[U(t-\tau) - U(t-2\tau)] - g$$

where U is

$$U(X) = \int G(X, X') f(X') dX'$$

Or

$$U(X) = \int \frac{1}{k\tau}[G(X, X')U(t-\tau) - G(X, X')U(t-2\tau)] - G(X, X') g dX'$$

Or write as three terms:

$$U(X) = \int \frac{1}{k\tau} G(X, X') U(t-\tau) dX' - \int \frac{1}{k\tau} G(X, X') U(t-2\tau) dX' - \int G(X, X') g dX'$$

We find that there are not data transfer between subdomain when the analysis solution be calculated in the subdomain. Only data exchanged on the fine grid. That is, on the coarse grid, data is not exchanged between subdomain. This way may improve the parallel speed up.

4. EXAMPLE

Consider one dimension example:

$$\begin{cases} u_t = \dfrac{d}{dx}(k \dfrac{du}{dx}) \\ u(t=0) = x^2 \\ u(x=0) = 0 \\ u(x=1) = 1 \\ u(x=0.5) = 0.25 \\ 0 < x < 1; t > 0 \end{cases}$$

We can get

$$u(t, x, k) = u_0 x + \int_0^x (x - y) f(y, t, k) dy$$

$$f(x, t, k) = (u_t - k_x u_x) / k$$

Here u_0 is constant, determined by u(x=1)=1 condition. Let xi=ih, h is step length of x, t_m=mτ, τ is step length of time t. u^k_i=u(xi,tk), the numeric formula for temperture u is

$$u_{i+1}^m = u_0 (i+1)h + \int_0^{ih} (h + ih - y) f(y, t_m, k) dy + \int_{ih}^{ih+h} (h + ih - y) f(y, t_m, k) dy$$

Or write as

$$u_{i+1}^m = u_i^m + h u_0 + h \int_0^{ih} f(y, t_m, k) dy + \int_{ih}^{ih+h} (h + ih - y) f(y, t_m, k) dy$$

We have numeric calculative formula of u as:

$$u_{i+1}^m = u_i^m + u_0 h + h^2 \sum_{j=0}^{i} f(i, m)$$ Here f(i,m)=f(x=x_i,t=t_m,k).

Above formula is easy calculated, if factor k is given. We consider simple case of that k is constant, we have $k_x=0$, and take $f(x,t,k)=x^2+2kt$.

The space step h=0.01; τ=0.1; by using condition u(x=0.5)=0.25, we get numeric result k=0.943396.

5. CONCLUSIONS

We derived inverse heat conduction problems calculation formula, gave the condition to determine the heat transfer factor k. The reduce model k=k(X) and k=k(u) are discussed, and analysis formula for k=k(u) is given. The main results in this paper are,
1. factor k solution condition is given;
2. Construct half analysis half numeric model;
3. On coarse grid, data is not exchange between subdomain when analysis formula is calculated. Only exchange data on the fine grid. In this way, the parallel efficient will be improved.
The paper give a numeric one dimension example, result shows our algorithm can work.

REFERENCE

[1] Jungai Tian, Tingyan Xiao; Application of Total Variation-based Regularization to Inverse Problem of a Kind of Heat Equation; Journal of Civil Aviation University of China, 2004, No3.

[2] Tao Min, Haixia Wu; Solving inverse problem of one-dimensional parabolic equation by total variation regularization method; Journal of Computer Applications, 2008, No.1;

[3] Cortes, O.Urquiza, G.Hernandez, J.A.Cruz, M.A.CERMA; Artificial Neural Networks for Inverse Heat Transfer Problems**;** Electronics, Robotics and

[4] Automotive Mechanics Conference 2007; Cuernavaca, Morelos, Mexico 2001.

[5] H. Demuth, M. Beale, and M. Hagan. Neural Network Toolbox5: User’s Guide. The Mathworks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098, 2007.

[6] Convergence models for parallel multi-grid solution of non-linear transient equation using virtual boundary forecast method, to be published, 1999

[7] Guo Qingping, Y. Paker et al. “Optimum Tactics of Parallel Multi-grid Algorithm with Virtual Boundary Forecast Method Running on a Local Network with the PVM Platform”. Journal of Computer Science and Technology, Vol.15, No.4, July 2000：355-359.

[8] Guo Qingping，Y. Paker et al. Parallel Multi-grid Algorithm with Virtual Boundary Forecast Domain Decomposition Method for Solving Non-linear Heat Transfer Equation. Lecture Notes in Computer Science , High Performance Computing and Networking, Spreinge Press, May 2000, Vol . 1823：568-571.

Jingjing Qi is a postgraduate student studying for doctor degree in the area of Computer Application at Wuhan University of Technology with research interests of distributed parallel computing. She graduated from Wuhan University of Technology in 2007. Her research interests are in inverse problems, specifically in heat conduction problems.

IMPROVED GAUSS-NEWTON ITERATION METHOD FOR NONLINEAR RANDOM EFFECT MODELS WITH RIGHT-CENSORED DATA *

Zong Xu-ping[1],Feng Guolin[2]
[1]School of Mathematics, Yangzhou University
Yangzhou, Jiangsu, 225002, China
[2]Chinese National Climate Center 100081, China

ABSTRACT

This paper presents modified Gauss-Newton iteration algorithm for the nonlinear regression models with random effects for Failure Time Data set. The convergence of the iteration is proved carefully. Simulation illustrated that our method is available. Our results may be regarded as an extension of Wei (1998) for exponential nonlinear regression models without failure time data.

Keywords: Gauss-Newton Iteration Method, Right-Censored Data, Nonlinear Random Effects Models

1. INTRODUCTION

The problem of analyzing time to event data arises in a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography. A common feature of these data is that they contain censored observations, especially right-censored data. And there is an enormous literature on dealing with these data sets. One simple method proposed by Aitkin (1981), we use this method to deal with nonlinear model for right censored data in this paper,. This type of model has been used for analyzing correlated survival observations (Hougaard 1986). For Cox's proportional hazards model (Cox 1972) with a gamma-fragility, inference procedure has been proposed by Klein (1992). Zhang et. al (1998) consider inference for a semi-parametric stochastic mixed model for longitudinal data, and Cai, Cheng and Wei (2002) discussed Cox model for analyzing univariate failure time data failure time data by using semi-parametric mixed-models. In this paper, we proposed the modified Gauss-Newton iteration method to deal with the nonlinear model with random effects for failure time data, especially right-censored data. The method has an advantage that it often convergences quickly, while the traditional Gauss-Newton iteration method strongly depends on the initial value. The simulation results show that our method is available in practice.

2. LIFE TIME NONLINEAR REGRESSION MODELS

Consider a life-testing experiment in which $m+n$ items are put on test and m items still survive at the conclusion of the test. Suppose that Y is an $(m+n)\times 1$ observed vector of y_i.Let the life times y_i $(i=1,2,\cdots,n+m)$ be independently and normally distributed with mean μ_i and common variance σ^2, and without loss of generality that the last m lifetimes are censored because of termination of the experiment.

A nonlinear model with random effects discussed in this paper is defined as follows:

$$Y = f(x,\beta) + Zu + \varepsilon, \varepsilon \sim N(0,\sigma^2 I_{n+m}) \tag{1}$$

where $f(x,\beta)=f(\beta)$ is known function with $f_{(i)}(\beta)=f(x_i,\beta)$ as its i-th element, X, Z are $(m+n)\times p$ and $(m+n)\times q$ design matrixes with x_i^T and z_i^T as the i-th row respectively and Z is a matrix with elements of 0 or 1; β is a $p\times 1$ unknown parameter vector having fixed value (fixed effect) defined, and u is a $q\times 1$ unobservable vector (random effect) with $u \sim N(0,\sigma^2\Sigma)$. In general Σ is a known diagonal matrix. ε is an $(n+m)\times 1$ vector of random errors, which are independent of the random effects u. Let

$$\phi(y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}, \Phi(y)=\int_y^{+\infty}\phi(t)dt, t_i=\frac{y_i-\mu_i}{\sigma}, S(y)=\frac{\phi(y)}{\Phi(y)}$$

From the above assumptions, $Y|u \sim N(f(\beta)+Zu,\sigma^2 I_{n+m})$, Following Robinson (1991), we use joint density function as the likelihood to deal with the random effects, see also Breslow and Clayton (1993) and Lee and Nelder (1996). So we have

$$L(Y,\beta,u)=\frac{1}{\sigma^n}\prod_{i=1}^{n}\phi(t_i)\cdot\prod_{i=n+1}^{n+m}\Phi(t_i)\cdot\frac{1}{(\sqrt{2\pi\sigma^2})^q}\exp\{-\frac{1}{2}u^T\Sigma^{-1}u\} \tag{2}$$

then the joint log-likelihood function can be written as

$$l(\beta,u)=-\frac{n+q}{2}\log 2\pi\sigma^2-\frac{1}{2}\log|\Sigma|-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-f(x_i,\beta)-z_i^Tu)^2$$
$$+\sum_{i=n+1}^{n+m}\log\Phi(t_i)-\frac{1}{2\sigma^2}u^T\Sigma^{-1}u \tag{3}$$

Differentiating the above formula to β, we get

$$\dot{l}(\beta)=\frac{1}{\sigma^2}\sum_{i=1}^{n}\tfrac{\partial f(x_i,\beta)}{\partial\beta}(y_i-f(x_i,\beta)-z_i^Tu)+\frac{1}{\sigma}\sum_{i=n+1}^{n+m}S(t_i)\tfrac{\partial f(x_i,\beta)}{\partial\beta}$$
$$=\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\tfrac{\partial f(x_i,\beta)}{\partial\beta}(y_i-f(x_i,\beta)-z_i^Tu)+\sum_{i=n+1}^{n+m}\tfrac{\partial f(x_i,\beta)}{\partial\beta}\sigma S(t_i)\right]$$
$$=\frac{1}{\sigma^2}\sum_{i=1}^{n+m}\tfrac{\partial f(x_i,\beta)}{\partial\beta}(g_i-f(x_i,\beta)-z_i^Tu) \tag{4}$$

where

$$g_i=\begin{cases} y_i, i=1,2,\cdots,n \\ \sigma S(t_i)+f(x_i,\beta)+z_i^Tu, i=n+1,\cdots,n+m \end{cases}$$

Differentiating (3) to u, we get

$$\frac{\partial l}{\partial u}=\frac{1}{\sigma^2}\sum_{i=1}^{n}z_i(y_i-f(x_i,\beta)-z_i^Tu)+\frac{1}{\sigma}\sum_{i=n+1}^{n+m}z_iS(t_i)-\frac{1}{\sigma^2}\Sigma^{-1}u$$
$$=\frac{1}{\sigma^2}\left[\sum_{i=1}^{n+m}z_i(g_i-f(x_i,\beta)-z_i^Tu)-\Sigma^{-1}u\right]=0$$

The solution to above equation is

$$\tilde{u}=(Z^TZ+\Sigma^{-1})^{-1}Z^T(G-f(\beta)) \tag{5}$$

where G is the vector with the element g_i. Substituting the above formula into (4), we get

$$\dot{l}_p(\beta)=\frac{1}{\sigma^2}D^T\Omega^{-1}(g_i-f(x_i,\beta))=\frac{1}{\sigma^2}D^T\Omega^{-1}e \tag{6}$$

* The research is supported by National Science & Technology Support Plan 2007BAC01.
Email: xpzhong@yzu.edu.cn

where of $e = G - f(\beta), \Omega = I_{n+m} + Z\Sigma Z^T, D = \partial f(\beta)/\partial\beta,$ it is called profile Score function. Then $\hat{\beta}$ can be regarded in form as the generalized least-square estimator of the following nonlinear model.

$$G = f(x,\beta) + e, e \sim (0, \sigma^2\Omega) \tag{7}$$

From a simple computing, we can get

$$\ddot{l}_p(\beta) = -\frac{1}{\sigma^2} D^T\Omega^{-1}D + \frac{1}{\sigma^2}[e^T\Omega^{-1}][W] \tag{8}$$

where $W = \partial^2 f(x,\beta)/\partial\beta\partial\beta^T$, $[\circ][\circ]$ denotes he array multiplication, see Wei (1998) Appendix A for details. We assume that regular conditions such as Wei (1998) for our model are satisfied, in particular, we assume that

$$\lim_n \frac{D^T\Omega^{-1}D}{n} = K(\beta) \cdot \tag{9}$$

Let $\hat{\beta}$ be the maximum likelihood estimate of β, then it follows from (3) that $\hat{\beta}$ satisfies

$$D^T(\hat{\beta})\Omega^{-1}\hat{e} = 0 \text{,} \tag{10}$$

Yu and Wong (2002) gave a detail proof that $\beta^i \to \hat{\beta}, a.e.$ under some regular conditions. It means that the solution to equation (6) exists and it is unique.

The above equation shows that in Euclidean space R^{n+m}, the "residual vector" $\hat{e} = e(\hat{\beta})$ is orthogonal to the space spanned by column vectors of $D(\hat{\beta})$ with respect to the matrix Ω^{-1} inner product. Combining this geometric interpretation with the geometric framework of nonlinear regression models presented by BW[2], we can introduce a modified BW geometric framework for our models (1) as follows.

Take $\eta = f(\beta)$ as a coordinate in Euclidean space R^{n+m}, then $\eta = f(\beta)$ may be called solution locus. It is easily seen that the tangent space T_β is spanned by the columns of $D(\beta)$. For any two vectors a and b in R^{n+m}, we define an inner product as $<a,b> = a^T\Omega^{-1}b$. Under this inner product, the corresponding normal space is denoted by $T_\beta^{'}$. We can define curvature arrays for the solution locus $\eta = f(\beta)$, connected with the model (1). To this aim, we choose the orthogonal basis for spaces T_β. Suppose that the QR decomposition of $D(\beta)$ under the inner product is given by $D(\beta) = (Q,N)\begin{pmatrix} R \\ 0 \end{pmatrix}$ where R and $L = R^{-1}$ are $p \times p$ nonsingular upper triangular matrices and the columns of Q and N are orthogonal basis for the tangent space and the normal space of locus $\eta = f(\beta)$ at β. The matrices Q and N satisfy $Q^T\Omega^{-1}N = 0, N^T\Omega^{-1}N = I_{n+m}$.

Now we define the intrinsic curvature array A^I and parameter-effects curvature array A^P as:

$$A^I = [N^T\Omega^{-1}][U], A^P = [Q^T\Omega^{-1}][U], U = L^T WL$$

Note that for an inner product space with weight Ω^{-1}, the projection operator of matrix D is $P_D = D(D^T\Omega^{-1}D)^{-1}D^T\Omega^{-1}$ and satisfies that $P_D^2 = P_D$ and and $P_N = NN^T\Omega^{-1}$ are orthogonal projection operators of tangent space. It is also easy to show that:

The geometric framework introduced above seems similar to that defined by BW[2], so it may be called the modified BW(MBW) geometric framework. But there are some differences between our MBW framework and BW framework.So we can study the model from the view of differential geometry, see Zong (2000).

3. MODIFIED GAUSS-NEWTON ITERATION METHOD

From (3) and (4), taking a first order Taylor series approximation to the likelihood equation $\dot{l}(\beta) = 0$, we obtain

$$l(\hat{\beta}) \approx \dot{l}(\beta_0) + \ddot{l}(\beta_0)(\hat{\beta} - \beta_0) \approx 0$$

$$\hat{\beta} \approx \beta_0 + (-\ddot{l}(\beta_0))^{-1}\dot{l}(\beta_0)$$

Therefore the Gauss-Newton iteration procedure can be expressed as

$$\beta^{i+1} = \beta^i + \{-\ddot{l}(\beta^i)\}^{-1}\dot{l}(\beta^i), i = 0,1,2,\cdots \tag{11}$$

Where β^i is the *i*-th iteration. In this iteration equation, the observed information $-\ddot{l}(\beta)$ is usually replaced by the expected information $D^T\Omega^{-1}D/\sigma^2$, this replacement is convenient and acceptable (e.g. Wei, 1998). Then the Gauss-Newton iteration procedure of $\hat{\beta}$ is given by

$$\beta^{i+1} = \beta^i + (D^T\Omega^{-1}D)^{-1}D^T(G - f(x_i,\beta))|_{\beta=\beta^i} \tag{12}$$

Equation (12) shows that the iteration procedure can be regarded as a formula of the generalized least squares estimator for the following general nonlinear regression

$$G = f(x,\beta) + \delta, \delta \sim (0,\Omega) \tag{13}$$

For the further discussion, the reader is referred to literature, such as Wei (1998). In general, the convergence of iteration (12) is quite quick, but it strongly depends on the initial value $\beta^{(0)}$. Wei (1998) gave a set of regular conditions under which Gauss-Newton iteration procedure of nonlinear least squares estimator has asymptotically numerical stability, i.e. $\beta^i \to \hat{\beta}, a.e.$ as $i \to \infty$, when the initial value $\beta^{(0)}$ is in a certain neighborhood of β_0. For the model in this paper, the asymptotically numerical stability can be obtained, see Yu and Wong (2002) for detail. In practice, the modified Gauss-Newton iteration procedure is often used. To introduce this method, we need the following theorem.

Theorem For any β satisfied $\dot{l}_p(\beta) \neq 0$, there exists a value $\lambda^* > 0$, such that $l_p(\beta + \lambda K(\beta)) > l_p(\beta)$ for $0 < \lambda < \lambda^*$, where $K(\beta)$ is the second term of (12), that is

$$K(\beta) = (D^T\Omega^{-1}D)^{-1}D^T(G - f(x_i,\beta)) \cdot \tag{14}$$

Proof . Denote $q(\lambda) = l_p(\beta + \lambda K(\beta)), \lambda > 0$, then we have

$$\begin{aligned} l_p(\beta + \lambda G(\beta)) - l_p(\beta) &= q(\lambda) - q(0) \\ &= \dot{q}(0)\lambda + \alpha\lambda \\ &= [\dot{q}(0) + \alpha]\lambda \end{aligned}$$

where $\alpha \to 0$ as $\lambda \to 0$. It follows from (3) that $\dot{q}(0)$ can be expressed as

$$\begin{aligned} \dot{q}(0) &= \sum_{r=1}^k [\frac{\partial l_p}{\partial \beta_r} \frac{d(\beta + \lambda K(\beta))_r}{d\lambda}]_{\lambda=0} \\ &= \sum_{r=1}^k \frac{\partial l_p}{\partial \beta_r} K_r(\beta) \\ &= [\dot{l}(\beta)]^T K(\beta) \end{aligned}$$

$$=[\dot{l}(\beta)]^T[-\ddot{l}(\beta)]^{-1}[\dot{l}(\beta)]>0$$

So there exists a value $\lambda^*>0$, such that $\dot{q}(0)+\alpha>0$ when $0<\lambda<\lambda^*$. This results in $l(\beta+\lambda K(\beta))>l(\beta)$.By this theorem, we can summarize the modified Gauss-Newton iteration as follows.

(a) Choose an initial value β^0, and compute $K_0=K(\beta^0)$. Find a $0<\lambda_0<1$,such that $l(\beta^0+\lambda_0 K_0)=\max_{0<\lambda\le 1} l(\beta^0+\lambda K_0)>l(\beta^0)$

(b) Let $\beta^1=\beta^0+\lambda_0 K_0$. Compute $K_1=K(\beta^1)$ and find a $0<\lambda_0<1$, such that $l(\beta^1+\lambda_1 K_1)=\max_{0<\lambda\le 1} l(\beta^1+\lambda_1 K_1)>l(\beta^1)$

(c) Let $\beta^2=\beta^1+\lambda_1 K_1,\cdots$

It is easily seen that by this procedure, we have $l(\beta^0)<l(\beta^1)<l(\beta^2)<\cdots<l(\beta^i)<l(\beta^{i+1})<\cdots$, and may have $l(\beta^i)\to l(\hat{\beta})$ under some regular conditions (wei, 1998). There are several ways for choose value λ_i. Wei (1998, p27) suggested quite simple values of λ s, which are $\lambda=2^{-1},2^{-2},2^{-3},\cdots$.

4. SIMULATION

Generate 40 numbers from the standard normal distribution e_{ij}, then generate 40 numbers x_{ij} from the normal distribution $N(0,1)$, Let $\beta=(1,2)^T,\sigma=1$, we regard σ as nuisance parameter. Assume the model is

$$y_{ij}=e^{\beta_0+\beta_1 x_{ij}}+u_i+e_{ij}$$
$$x_{ij}\sim N(0,1),\quad u_i\sim N(0,i\sigma^2),\quad \varepsilon_{ij}\sim N(0,\sigma^2),$$
$$i=1,\cdots,10,\quad j=1,\cdots,4$$

We can get the 40 response values y_i, let y_{39},y_{40} be censored by 5;

By using the modified Gauss-Newton iteration procedure $\beta^1,\beta^2,\cdots\beta^i,\beta^{i+1}$,until $\left\|\beta^{i+1}-\beta^i\right\|^2<10^{-8}$, then stop the iteration;

To repeat the step above 1000 times, then we can get 1000 values of β, if the mean is close to the real value $\beta=(1,2)^T$, and the variance is very small, that means the modified Gauss-Newton iteration procedure is successful. By using the MATLAB code, we can get that the mean of β is $\overline{\beta}=(1.0040,2.0255)$; while the variance of β is $s=(1.0\times10^{-3},2.1\times10^{-3})$. This result illustrates that the method is quiet available. The matlab code cold be obtained by email from the author.

This paper proposes the Modified Gauss-Newton iteration algorithm for the nonlinear regression models with random effects for Failure Time Data set, we can also use this method to deal with the other models, such as nonlinear models with random effects, measurement error models and failure data, the model referred to Zong (2002).

REFERENCES

[1] Aitkin, M. (1981) A note on the regression analysis of censored data, *Technometrics*, 23, 161-163.

[2] Bates, D.M. and Wates, D.G. (1980)Relative curvature measures of nonlinearity. *J. Roy. Statist. Soc. Ser.B*, 42, 1-25.

[3] Breslow, N.E. and Calyton, D.G., (1993) Approximate inference in generalized linear mixed models. *JASA*, 88, 9-25.

[4] Cai, T.; Cheng, S. C. and Wei, L. J. (2002) Semiparametric Mixed-effects Models for Cluster Failure Time Data, *JASA*, 97, 514-522.

[5] Cox, D. R. (1972), Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 34, 187-220.

[6] Hamilton, B. C. (1986) Confidence regions for parameter subsets in nonlinear regression. *Biametrika*, 73, 57-64.

[7] Hougaard, P. (1986) A class of Multivariate Failure Time Distributions, *Biometrika*, 73, 671-678.

[8] Klein, J. P. (1992) Semiparametric Estimation of Random effects using Cox model based on EM algorithm, *Biometrics*, 48, 795-806.

[9] Lee, Y. and Nelder, N.A., (1996)Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B*, 58, 619-678.

[10] Robinson, G.K., (1991)That BLUP is a good thing: the estimation of random, *Statist. Sci.*, 6, 15-51.

[11] Zhang D., Lin, X. Raz, J. and Sowers, M. Semiparametric Stochastic Mixed Models for longitudinal. *JASA*, 93, 710-719.

[12] Wei, B. C. (1998). Exponential Family Nonlinear Models. Singapore: Springer-Verlag Singapore..

[13] Yu, Q. Q. and Wong, G. Y. C. (2002) Asymptotic Properties of a modified Semi-parametric MLE in linear Regression with Right-censored Data, *Acta Mathematics Sinica*, 18, 405-416.

[14] Zong, xu-ping, et. al. (2000), Confidence on nonlinear models with random effects, *Applied Mathematics*,Vol. 13, N0. 4.,100-105.

[15] Zong Xu-ping, et. al. (2002) Estimation in linear models with random effects and errors-in-variables, *Ann. Inst. Stat. Math.*, Vol. 54, NO. **3**, pp595-606.

Zong xuping: Associated Professor of School of Mathematics, Yangzhou University, is the common fellow of Association of Jiangsu Probability and Statistics, and he is the fellow of Chinese Association of Probability and Statistics. He graduated from Southeast University in 1987; He obtained PHD in 2000. He was the Visiting Scholar of HongKong University (1998.3~1998.9), the visiting Professor of Yale University (2003-2004). He has published a book of Probability and Statistics, which was selected to be the textbook by Chinese National Ministry. His research interests are in Statistical nonlinear model, statistical diagnostics and statistical computing.

A BACK PROPAGATION NEURAL NETWORK PARALLEL ALGORITHM FOR INVERSE HEAT CONDUCTION PROBLEMS

Lin Jiansheng, Guo Qingping, Qi Jingjing
Department of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei 430063.P.R.China
Email: ljs_forest@sina.com

ABSTRACT

Inverse heat conduction problem (IHCP) is a nonlinear, ill-posed problem, requiring massive amount of computation, which is difficult to be described and solved properly with existing mathematical theories. This paper creatively introduces neural networks into the solution, and proposes a new BP neural network (BPNN) algorithm. Using a temperature field as an input vector, it outputs the coefficients of heat conduction. The new algorithm succeeds to solve this problem.

In order to solve the computing problem of requiring huge amounts of computation when training, this paper also raises a BP neural network parallel algorithm. The parallel algorithm adopts the model of data parallelisms, to allocate the training sample set for every node of the cluster on average; every node on the cluster has an entire copy of the BPNN. The parallel program mode is a master-slave mode based on MPI. The new parallel algorithm greatly increases the speed of training.

Keywords: IHCP, BPNN, NN, Neural Network, Parallel, MPI.

1. INVERSE HEAT CONDUCTION PROBLEM

Inverse heat conduction problem is a branch of the inverse problem's research, which is an inverse process of the forward heat conduction problem. We call the problem, whose aim is to compute the coefficient of heat conduction by researching the time-varying law of the object's interval or boundary temperature value, an inverse heat conduction problem (IHCP).

IHCP is a nonlinear, ill-posed problem, requiring massive computing amount, which makes the solution to IHCP very complicated and difficult. In the real engineering domain, we can barely gain the analytical solutions. Turning to numerical computing for help is the only feasible method. In the research paper, Jonas and Abdullah hold the view that computational amount of IHCP far exceeds the forward problem, no matter the inverse problem is linear inverse or nonlinear.

1.1 Description of Heat Conduction Problem

The IHCP researched in this paper, is based on the solving of coefficient of heat conduction in the internal combustion engine's heat-barrier material, which is a mixture of ceramic and metal. The coefficient of heat conduction of this kind of material is a multi-dimensional, nonlinear, time-varying function.

Under the rectangular coordinate, the temperature of the material can be expressed as follows:

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}(k\frac{\partial u}{\partial x})+\frac{\partial}{\partial y}(k\frac{\partial u}{\partial y}) \tag{1.1}$$

where, u=u(x,y,t) represents the temperature at time t, coordinate point (x,y),and k=k(x,y,t), represents the coefficient of heat conduction of the material.

Besides, boundary conditions are content with the first boundary conditions. We can get the mathematical model of two-dimensional heat condition problem as follows:

$$\begin{cases}\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}(k\frac{\partial u}{\partial x})+\frac{\partial}{\partial y}(k\frac{\partial u}{\partial y}) \\ u(x,y,0)=0 \\ u(0,y,t)=\sin(2\pi t)\sin(\pi y) \\ u(1,y,t)=0 \\ u(x,0,t)=0 \\ u(x,1,t)=0 \\ 0<x,y<1,t>0\end{cases} \tag{1.2}$$

When the coefficient of heat conduction k=k(x,y,t) changes by temperature, Eq. (1.2) is a nonlinear heat conduction problem. In the engineering domain, the coefficient of heat conduction is generally regarded as a quadratic function of temperature u. And in the ceramic-metal material we researches, the coefficient of heat conduction is also a function of x. So the coefficient of heat conduction can be represented as:

$$k=k(x,u)=(1-x^2)(a+bu+cu^2)+x^2(d+eu+fu) \tag{1.3}$$

where, a, b, c, d, e and f are parameters associated with the material's properties. Obviously, the coefficient of heat conduction k=k(x,y,t) is uniquely determined by parameter a, b, c, d, e and f.

1.2 Solution to Heat Conduction Problem

The solution to two-dimensional heat conduction problem, i.e. mathematical model (1.3) is under the premise of knowing the value of parameter a, b, c, d, e and f. Assigning a group of value to parameter a, b, c, d, e and f, for example, setting a = 0.1, b = -0.01, c = 0.001, d = 1.0, e = 0.1, f = 0.01, we can get $k=(1-x^2)(0.1-0.01u+0.001u^2)+x^2(1.0+0.1+0.01u)$. Applying finite difference method to model 1.2, we can get difference schema below:

$$(4+\rho)\, u_{ij}^{(n+1)} = u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n+1)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n+1)} + \mathrm{dij} \tag{1.4}$$

The solution to difference schema (1.5) is also a big problem. Fortunately, Professor Guo Qingping has studied out a multi-grid parallel algorithm [1][2], which has solved it properly. When the coefficient of heat conduction is given, applying this algorithm, we can compute the temperature u=u(x,y,t) at grid point u(x,y,t). The algorithm studied by Professor Guo, is a basis of this paper.

1.3 Description of Inverse Heat Conduction Problem

Opposite to heat conduction problem, IHCP is to compute the coefficient of heat conduction, i.e. to compute the value of parameter a, b, c, d, e and f, under the premise of knowing the temperature of grid point by experiment or other ways.

Unfortunately, different from forward problems, which have gained desirable mathematical theory supports and solution, IHCP remain a nonlinear, ill-posed problem, which can't be described and solved properly by the existing mathematical theories. Although, there have existed Tikhonov regularization [3], conjugate gradient method, Gauss-Newton iterative method [4], and some other methods based on these methods above [5]. These methods more or less have problems like:

(1) they are only aimed at one specified, simple problem,

without broad scope of application.
(2) their steps are very complicated, requiring abstruse theory knowledge.
(3) they lack obvious physical meanings.

So, raising a new algorithm with higher efficiency, and broader scope of application, is extremely urgent. In recent years, artificial neural network, which develops rapidly in the domain of artificial intelligence and image processing, owns strong ability of learning and adaptability. These features of artificial neural network offer new inspiration to solve IHCP.

2. NEURAL NETWORK

2.1 Introduction to Neural Network

Artificial neural network (ANN) is a kind of information processing system, which imitates the structure of human's brain. It is composed of massive, simple nonlinear processing unit, which is connected on another, according to human brain's structure. ANN has the features of distributed storage, parallel processing, benign adaptability, self-organization, fault tolerant, strong ability of memorization, etc.

After several years' development, now, several kinds of network structure have been raised, including Perception, Hopfield network, ART network and BP network, etc. These neural networks can be divided into three kinds: forward network, feedback network, and self-organizing network. This thesis will adopt forward neural network, which has the ability of approximating any nonlinear function infinitely.

2.2 Approximating Ability of Forward Neural Network

A forward neural network with n input units and m output units can perform a mapping from an n-dimensions space R^n to an m-dimensions space R^m. According to the original theorem proposed by Kolmogorov: For all $n \ge 2$, and for any continuous real function g of n variables in the domain [0,], $g:[0,1]^n \to R$, there exist 2n + 1 continuous, monotonously increasing functions of one variable in [0,1], by which g can be reconstructed according to the following equation:

$$g(x_1,x_2,\dots,x_n)=\sum_{j}^{2n+1}\psi_j\Big(\sum_{i=1}^{n}\phi_{ij}(x_i)\Big) \tag{2.2}$$

where $x_i \in [0,1]$, and ψ_j is continuous functions of one variable.

Many researchers attempt to improve this theorem since its original formulation, We can replace $\phi_{ij}(x_i)$ with $b_i\phi_j(x_i)$, where b_i is a constant. Therefore, Eq.(2.2) can be rewritten as

$$g(x_1,x_2,\dots,x_n)=\sum_{j}^{2n+1}\psi_j\Big(\sum_{i=1}^{n}b_i\phi_j(x_i)\Big) \tag{2.3}$$

Eq.(2.3) corresponds to a three-layered forward network architecture with a single output. Based on the theoretical foundation above the following theorem can be obtained:
Let $\phi(x)$ be a non-constant, bounded and monotonously increasing continuous function. There exist an integer k and sets of real constants α_i, θ_i and w_{ij}, where (i = 1, 2, . . . ,k), (j = 1,2,. . . ,n), such that the expression:

$$\hat{g}(x_1,x_2,\dots,x_n)=\sum_{j}^{m}\alpha_i\phi\Big(\sum_{i=1}^{n}w_{ji}x_i-\theta_i\Big)$$

can be defined to meet $\left|g(x_1,x_2,\dots,x_n)-\hat{g}(x_1,x_2,\dots,x_n)\right|<\varepsilon$

where ε is an arbitrary little real number greater than 0. From the theoretical description[6][7] above, for the condition $\varepsilon > 0$, there exists a three-layered forward network, in which the activation function of the hidden layer is $\varphi(x)$, w_{ji} is the neuron's weight, the output layer is linear. The theorem presented above provides the foundation for applying the BPN network to the solution of IHCP.

3. CONSTRUCTION OF BP NEURAL NETWORK ALGORITHM FOR IHCP

3.1 Construction of BP Neural Network

According to Eq.(1.3), we know the truth: if the value of parameter vector (a,b,c,d,e,f) is given, i.e. (a^i,b^i,c^i,d^i,e^i,f^i), the coefficient of heat conduction is also uniquely determined. Then, adopting the numerical solution studied by Professor Guo, based on difference schema (1.4), we can compute the temperature u^i(x,y,t) at grid point (x,y,t). Using a certain rule, we can mark all the grid point as u^i_1， u^i_2， ...， u^i_n. Vector (u^i_1， u^i_2， ...， u^i_n) is called temperature field U^i. Obviously, the mapping from the coefficient k^i of heat conduction to temperature field U^i is a one-to-one mapping, which can be expressed by a BPNN. So, we construct a BPNN as Fig.1:

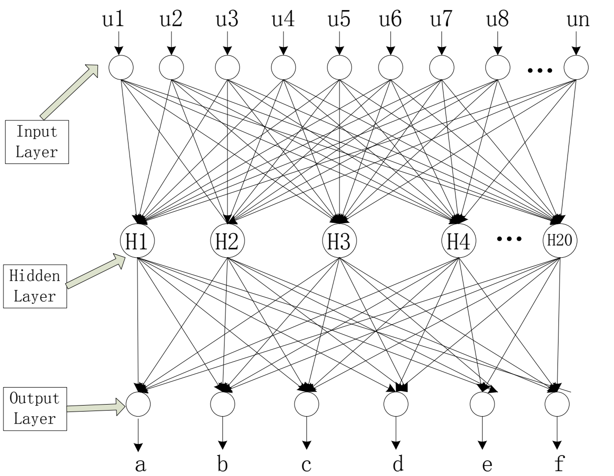


Figure 1. Neural network for IHCP

The BPNN has three layers, i.e. input layer, hidden layer and output layer. The input data is temperature field U, and $U=(u_1,u_2,\dots,u_n)$, which means that U is make up of the temperature u(x,y,t) at all the grid point (x,y,t). So the number of neutron at input layer equals to the number of grid point. Every neutron in the input layer corresponds to the temperature at a grid point u_i. If we divide the interval into 100, 100, 10 grid points separately at x, y and t axis, then we have 10^5 grid points, which means that the number of neutron at input layer is 10^5, and the input data of the neutron in the input layer is u_i. The output data is the coefficient of heat conduction k. According to Eq.(1.3), k is determined by 6 parameters, i.e. a, b, c, d, e, f. So the number of neutron in output layer equals to the number of parameter in Eq.(1.3), i.e. 6. Every neutron in output layer corresponds to one of the 6 parameters, and the output data of a neutron in output layer is the computing value of the corresponding parameter.
Then, we should determine the number of hidden layer and the number of the neutron in every hidden layer. Unfortunately, there are no certain rules to comply. Usually, it's adequate to solve the problem using only one hidden layer. With too many hidden layer, the algorithm will be too complicated. So, we set the number of hidden layer as 1, and the number of neutron in the hidden layer is set as 20 initially.

We choose sigmoid function

$$f(x) = \frac{1}{1+e^{-x}}$$

as the activation function. Sigmoid function has benign characters:

(1) The value range of x is $(-\infty,+\infty)$, and the value range of the function is [0,1].

(2) The function is differentiable, and its derived function is $f'(x)=f(x)(1-f(x))$, which is easily computed by the value of the function.

3.2 Construction of Training Sample Set

In order to train the BPNN constructed before, we need to construct the training sample set. Generally, there two methods to construct the training sample: (1) by experiment, (2) by simulative computation to the heat conduction problem.

Toward to the BPNN we constructed, the number of the neutron in the input layer reaches $O(10^5)$. In order to train the network more effectively, the sample size should reach $O(10^5)$ at least. It's impossible to get the enormous training sample by experiment. So we will obtain the training sample set by simulative computation to the heat conduction problem.

From the analysis in 1.1, we know that the solution of the heat conduction problem is to compute the temperature field, when the coefficient of heat conduction is given. Let $(a,b,c,d,e,f)=(a^i,b^i,c^i,d^i,e^i,f^i)$, where$(a^i,b^i,c^i,d^i,e^i,f^i)$ are the values given, the coefficient of heat conduction shown in Eq.(1.4) determined also, then, adopting the numerical solution studied by Professor Guo, we can compute the temperature field $u=(u_1,u_2,\ldots u_n)$. If we continuously give a group of different value to (a,b,c,d,e,f), we will obtain different temperature field U. Every temperature field U corresponds to a group of different value (a,b,c,d,e,f), we mark them as $\{U^i,k^i\}$. Thus, we can constitute the training sample set D :

$$\{U^1,k^1\},\ \{U^2,k^2\},\ \ldots,\ \{U^p,k^p\},$$

where p is the sample set size, $\{U^i,k^i\}$ is a example of the train sample set D.

3.3 BP Neural Network Algorithm

In order to describe conveniently, we define the weight matrix W^1 in the hidden layer as:

$$W^1 = \begin{bmatrix} w^1_{1,1} & w^1_{1,2} & \cdots & w^1_{1,n} \\ w^1_{2,1} & w^1_{2,2} & \cdots & w^1_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ w^1_{20,1} & w^1_{20,2} & \cdots & w^1_{20,n} \end{bmatrix}$$

the weight matrix W^M in the output layer as:

$$W^M = \begin{bmatrix} w^M_{1,1} & w^M_{1,2} & \cdots & w^M_{1,20} \\ w^M_{2,1} & w^M_{2,2} & \cdots & w^M_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ w^M_{m,1} & w^M_{m,2} & \cdots & w^M_{m,20} \end{bmatrix}$$

θ^1_j as the threshold in the hidden layer, θ^M_j as the threshold in output layer.

Given a sample d: {U,k}, where, $u=(u_1,u_2,\ldots u_n)$ is the input vector, and k=(a,b,c,d,e,f) is the target vector, after the network's computation, we can get the output vector (a',b',c',d',e',f'). Out of expressing conveniently, we define $(t_1,t_2,t_3,t_4,t_5,t_6)=(a,b,c,d,e,f)$ as the target vector, and $(y_1,y_2,y_3,y_4,y_5,y_6)= (a',b',c',d',e',f')$ as the output vector. After the computation of the network, we can get the output in the hidden layer as follows:

$$o^1_j = \sum_{i=1}^{n} w^1_{j,i} u_i \tag{3.1}$$

and get the output in the output layer as follows:

$$y_j = \sum_{i=1}^{6} w^M_{ji} o^1_i \tag{3.2}$$

We define the error function as follows:

$$E_d = \frac{1}{2}\sum_{i=1}^{6}(t_i - y_i)^2 \tag{3.3}$$

Then, in order to modify the weight and threshold, we use gradient descent method to back propagate the error E_d.

(1) In the output layer M, the error δ^M_j of jth neutron is

$$\delta^M_j = y_j(1-y_j)(t_j - y_j) \tag{3.4}$$

So, towards sample d, according to gradient descent method, the training rule of weight in the output layer is

$$\Delta w^M_{j,i}(d) = \eta \delta^M_j o^1_i \tag{3.5}$$

and the training rule of threshold in the output layer is

$$\Delta\theta^M_j(d) = -\eta\delta^M_j \tag{3.6}$$

(2) In the hidden layer, the error δ^1_j and the error δ^M_j in the output layer meet as follows:

$$\delta^1_j = -o^1_j(1-o^1_j)\sum_{l=1}^{6}\delta^M_l w^M_{l,j} \tag{3.7}$$

According to Eq.(3.7), given the error in the output layer, we can compute the error of every neutron in the hidden layer. So, towards sample d, according to gradient descent method, the training rule of weight in the hidden layer is

$$\Delta w^1_{j,i}(d) = \eta\delta^1_j u_i \tag{3.8}$$

and the training rule of threshold in the output layer is

$$\Delta\theta^1_j(d) = -\eta\delta^1_j \tag{3.9}$$

After having trained all the samples in the training sample set D, we can compute the total weight increment $\Delta w^k_{j,i}$ and total threshold increment $\Delta\theta^k_j$ as follows:

$$\Delta w^k_{j,i} = \sum_{d\in D}\Delta w^k_{j,i}(d) \tag{3.10}$$

$$\Delta\theta^k_j = \sum_{d\in D}\Delta\theta^k_j(d) \tag{3.11}$$

where k equals to 1 or M.

And we can compute the total error E as follows:

$$E = \sum_{d\in D} E_d \tag{3.12}$$

Then, we can refresh the weight $w^k_{j,i}$ and threshold θ^k_j in the output layer and hidden layer as follows:

$$w^k_{j,i} \leftarrow w^k_{j,i} + \Delta w^k_{j,i} \tag{3.13}$$

$$\theta^k_{j,i} \leftarrow \theta^k_{j,i} + \Delta\theta^k_{j,i} \tag{3.14}$$

where k equals to 1 or M.

From above, we conclude the BP neural network algorithm to solve IHCP as follows:

Step 1: Create a BPNN with n neutrons in the input layer, 20 neutrons in the hidden layer, and 6 neutrons in the output layer. Set the sigmoid function as the activation function. Initialize weights and thresholds with some little random numbers between -0.05 and 0.05, define a variable named num and initialize it as num=1, which records the number of train.

Step 2: Given a new sample d: {U,k}, where, $u=(u_1,u_2,\ldots u_n)$, k=(a,b,c,d,e,f), input the input vector$(u_1,u_2,\ldots u_n)$.

Compute the outputs in the hidden layer and output layer using Eq.(3.1), Eq.(3.2) separately, the error E_d using Eq.(3.3)

Step 3: Back propagate the error to compute the error produced by the neutrons in the output layer and hidden layer using Eq.(3.4), Eq(3.7) separately.

Step 4: Compute weight and threshold increments corresponding to sample d in the output layer using Eq.(3.5) and Eq.(3.6), in the hidden layer using Eq.(3.8) and Eq.(3.9).

Step 5: Judgment: if all the samples in the training sample set D have been trained, compute the total error E using Eq.(3.12)
compute total weight increments and total threshold increments in the output layer and hidden layer using Eq.(3.10) and Eq.(3.11) separately,
refresh weights and thresholds in the output layer and hidden layer using Eq.(3.13) and Eq.(3.14) separately,
refresh the variable num as num= num+1, go to the next step,
otherwise, go to step 2.

Step 6: Judgment: if the total error E is less than a given precision, which means that the train succeeds, output weights and thresholds, and terminate the program; if the variable num is more than a given max number N, which means that the train failed, terminate the program, otherwise go to step 2 to train continually.

4. BP NEURAL NETWORK PARALLEL ALGORITHM BASED ON CLUSTER

Because of the complexity of IHCP, the number of neutrons in the BPNN we constructed reaches $O(10^5)$, and the sample size should reach $O(10^5)$ at least. It requires huge amounts of computation to train the BPNN. It's impossible for a personal computer to undertake it. So, we should parallelize the algorithm, and turn to the PC cluster for help.

4.1 Neural Network Parallel Algorithm Base on Cluster

The PC cluster is a group of personal computers connected by the network, which offers some certain service for the users as an entity. A personal computer in the PC cluster is called a node. The parallel algorithm in this thesis is based MPI, which is configured on the PC cluster.

In order to reduce the communication, balance the load of every node, improve the performance of the parallel algorithm, we adopt the way of data parallelisms. We allocate the training sample set for every node on average, and every node have an entire copy of the neural network algorithm.

The parallel program mode is a master-slave mode. When the program runs, it will generate one master process and several slave processes. The master process is responsible for initializing weights and thresholds, allocating computing tasks, training the network, gathering errors, weight and threshold increments sent by slave processes, broadcast new weights and thresholds. Slave processes are in charge of training the network, sending errors, weight and threshold increments to the master process, receiving new weights ad thresholds from the master process. Obviously, communication only exits between the master process and slave processes, and there is no communication between one slave process and another slave process.

Suppose there are n personal computers in the PC cluster, i.e. $node_0, node_1, \ldots, node_{N-1}$. Using $node_0$ as master node, $node_1, \ldots, node_{N-1}$ as slave nodes, we divide the training sample set into N subsets, i.e.$D_0, D_1, \ldots, D_{N-1}$, and subset D_i is allocated for $node_i$ (i=0，1,…,N-1). We can design the neural network parallel algorithm as follows.

4.2 Algorithm for the Master Node

Step 1: Create a BPNN with n neutrons in the input layer, 20 neutrons in the hidden layer, and 6 neutrons in the output layer. Set the sigmoidal function as the activation function. Read subset D_0 from local memory. Initialize weights and thresholds with some little random numbers between -0.05 and 0.05, define a variable named num and initialize it as num=1, which records the number of train, set a terminated variable called STOP and initialize it as STOP=0, which means that it needs to train continually.

Step 2: Broadcast the value of STOP to slave nodes. Judgment: if STOP=0 is true, broadcast weights and thresholds to slave nodes, go to the next step; if STOP=1 is true, which means that the train is completed, terminate the program.

Step 3: Train the network on subset D_0. For every sample d in subset D_0, compute error E_d using E.q(3.3) , and back propagate it to compute the sample's weight increments and threshold increments using E.q(3.5), E.q(3.6), E.q(3.8), E.q(3.9) separately.
After having all the samples in subset D_0, compute the local total weight increment

$$\Delta w_{j,i}^k(node_0) = \sum_{d \in D_0} \Delta w_{j,i}^k(node_0)$$

compute the local total threshold increment

$$\Delta \theta_j^k(node_0) = \sum_{d \in D_0} \Delta \theta_j^k(d)$$

and compute the local total error

$$E(node_0) = \sum_{d \in D_0} E_d \ .$$

Step 4: Gather local total errors, local total weight increments and local total threshold increments on every slave node.
Compute the total error as

$$E = \sum_{i=node_0}^{node_{N-1}} E(p)$$

compute the total weight increment as

$$\Delta w_{j,i}^k = \sum_{p=node_0}^{node_{N-1}} \Delta w_{j,i}^k(p)$$

compute the total threshold increment as

$$\Delta \theta_j^k = \sum_{p=node_0}^{node_{N-1}} \Delta \theta_j^k(p) \, .$$

Then, refresh the variable num as num= num+1, weights and thresholds using Eq.(3.13) and Eq.(3.14).

Step 5: Judgment: if the total error E is less than a given precision, which means that the train succeeds, set variable STOP as STOP=1; if the variable num is more than a given max number N, which means that the train failed, set variable STOP as STOP=1. Go to step 2.

4.3 Algorithm for the Slave Node

Suppose this slave node is $node_i$. In order to reduce unnecessary communication, before training the network, the subset D_i has been stored on this node computer. It needn't to receive the subset when training.

Step 1: Create a BPNN with n neutrons in the input layer, 20 neutrons in the hidden layer, and 6 neutrons in the output layer. Set the sigmoidal function as the activation function. Read subset D_i from local memory.

Step 2: Wait and receive the value of variable STOP from the master node. Judgment: if STOP=0 is true, receive weights and thresholds from the master node, go to the next step; if STOP=1 is true, which means that the train is completed, terminate the program.

Step 3: Train the network on subset D_i. For every sample d in subset D_i, compute error E_d using E.q(3.3) , and back propagate it to compute weight and threshold increments using E.q(3.5), E.q(3.6), E.q(3.8), E.q(3.9) separately.
After having all the samples in subset D_i, compute the local total weight increment
$\Delta w_{j,i}^k(node_i) = \sum_{d \in D_i} \Delta w_{j,i}^k(d)$,
compute the local total threshold increment
$\Delta \theta_j^k(node_i) = \sum_{d \in D_i} \Delta \theta_j^k(d)$,
and compute the local total error
$E(node_i) = \sum_{d \in D_i} E_d$.
Step 4: Send the local total errors, local total weight increments and the local total threshold increments to the master node.
Go to step 2.

5. CONCLUSIONS

Inverse heat conduction problem (IHCP) is a nonlinear, ill-posed problem, requiring massive amount of computation. The problem is difficult to be described and solved properly with existing mathematical theories. The existing methods like Tikhonov regularization, conjugate gradient method and Gauss-Newton iterative method are commonly limited in a certain area . It is extremely urgent to develop a new algorithm with higher efficiency and broader scope of application.

This paper creatively introduces neural networks into the solution, and proposes a new BP neural network (BPNN) algorithm. Using a temperature field as an input vector, it outputs the coefficient of heat conduction. The training of it requires a training sample set with massive samples. This paper creatively gains the training sample set by the method of computing the corresponding heat conduction problem, when by the method of experiment impossible.

In order to solve the computing problem of requiring huge amounts of computation when training, this paper also raises a BP neural network parallel algorithm. The parallel algorithm adopts the model of data parallelisms, to allocate the training sample set for every node of the cluster on average; every node on the cluster has an entire copy of the BPNN. The parallel program mode is a master-slave mode based on MPI. The new parallel algorithm greatly increases the speed of training.

REFERENCES

[1] Guo Qingping, Y. Paker et al. “Optimum Tactics of Parallel Multi-grid Algorithm with Virtual Boundary Forecast Method Running on a Local Network with the PVM Platform”. Journal of Computer Science and Technology, Vol.15, No.4, July 2000：355-359.

[2] Guo Qingping，Y. Paker et al. Parallel Multi-grid Algorithm with Virtual Boundary Forecast Domain Decomposition Method for Solving Non-linear Heat Transfer Equation. Lecture Notes in Computer Science , High Performance Computing and Networking, Spreinge Press, May 2000, Vol . 1823：568-571.

[3] Engl,Heinz W, Zou Jun. New approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction. Inverse Problems. v16 ,n6, Dec 2000: 1907-1923.

[4] Xie Jianli, Zou Jun. An improved model function method for choosing regularization parameters in linear inverse problems. Inverse Problems, V18, N3, June, 2002: 631-643.

[5] Liu Hongyu, Zou Jun. Iterative choices of regularization parameters in linear inverse problems. Inverse Problems, v14, n5, Oct 1998: 1247-1264.

[6] Cybenko, G. Continuous valued neural networks with two hidden layers are sufficient (Technical Report). Department of Computer Science, Tufts University, Medford, MA, 1988.

[7] Cybenko, G. Aporoximatation by superposition of a sigmoidal function. Mathematics of Control, Signals, and System, 1989，2：303-314.

[8] Tom M. Mitchell. Machine Learning, simplified Chinese translation edition. Beijing: China Machine Press, 2003：60-90.

Lin Jiansheng is a master candidate of school of computer science and technology. He had received the B.S degree in Information and Computation Science from Chongqing Institute of Technology, Chongqing, China, in 2007. Now he goes on further study in his tutor Qingping Guo’s parallel processing and computing lab, and focuses his study on parallel computing and numerical computing.

AN AUTOMATIC PORTAL GENERATION ALGORITHM FOR BSP TREE

Wu Chengwei , Guo Qingping
Distributed Parallel Processing Lab, Wuhan University of Technology
Wuhan, Hubei 430063, China
Email: lizhipaoniu@163.com

ABSTRACT

Automatic portal generation (APG) for the scene based on BSP Tree is always a knotty problem. This paper firstly gives an introduction of the BSP + Portal + PVS technology, and analyzes the purpose and function of APG; then, points out the disadvantages of the previous methods for APG. Finally, the paper proposes a new method by means of twice coherent cutting processes using recursive traversal to implement APG. The result of the experiment shows that the algorithm can avoid redundant portals effectively, and ensure the indices of the two leaves connected by the portal are correct and valid, providing a guarantee for the PVS analysis followed.

Keywords: BSP, Automatic portal generation, PVS.

1. INTRODUCTION

In the computer simulation, the scene may be very large and complicated. If you render every graphic unit of it, it will inevitably waste massive computer resources. The optimal strategy is to render the units which are visible to user only, and remove the invisible ones. In many indoor games, the integrated approach which combines BSP, Portal and PVS, is adopted to achieve the aim.

BSP[1] (binary space partitioning) is a very mature technology for space division. Its ultimate goal is to divide the whole scene into convex sets (we call it "cell"). Ensuring that each set is convex can bring us many advantages when we deal with object location and collision detection. Because there is no special ambiguity in convex body in comparison with concave one.

Portal, as a matter of fact, is a polygonal entry which connects two rooms, and is invisible in the scene. The basic idea with a portal engine is that when you render a scene from a viewer's position with a viewing frustum and encounter a portal polygon, the portal clips the viewing frustum. Then the adjacent sector is rendered from the same viewer's position but with the new viewing frustum. Many objects of the scene can easily be culled away since the viewing frustum is limited only to be exactly what you see.

A portal engine has a very flexible structure that provides some nice features, but there are some problems with it, especially when all the clipping occurs when you are drawing the scene. A similar but much better solution is calculating what needs to be drawn in the pre-rendering of the map. Although the granularity of PVS is a bit rough, it pays to do in complicated situation since it can sharply reduce the consumption of CPU resources. For each leaf of BSP scene, a Potential Visible Set (PVS) is created. This PVS is the set of leaves that is visible from the first leaf. To calculate the PVS we need to do standard ray tracing between the leaves, to see if any point in a leaf is visible from another one.

2. PRESENT SITUATION

The BSP algorithm is quite simple, the key point of which depends on the strategy to choose the optimal dividing plane. And the strategy can have a direct impact on the balance of the final BSP tree, causing an ineffective ray tracing. Potential Visible Set (PVS) process starts from the cell adjacent to the current one, including the same portal with the former one. And then the whole scene is traversed. The process can be easily done by DFS.As to whether cell A and B can be visible to each other, we can construct sampling points on all the portals of A and B, and carry out a race tracing.

The difficulty of BSP + Portal + PVS technology lies in the automatic generation of portal. According to the description of PVS technology, we can infer that our objective is acquiring all the portal polygons and the indices of two cells connected by each portal.

Andreas Brinck, a staff of the DICE company, firstly came up with a solution to the problem, detailed in the paper named *Binary Space Partioning Trees and Polygon Removal in Real Time 3D Rendering*[2]. But unfortunately, there are some bugs and disadvantages in the method, such as redundant portals, lack of cell indices, among which the most fatal one is that irrational recursions may happen during runtime.

In 2003, Huling and Li Zhenxiao improved the algorithm proposed by Andreas Brinck, and posted an article named *Automatic Portal Generation*[3] on GameDev to describe there new idea. Nevertheless, the improved algorithm still causes plenty of redundant portals, and can't ensure that the indices are correct either. In the article *How to remove redundant portals*[4], Saya described the drawbacks of this algorithm, and gave a way to remove redundant portals.

Recent years, Alon Lerner[5] and Denis Haumont [6] put forward AGP algorithms based on half-edge traversal and characteristic of distance field respectively. These two solutions are omitted from consideration since they don't aim to BSP scenes.

3. ARITHMETICAL STATEMENTS

According to the qualities of BSP tree, we can simply derive qualities of portal as follows,

1. Each portal must be on the dividing plane, that is to say, they are generated in a nonleaf node and get their sizes constrained in the space range of the subdomain of that node;
2. Each portal may finally be divided into several ones;
3. Each portal will finally be clipped by the graphic units of cell, which are usually triangles in order to ensure that every portal locates within the space of the cell it subordinate, especially the portal, which is out of the cell, will be clipped to vacuum;
4. Each portal should connect two cells which located on different sides of the division surface where the portal is.

In the following, the drawbacks of previous algorithms will be analyzed, and new solution will be put forward.

3.1 Drawbacks of the Previous Algorithms and Causes

Both Andreas's and Hu's methods implement portal generation in one single traversal. Here, we take Andreas's method as an example to analyze the drawbacks and causes of this kind of solution in Figure 1.

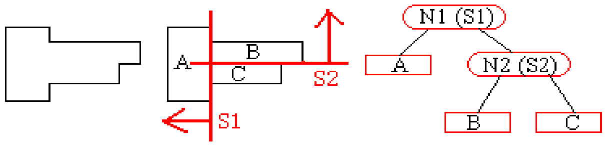


Figure 1. Scene 1

The process of portal generation is as follows:

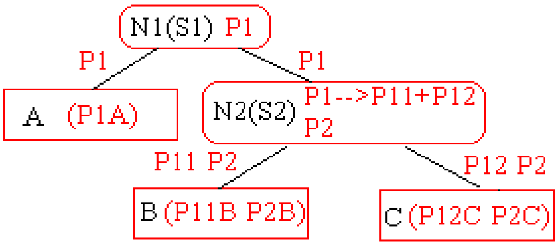


Figure 2. Process of Portal Generation for Scene 1

1) Generate new portal P1 on plane S1;
2) Put P1 into both N1's positive and negative subtrees;
3) P1 enters into A, and gets cut into P1A;
4) P1 enters into N2, and gets cut into P11 and P12 by S2 as P1 spans S2;
5) Generate new portal P2 on plane S2;
6) P11 and P2 enter into B, and get cut into P11B, P2B; separately;
7) P12 and P2 enter into C, and get cut into P12C, P2C separately.

Finally we get five portals: P1A, P11B, P12C, P2B and P2C. However, a careful observation will reveal P1A and P2C are redundant. The reason for redundance of P1A is plane S2 cuts it into two parts, while P2C is redundant since it is a duplicate of P2B.Moreover, the cutting processes to portals in the positive and negative subtrees are separate, as a result, we cannot get the connection information of portals well.

Hu's method improves on Andreas's, and can avoid some redundant portals such as P2C by means of passing addresses, and get the connection information of portal well. However, redundant portals such as P1A still exist , which can only be eliminated in the extra removing process.

The root cause to cause these drawbacks lies in the cutting processes in both subtrees are independent,Lacking coherence.

3.2 Our Solution

We can infer from the 4th quality that, supposing that certain portal is generated from a dividing plane of a inner node P, we can put the portal in the positive or negative subordinate tree to clip recursively. As it goes usually, we assume that it is put in the positive one.If the negative one of node P is not taken into consideration, the portal that is clipped down will be the ultimate one and the index of the last cell should be that of the cell on the positive side of the Portal. Then, if we now put the Portal above into the negative subordinate tree of Node P to clip it recursively, the final Portal will be final no matter which subordinate tree is compared to. The following is the psuedocode of the algorithm.

```
struct BSPNode {
        Long          index;
        BSPNode*    pPosit, pNegat;
}
struct Portal {
        Long        positIndex, negatIndex;
        Polygon     poly;
        BSPNode *   pGenNode;
}
```

```
GeneratePortals(BSPNode* pNode, Portal ptl[])
    if pNode is not a leaf;
        Portal pPosit[] = {}, pNegat[] = {};
        for each portal P of ptl[];
            calculate which side is P in relative to
             the dividing plane of pNode;
            if is in the positive side, push P into pPosit;
            if is in the negative side,push P into pNegat;
            if is spanning the dividing plane;
              Split P into positPart and negatPart
              push positPart into pPosit
              push negatPart into pNegat
        create a new portal newP on dividing plane, and
         clip it with all the ancestor nodes of pNode;
        newP.pGenNode = pNode;
        push newP into pPosit;
        GeneratePortals(pNode->pPosit, vPosit);
        GeneratePortals(pNode->pNegat, vNegat);
    if pNode is a leaf
        for each portal P of ptl[]
            clip p with all graphic units of pNode;
        if P is reasonable
            P.positIndex = pNode.index
            RefinePortal(P, P.pNodeGen);
```

```
RefinePortal(Portal P, BSPNode* pNode)
    if pNode is not a leaf
        calculate which side is P in relative to
        the dividing plane of pNode;
        if is in the positive side
            RefinePortal(P, pNode->pPosit);
        if is in the negative side
            RefinePortal(P, pNode->pNegat);
        if is spanning the dividing plane
            Split P into positPart and negatPart
            RefinePortal(positPart, pNode->pPosit)
            RefinePortal(negatPart, pNode->pNegat);
    if pNode is a leaf
        clip p with all graphic units of pNode
        if P is reasonable
            P.negatIndex = pNode.index
            Save P;
```

When we call GeneratePortal function for the first time, in other words, while the parameter pNode is equal to the root node of the BSP tree, the parameter ptl[] is an empty set.

3.3 Algorithm Running Flow

The common configurations of the rooms in indoor scene are parallel, staggered and some special ones. To describe easily, here define P(A,B) which means that portal P connects cell A (positive side) and cell B(negative side). If either A or B equals NIL, this means it has not be assigned.

a. Parallel Scene

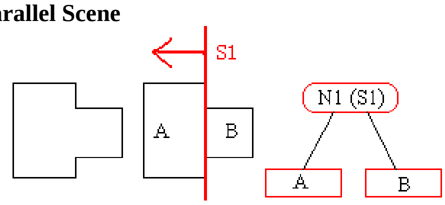


Figure 3. Parallel Scene

1) Generate P1(NIL,NIL) on plane S1;
2) Put P1(NIL,NIL) into the positive subtree of N1;
3) P1(NIL,NIL) enters into A,and gets cut into P1(A,NIL);
4) Put P1(A, NIL) into the negative subtree of N1;
5) P1(A,NIL) enters into B, gets cut into P1(A,B);

The final portal generated is P1(A,B).

b. Staggered Scene

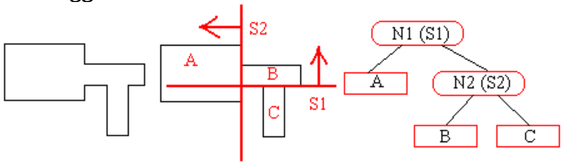


Figure 4. Staggered Scene

1) Generate P1(NIL,NIL) on plane S1;
2) Put P1(NIL,NIL) into the positive subtree of N1;
3) P1(NIL,NIL) enters into A, and gets cut into P1(A,NIL);
4) Put P1(A,NIL) into the negative subtree of N1;
5) P1(A, NIL) enters into N2;
6) P1(A,NIL) enters into B, gets cut into P1(A, B);
7) Generate P2(NIL, NIL) on plane S2;
8) Put P2(NIL, NIL) into the positive subtree of N2;
9) P2(NIL,NI) enters into B, and gets cut into P2(B,NIL);
10) Put P2(B,NIL) into the negative subtree of N2;
11) P2(B,NIL) enters into C, and gets cut into P2(B,C);

The final portals generated are P1(A,B), P2(B,C).

c. Special Scene

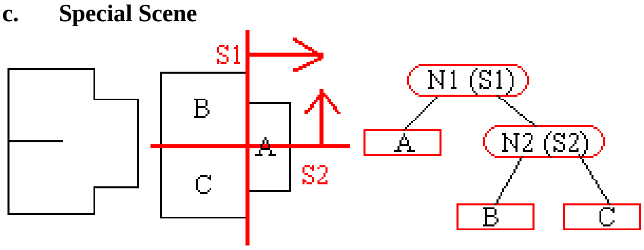


Figure 5. Special Scene

1) Generate P1(NIL,NIL)on plane S1;
2) Put P1(NIL,NIL)into the positive subtree of N1;
3) P1(NIL,NIL)enters into A,and gets cut into P1(A,NIL);
4) Put P1(A, NIL) into the negative subtree of N1;
5) P1(A, NIL) enters into N2, and gets cut into P11(A,NIL) and P12(A,NIL) by S2 since P1 spans S2;
6) P11(A,NIL) enters into B, gets cut into P11(A,B);
7) P12(A,NIL) enters into C, gets cut into P12(A,C);
8) Generate P2(NIL, NIL)on plane S2;
9) Put P2(NIL,NIL) into the positive subtree of N2;
10) P2(NIL,NIL) enters into B, and gets cut into P2(B,NIL);
11) Put P2(B,NIL) into the negative subtree of N2;
12) P2(B,NIL) enters into C, and gets cut into P2(B,C);

The final portals generated are P11(A,B), P12(A,C), P2(B,C).

4. COMPLEXITY ANALYSIS

There are many unpredictable factors that affect function GeneratePortals, such as the number of the graphic units in the scene, noted as Nu, the number of the cells and the depth of BSP tree, noted as Nc and D separately. In addition, the probability that portal is clipped by inner nodes and cells into vacuum and the number of ultimate portals generated noted as Np, plus the former factors are all unpredictable ones that have no absolute functional relation with each other. And therefore, we can not use a specific expression to identify the complexity of function Generate Portals, but to get an approximate formula.

As Portal is generated from inner nodes, we can suppose that Np and Nc are of the same order of magnitude. As for therelation between Nc and D, D can be seen as lgNc approximately, then equal to lgNp. From the pseudocode, we can know that if function RefinePortal and the clipping to Portal in a cell is out of consideration, the approximate complexity of function GeneratePortals will be

$$T(Np) = 2 * T(Np/2) + Np \qquad (1)$$

This is equal to that of merge sort, and therefore $T(Np) \in O(Np*lgNp)$.

The recursive process of function RefinePortal is similar to Binary Search,its complexity being lgD, equal to lgNp. The times that RefinePortal is called equals Np, the number of portal, and therefore, the whole complexity of calling function RefinePortal is O(Np*lgNp).

As it goes like this, the whole complexity of cells clipping portal is equal to the production of the number of cell, and the number of portal, and the average number of graphic units each cell, therefore,the whole clipping complexity

$$T = O\ (Np*lgNp+Np*lgNp+Nc*Np*Nu/Nc) \qquad (2)$$

As Nu>>lgNp, T = O(Np*Nu).

5. PERFORMANCE

In the testing stage,we have got the hardware platform as follows:

CPU : Intel(R) Pentium(R) 4 CPU 2.40GHz
Memory : 1GB
VGA Card : NVIDIA GeForce2 MX400.

The choosen maps for testing contains 1250 triangles in total, and every one thousand times' test takes 3.8~3.9s.

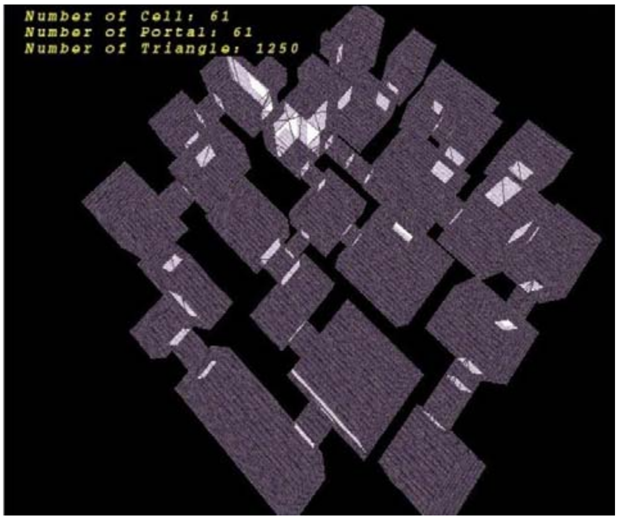


Figure 6. Automatic Portal Generation
(The white polygons are portals)

Table 1. The Connect Information of Portals for Figure 6.
("Ptl Idx" represents the index of portal,"Pos Idx" and "Neg Idx" represent the indices of positive and negative cells connected by the portal separately)

Ptl Idx	Pos Idx	Neg Idx	Ptl Idx	Pos Idx	Neg Idx	Ptl Idx	Pos Idx	Neg Idx
0	0	51	21	18	23	42	38	39
1	0	2	22	19	20	43	40	42
2	1	6	23	21	26	44	40	41
3	1	2	24	22	23	45	41	47
4	3	6	25	23	58	46	43	48

5	3	4	26	24	26	47	43	44
6	5	6	27	24	25	48	44	45
7	7	10	28	27	31	49	46	48
8	7	8	29	27	29	50	46	47
9	9	12	30	28	45	51	49	52
10	9	10	31	28	29	52	49	50
11	11	15	32	30	33	53	51	53
12	11	12	33	30	31	54	52	53
13	13	54	34	32	36	55	54	56
14	13	15	35	32	33	56	55	60
15	14	22	36	34	36	57	55	56
16	14	15	37	34	35	58	57	60
17	16	17	38	35	38	59	57	58
18	17	36	39	37	59	60	59	60
19	18	35	40	37	38			
20	18	58	41	38	42			

6. CONCLUSIONS

Automatic portal generation for the scene based on BSP Tree is always a thorny issue. A new method by means of twice coherent cutting processes using recursive traversal is brought forth in this paper to avoid redundant portals and ensure the indices of the two leaves connected by the portal are valid, providing a guarantee for the PVS analysis followed.

Based on the chance that the portal may be dug by coplanar scene graphic units, it's probably that the final portal is concave. In such a case, you can carry out a convex decomposition[7] in order to facilitate the analysis of PVS. The best solution is to avoid protruding pixel region while modeling.

REFERENCES

[1] Schumacker Robert A, Brand Brigitta, Gilliland Maurice G. Study for Applying Computer-Generated Images to Visual Simulation. AFHRL-TR-69-14, Air Force Human Resources Laboratory, 1969.

[2] Samuel Ranta-Eskola. Binary Space Partioning Trees and Polygon Removal in Real Time 3D Rendering[D]. Uppsala University, 2001.

[3] Huling. Automatic Portal Generation[EB/OL]. http://www.gamedev.net/reference/articles/article1891.asp, 2009.

[4] Saya3d. How to remove the redundant portals[EB/OL]. http://dev.gameres.com/Program/Visual/3D/saya3dProtal.pdf, 2009.

[5] Alon Lerner. Breaking the Walls: Scene Partitioning and Portal Creation[C]. In Proceedings of the 11th Pacific Conference on Computer Graphics and Applications, 2003, pp.303~312.

[6] Denis Haumont, Olivier Debeir. Volumetric cell-and-portal generation[J]. Computer Graphics Forum, Vol.22, No.3, Sept. 2003, pp.303~312.

[7] Philip Schneider, David H. Eberly. Geometric Tools for Computer Graphics[M]. Publishing House of Electronics Industry, 2005, pp.565~589.

Wu Chengwei is a master candidate of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology, focuses his study on parallel computing and game engine.

OPTIMIZING FUZZY CONTROLLER OF ENGINEERING SHIP WITH GENETIC ALGORITHM*

MEI Jun-jin[1,2], ZHANG Jian[2], LI Zhi-jun[2], SHEN Ke-yu[2]
[1] School of Computer,Huangshi Institute of Technology, Huangshi, 435003,China
[2] School of Automation, Wuhan University of Technology, Wuhan, 430063,China
Email: cims-hs2004@163.com; junwenlee@sohu.com

ABSTRACT

Adjusting the length of anchor chain controls the course of engineering ship. It limits the moving scope of the ship and makes it not drift about wantonly. But while moving and mooring the ship, because wind, wave, flow and various kinds of random influence is difficult to be predicted, it is difficult to control of the course. So fuzzy controller is used to improve the performance of control in this thesis, and genetic algorithm is adopted to optimize the rule of fuzzy controller, which enables the controller to produce the better result in case of no priori knowledge. Finally, the mathematics model of the ship is set up and simulation is carried on to the system.

Keywords: Engineering Ship, Fuzzy Control, Genetic Algorithm, Optimize Rules.

1. INTRODUCTION

The geotextiles-laying vessel in Changjiang River is designed for the special demand of the Changjiang River channel when renovate the soft foundation, such as being sandy, and riverbed construction. It guarantees the project of renovating channel goes on smoothly. Figure 1 is the picture of the actual project.

Figure 1. Picture of Actual Project

The equipment of the ship includes orienting system, moving and locating system etc. The moving and locating system is used to make the ship's moving in a limited scope and guarantee the ship will not drift about wantonly with wind, wave, flowing. Because the ship may be affected by wind, wave and flow that come from any direction, it needs 6 anchors to be assigned in all directions (the form of radiation) in the course of orienting. As Figure 2 shows.

The moving and locating system relocates the ship through adjusting the length of anchor cables. The head (T5) and tail (T6) anchor winches are in middle of the ship's head and tail separately. By adjusting the length of the head and tail anchor cables, the total effort of longitudinal flow can be counteracted. Four anchor winches in the larboard and starboard are responsible for overcoming ship's offset on either side. During the process of moving control, the differential GPS surveys ship's track data and transmits it to monitor platform, according to the offset of survey track and the expect track, the monitor platform makes one or more anchor winches change the length of cable to counteract the offset.

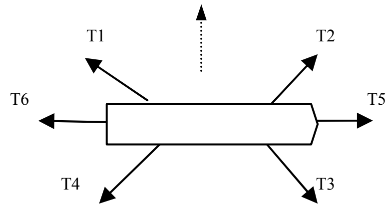


Figure 2. Mooring System

During the process of moving and locating ship, it has some difficulties to control the course because various kinds of random influence, such as wind, wave and flow are difficult to be predicted. In this case, every anchor winch is standby and ready to adjust the length of the cables. Once the offset is detected, the anchor winches begin to move. Then the ship is in the dynamic equilibrium. Since acting force of ship is complicated and coordination among pulling force and angle of six anchor's cables difficult, a lot of randomness factors will inevitability be neglect and system's control performance will be influenced if a fixed model is adopted. However, adopt the fuzzy control system can realize the control easily.

2. FUZZY CONTROL SYSTEM DESIGN

2.1 Input and Output Design

Error signal, detected by GPS, and calculated change of error is system's input signal, that is: $X_m=[e,ec]^T$, $e=[-15，15]$，$ec=[-0.8,0.8]$

The speed of six anchor winches receiving or putting cables is System's output signal, $U=[u_1, u_2, u_3, u_4, u_5, u_6]$. $u=[-2, 2]$

The input and output are divided into 7 fuzzy grades separately, that is NB (Negative Big), NM (Negative Middle), NS (Negative Small), ZE (Zero), PS (Positive Small), PM (Positive Middle), PB (Positive Big).

2.2 Selection of Membership Function (MF)

When select the membership function, if the error is large a fuzzy set of the low resolution will be adopted, and otherwise if the error is small a fuzzy set of the high resolution will be

* Supported by Nation Natural Science Foundation of China(69874018,79970025).

adopted. When $x_i = [x_i^-, x_i^+]$ (i=1,2,…,n) and the domain of variable x_i is divided into N_i section, gausses MF is adopt as:

$$\mu_A(x) = e^{-\frac{(x_i - x_0{}^j)^2}{2\sigma^{j2}}} \quad (1)$$

Where $\sigma^j = (1-1/4)(x_i^+ - x_i^-)/N_i$, and $x_0^j = x_i^- + (x_i^+ - x_i^-) * j/N_i$, $j=0,1,2,\cdots,N_i$. σ influences the form of membership function curve directly, and the different form of MF curve can cause different control characteristics. Since the domain of variable x_i is divided into N_i section, there will be N_i+1fuzzy set in the domain of variable x_i.

2.3 Generating Fuzzy Rules

In a system that has two input signal and six output signal which can also be called as MIMO system, rules forms as follow: R =(R^1_{MIMO}, R^2_{MIMO}, …, R^n_{MIMO}), where R^l_{MIMO}: If x_1 is F^l_1 and x_2 is F^l_2, Then $y_1 = \theta_{l1}$, $y_2 = \theta_{l2}, \cdots y_{l6} = \theta_{l6}$, R^l_{MIMO} belong to fuzzy set X *X, and the consequent is the combining of n control signals, which are independent. So R^l_{MIMO} can be taken as rules of n MISOs, that is R^l_{MIMO}={R^{l1}_{MIMO}, R^{l2}_{MIMO}, …, R^{ln}_{MIMO}},where R^{lj}_{MIMO}:If x_1 is F^l_1 and x_2 is F^l_2, then $y_j = \theta^l_j$, so the system is simplified as:

R^l: If x_1 is F^l_1, and x_2 is F^l_2, then $y_i = \theta^l_i$

Since [e, ec] have been divided into seven fuzzy set, the total number of rules is 7×7=49.

2.4 Defuzzifier

Take the MIN-MAX-Center[2] method to defuzzyfy the output, as follow

$$y_{jo} = \frac{\sum_{i=1}^{n} u_i(y_i) \cdot y_i}{\sum_{i=1}^{n} u_i(y_i)} \quad (2)$$

3. FUZZY CONTROL SYSTEM DESIGN

In recent years, evolution-based knowledge gained a great deal of popularity due to its inherent ability in efficient and parallel search of complex and multi-modal landscapes. Application of genetic algorithms (GA) to optimization of fuzzy logic control systems can automatically generate the consequent part of fuzzy rules without prior knowledge, and produce better rules to reach the satisfactory result. The systematic block diagram shows as Figure 3[9].

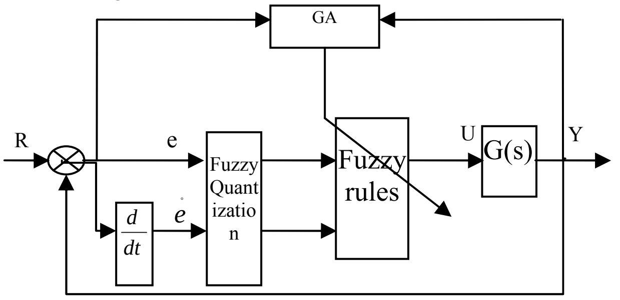


Figure 3. GA optimize Fuzzy Controller

3.1 Coding

Generally, when GA is applied to the continuous system that is multi-dimension and high accuracy, performance will be relatively bad if the binary code is used as genes. Consider the character and application of fuzzy controller, the real code has higher quality than the binary code. The fuzzy rules which adopt the decimal code to express the chromosomes show as follows: NB is represented by 1; NM is represented by 2; NS is represented by 3; ZO is represented by 4; PS is represented by 5; PM is represented by 7, and the 0 represents the rule does not exit. Since forty-nine rules have been got, then forty-nine decimal numbers are needed to represent them. Then an initial chromosome which can represent the whole rules is form as a number of forty-nine decimal numbers. The initial N populations are generated at random, that means N rules set are produced. After selection, crossover, mutation of the original population which is made up of N individual, the individual will be got, that is the optimum rules set[6].

3.2 Selection of Fitness Function

Fitness Function is used to evaluate the quality of solution while GA is in the evolving and searching progress. The expect output of the system is R, and the actually output is Y. Define the evaluation function [5,6] as:

$$J = \sum_{t=1}^{n} [R - Y(i)]^2 \quad (3)$$

where T is sample time; n is member of sample.

After evaluation function of the system is confirmed, the fitness function can be defined. The purpose of GA optimization is to adjust the control parameter and minimize the value of evaluation function J. Since the values of fitness function that are compared and ordered is the basis of calculating the selection probability, it should not be the negative and must be increased as the quality of the solutions is improved. Therefore it is necessary to change evaluation function to a form of finding max value and value is negative. The value of fitness function is got as:

$$f = \frac{P}{1 + J} \quad (4)$$

where P can be any positive number.

3.3 Selecting Parent Individual

The classic roulette wheel selection [6] is a method that select individual by the value of fitness. Whatever the fitness of individual is, the individual has the possibility to be selected. Thus the fine gene is preserved. The individual is select by the way as fellow:

$$selection \quad prbability \le \sum_{i=0}^{i} individual \quad fitness[i] / sum \quad of \quad individual \quad fitness \quad (5)$$

3.4 Crossover

Stepped crossover [3] is used to realize the crossover operation. The stepped crossover decide which parent's gene will be inherited by the new generation at the corresponding position by setting the mask word. The mask word is generated at random and it is made up of 0 and 1, keeping the same structure with the individual. The operation of stepped crossover is as follow: assume that the two parent individual is known as Tl and T2, and mask word at ij is known as M_{ij}. if M_{ij}=0, corresponding gene at ij does not cross, namely $Tl'_{ij} = Tl_{ij}$, $T2'_{ij} = T2'_{ij}$, on the contrary, if M_{ij} =1, crossover is happened in corresponding gene, namely $T1'_{ij} = T2_{ij}$, $T2'_{ij} = T1_{ij}$.

3.5 Mutation

The mutation gene is chosen as: evenly generate a random number r to very individual at the rage of 0 and 1. When r<p_m(mutation probability), the mutation is happened. Then the gene of X^i_r individual at the position of p will mutate. Since the code of fuzzy rule must be a integer, the mutation is defined as:

$$\varphi = (X^i_t, P) = round[X^i_t(P) + C(-0.5, 0.5)] \quad (6)$$

Where C(-0.5,0.5) is a evenly distributed noise at the range of

-0.5 and 0.5. Limiting the change of mutation value at [- 0.5,0.5] make the max change of rule in an integer, that is to say the max variation amount can only be +1 or -1. The function *round(x)* is used to get the nearest integer of *x*. The mutation method said reasonable control rule can only make small change near the initial point. Too big change will cause unreasonable rules to be produced, so the value of mutation can only be +1 or -1.

4. THE MATHEMATICAL MODEL OF THE CONTROLLED PLANT

4.1 Mathematical Model of Vessels Movement

$$\begin{cases} (m+m_x)\dot{u}-(m+m_y)vr = X_H + X + T_X \\ (m+m_y)\dot{v}+(m+mx)ur = Y_H + Y + T_Y \\ (I_{ZZ}+J_{ZZ})\dot{r} = N_H + N + T_N \end{cases} \tag{7}$$

Where X_H is the force of dynamic fluid in X direction; Y_H is the force of dynamic fluid in Y direction; N_H is the moment of force of dynamic fluid; X is the force of wind, wave and flow in X direction; Y is the force of wind, wave and flow in Y direction; N is the moment of force of wind, wave and flow of Z axis; T_X is the force of hawser in X direction; T_Y: is the force of hawser in Y direction; T_N is the moment of force of hawser of Z axis; m, m_x are m_y are the mass of the ship and adding mass in X and Y direction. I_{zz}、J_{zz} are moment of inertia and adding moment of inertia of the ship to Z axis. u, v, r are the transverse, longitudinal speed and winding z axle's rotating angular speed of Z axis.

4.2 Calculating the Force of Hawser

Form Figure 4, the vector relation can be got as follow[8]:

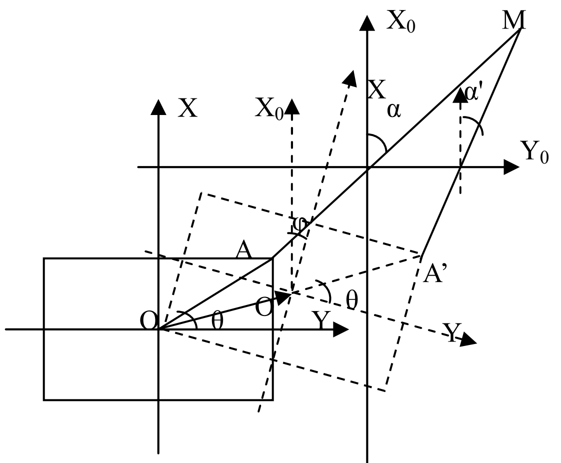


Figure 4. Analysis of one Cable's Force

$$\overrightarrow{OO'}+\overrightarrow{O'A'}-\overrightarrow{OA}=\overrightarrow{MA'}-\overrightarrow{MA} \tag{8}$$

From analyzing one cable's force to the ship, following equations can be got:

$$\begin{cases} T_x = T_H \cos(\varphi-\alpha) \\ T_y = T_H \sin(\varphi-\alpha) \\ T_n = T_H a\cdot\sin(\varphi-\alpha) \end{cases} \tag{9}$$

Where T_H is determined by the length of the anchor cable and the horizon distance of hawse (A) and anchor (M).

Thus the resultant force of six anchor cables is

$$\begin{cases} T_X = \sum_{i=1}^{6} T_{xi} \\ T_Y = \sum_{i=1}^{6} T_{yi} \\ T_N = \sum_{i=1}^{6} T_{ni} \end{cases} \tag{10}$$

5. SIMULATION

Simplify the output of the fuzzy controller as the speed of four anchor winches when carry on simulation, namely v1, v2, v3 and v4, and assume v1, v2, v3 and v4 have the following mathematics relation: $v_1=-v_4$, $v_2=-v_3$, $v_1=kv_2$. Thus the system is simplified as an two-input and one-output system. While simulating, the sample time is set 0.5s, that is to say every 0.5s a command is sand to the controller. Get thirty sets input data in study, and use them to calculate the individual fitness at each generation. Assume the length of the ship is L=152m, wide of ship is B=25m, and the depth of draft is d=1.6m. After optimizing the fuzzy controller's rules offline, the simulation curve is got as Figure 5.

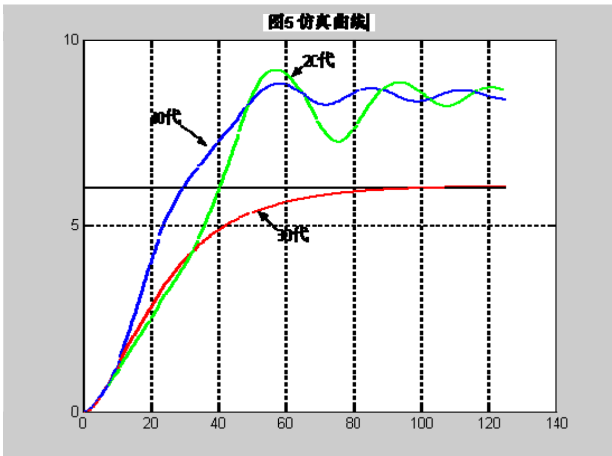


Figure 5. Simulation Curve

It is obvious that applying GA to optimize fuzzy controller's rule let the rules of the fuzzy controller can be set at random initially and can be find out as optimum rules (or rules that inferior to optimum rules) after automatically update (or evolution), let the system find out the optimum rules form according to the goal function of the control system established, make rules form set up more scenically and make improvement of performance of fuzzy controller.

REFERENCES

[1] Wang Xiaoping. Genetic Algorithms ——Theory、Applications and software Implementation[M]. Xi'an: Xian Jiaotong University Press, 2002.

[2] Li Shiyong. Fuzzy Control • Neuron control and Intelligent Cybernetics[M]. Harbin: Harbin University of Technology Press, 1998

[3] Zhang Yuan, Li Zhijun. Using genetic algorithm to

optimize fuzzy control rules in workship′s controller [J]. Ship & Ocean Engineering, 2005, (2):28~31
[4] Shi Yinghai etc. Design of Fuzzy Controller Based on Genetic Algorithms，Journal of Shenyang Institute of Chemical Technolgy. Vol.15 No4 P273~278
[5] Haitao Liu,Sizong Guo. Research on Fuzzy Number Ranking Methods in Management Decision Making Analysis[J]. Advances in Systems Science and Applications(2007),Vol.7, No.1:110~116
[6] Ma Zhuang etc. Research on the Autopilot Fuzzy Control Based on the Genetic Algorithm[J]. Navigation Of China, SerialNo.45 P94~99
[7] Tao Huang,Jianping Fan,Cheng Wang. Fuzzy-pid Control Research on Vehicle Active Suspension Based on the Nonlinear Model[J]. Advances in Systems Science and Applications, 2006, 6(4):667~674
[8] Guo Xiaohao. Development and Construction of the Geotextiles-Laying Vessel in Changjiang River[J]. Ship & Ocean Engineering. 2003 , (1):1~6
[9] Mianyun Chen, Zhijun Li, Long Zhou etc. A Control Approach to “Poor” Information System. Advances in Systems Science and Applications(2001),Vol.1, No.2:152~158

MEI Jun-jin was born in 1978, graduated from Wuhan University of Technology, Wuhan City, China, in 1999 and received his M.S. degree in computer science from Wuhan University of Technology,China, in 2006.He is currently a lecturer in the Institute for Computer ,Huangshi Institute of Technology , Huangshi, China . He has published ten books, over 21 Journal papers, His research interests include the computer application、structural system design and computer vision.

A NOVAL AUDIO CLASSIFICATION ALGORITHM BASED ON GA AND SVM WITH COMBINED KERNEL FUNCTION

LI Jing, Wan Juan, Zhang Yunlu
School of Computer Science, Wuhan University, Wuhan, Hubei, 430079, China
Email: leejingcn@msn.com

ABSTRACT

Audio classification is an important access to extract audio structure and content, and is a premise for audio content analysis, retrieval and further treatment. Support Vector Machine (SVM) is a valid statistic learning method. In this paper, learning algorithm of SVM is introduced to construct classifier, and construct a new kernel function and use Genetic Algorithm (GA) to optimize the parameters of classifier model. This paper proposes a new audio classification algorithm, GA-CBSVM based on GA and SVM with combined kernel function to classify speech, music and their mixed audio. The experimental results show that GA-CBSVM is excellent for audio classification and the average of classification accuracy is up to 93.08%.

Keywords: Audio classification, Support vector machine, Combined kernel function, Genetic algorithm

1. INTRODUCTION

Audio information plays an important role in multimedia data. Because the original audio signal is merely a binary flow of non-semantic and unstructured, in addition to contain some limited information such as sampling frequency, quantitative accuracy and encoding method, so it lacks description of the content semantics and structural organization. Audio classification is an important means of extracting structural information and content semantics from audio signal, and is the key in depth processing, analysis and content-based retrieval of audio[1].

Audio classification method based on statistics learning algorithm is focal point and will be main research directions in the present and future. SVM is an algorithm based on statistics learning, and a classification method with the principle of structural risk minimization. Due to its better generalization ability, SVM can achieve better accuracy in the case of small samples[2]. Therefore, this paper introduces the learning algorithm of SVM as the theoretical basis for constructing classifier.

Because the kernel function of SVM has important impact on classification performance, this paper constructs a new combined kernel function with good performance, and the SVM by using this new kernel function have both good learning ability and good generation ability.

At the same time, the parameters of SVM classifier have also great impact on classification accuracy, and there is no uniform criterion for selection. So this paper proposes a parameters optimization algorithm based on genetic algorithm, GA-CBSVM, taking full advantage of global search capability of genetic algorithm, and combining GA and SVM with combined kernel function. This algorithm is used to classify speech, music and their mixed audio.

2. AUDIO FEATURE ANALYSIS AND EXTRACTION

The selection and extraction of audio feature is base of analyzing audio information, the selected features should capture the audio frequency-domain and time-domain traits of different audio classes, and is of robust and general to environment.

Prior to feature extraction, all audio signals are firstly converted into a general format of wav, 8-KHz, 16-bit, and mono-channel. Secondly audio signals are pre-emphasized with parameter 0.97, thirdly it is segmented into non-overlapping audio clips of one second long, and then each clip is further divided into frames by hamming-window. The frames are 256 samples each, with 50% overlap in each of the two adjacent frames.

Two types of features are computed from each frame: (i) perceptual features, composed of Zero Crossing Rate, Sub-band energy ratio, Frequency Centroid, Bandwidth, and (ii) Mel-frequency cepstrum coefficients (MFCCs).

After extracting features from each frame, calculate statistical characteristics of all frames in one clip as the clip features, including High ZCR Ratio(HZCRR)[12], Low frequency energy Ratio(LFER)[12], Sub-band energy ratio, Frequency Centroid, Bandwidth, 12 order MFCCs.

At last the paper introduces standard deviation to construct feature set.

Firstly Sub-band energy ratio, Frequency Centroid, Bandwidth, HZCRR, LFER these features are concatenated to form a 8-dimensional vector, and then the vector is normalized according to

$$x_i' = \frac{x_i - \mu_i}{\delta_i} \tag{2-1}$$

where the mean μ_i and standard deviation δ_i are calculated over all the training data, named "Perc". 12 order MFCCs form a 12-dimensional vector, named "Ceps".

The FeaturePerc and FeatureCeps sets are weighted and then concatenated into another a 20-dimensional vector, named "PercCeps".

PercCeps=[Perc/s1,Ceps/s2] (2-2)

where s1 equals 8, s2 is calculated by $s2 = \sum_{i=1}^{12} \delta_i$, where δ_i is the i-th standard deviation of MFCCs.

So each clip is expressed as a 20-demensional feature vector.

3. THE CLASSIFICATION BASED ON GA AND SVM WITH COMBINATED KERNEL FUNCTION

3.1 Basic Theory of Support Vector Machine

Given a set of training vectors belonging to two separate classes, $(x_1, y_1), K, (x_i, y_i)$, where $x_i \in R^n$ and $y_{i \in \{-1,1\}}$, one wants to find a hyper-lane $wx + b = 0$ to separate the data which maximizes the margin(the distance between the hyper-plane and the nearest data point of each class).

The solution optimal separating problem of SVM is given by the saddle point of the Lagrange functional,

$$L(\omega, b, \alpha) = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{l}\alpha_i\{y_i[(\omega \cdot x_i) + b] - 1\} \tag{3-1}$$

with Lagrange multipliers α_i. The solution is given by,

$$\bar{\omega} = \sum_{i=1}^{l}\bar{\alpha}_i y_i x_i, \qquad \bar{b} = -\frac{1}{2}\bar{\omega}\cdot[x_r + x_s] \tag{3-2}$$

where x_r and x_s are support vectors which belongs to class +1 and -1, respectively.
In linearly non-separate but nonlinearly separate case, the SVM replaces the inner product $x \cdot y$ by a kernel function K(x, y), and then constructs an optimal separating hyper-plane in the mapped space. Possible choices of kernel functions include:

(1) Polynomial Function

$$K(x \cdot y) = ((x \cdot y) + 1)^d \tag{3-3}$$

where, the parameter d is the degree of the Polynomial.

(2) Gaussian Radial Basis Function:

$$K(x \cdot y) = \exp\left(-\frac{|x - y|^2}{s^2}\right) \tag{3-4}$$

Where, the parameter s is the width of the Gaussian Function.

(3) Multi-Layer perception function:

$$K(x \cdot y) = \tanh(a(x \cdot y) + b) \tag{3-5}$$

Where, a and b are the scale and offset parameters.

3.2 Combined Kernel Function

Different kernel function confirms different non-linear transform and feature space, the selection of which is the key to classification accuracy.

Each kernel has both advantages and disadvantages. At present the kernel function of SVM is divided into two categories: global kernel function and local kernel function. Global kernel function possesses good generalization ability, and local kernel function possesses good learning ability.

Considering the global traits of Polynomial function and local traits of Gaussian Radial Basis function, this paper constructs a new function with good performance by linearly combining the above two functions. SVM that adopts this new combined kernel function——CBSVM have both good generalization ability and good learning ability. The formula of the new function is as follows:

$$K_{new}(x, x') = rK_{poly}(x, x') + (1 - r)K_{rbf}(x, x'), 0 < r < 1 \tag{3-6}$$

where

$K_{poly}(x, x')$ is polynomial function

$K_{rbf}(x, x')$ is Gaussian Radial Basis function.

So in this paper, the kernel function of SVM adopts the new combined kernel function.

3.3 A Parameters Optimization Algorithm Based on Genetic Algorithm

The parameters of SVM model have no unified selection criteria and theory, but these parameters have an important impact on classification accuracy of SVM. The paper proposes a parameters optimization algorithm based on genetic algorithm, named "GA-CBSVM", making full use of global search capability of GA, in order to get the best model of the optimal parameters.

The steps of Algorithm are as follows:

(1) Randomly generate a group of SVM parameters, and encode them by binary- encode scheme to form initial population.

(2) Each individual in the population is decoded to obtain the value of parameters, and then the audio feature set of training sample are used to train SVM, with which forecast testing sample to obtain recognition rate of testing sample data. Note down the largest recognition rate of the current population. The formula of recognition rate is as follows:

$$RR = \frac{\text{the number of samples that correctly classified}}{\text{the total number of testing samples}} \tag{3-7}$$

(3) Implementing genetic manipulation on the population, including selection, crossover and mutation to generate next population.

(4) Set the maximum number of iteration for 1000. Repeating the above processes and comparing the largest RR of the current population to the largest RR of all generations, in order to update the best individual. Until up to the maximum number of iterative.

The parameters optimization algorithm based on GA is as follows.

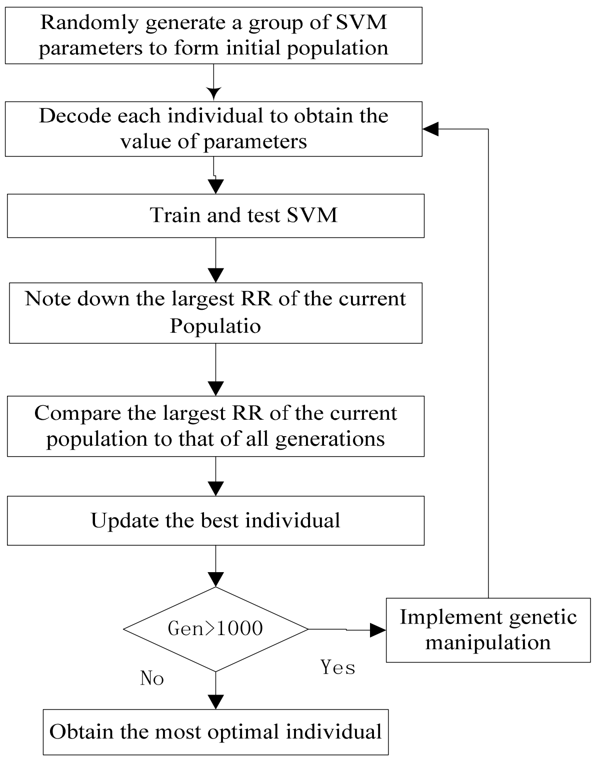


Figure1. Parameters Optimization Algorithm Based On GA

3.4 Classification of Speech, Music and Their Mixed Audio

After extracting features from audio, we need two classifiers based on GA-CBSVM to classify these three categories of audio type. First, the audio clip is classified into speech and non-speech classes. Where, non-speech means pure music. Then, speech is further classified into pure speech and speech over music. The flow of classification is shown in Figure 2.

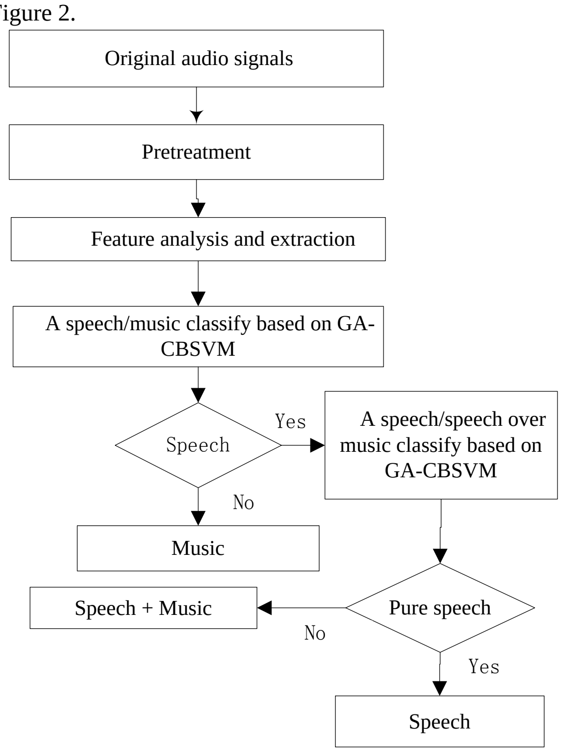


Figure 2. The Flow of Classification

4. EXPERIMENTAL RESULTS AND ANALYSIS

The data used in our experiments is collected from music CDs and Chinese reading documents. Firstly the initial audio signal is converted into a general format of wav, 8-KHz, 16-bit, and mono-channel by audio editing tools. Audio data have 3485 seconds in total length, in which pure music is 1484s, pure speech is 1055s and their mixed clip is 946s.

Table 1. Experimental results

	Recognition Rate		
	HMM	SVM	GA-CBSVM
music/speech	91.98%	93.70%	94.64%
speech/music + speech	84.52%	90.00%	91.53%
Average	88.25%	91.85%	93.08%

The classification accuracy obtained from classified based on GA-CBSVM is compared to that by HMM and SVM[10][12]. The results are shown that the classification accuracy improved to a certain extent by the classifier based on GA-CBSVM, and higher than 4.83 % and 1.23 % respectively by comparing with HMM and SVM.

5. CONCLUSIONS

Audio classification is a research hotspot of the current content-based analysis of the audio, and has an important application significance on audio retrieval, video-assisted analysis and so on. This paper proposes an algorithm based on GA-CBSVM for classification of an audio clip , the proposed approach classified audio clips into pure music, pure speech and their mixed clips. The experimental results have shown that the algorithm based on GA-CBSVM has good classification performance, and the average of accuracy is up to 93.08%.

REFERENCES

[1] Li Hengfeng, Li Guohui. Content-Based audio retrieval and classification, Computer Engineering and Application, 2000, (7):54~56.

[2] Lie Lu, Hong-Jiang Zhang, Stan Li, Content-based Audio Classification and Segmentation by Using Support Vector Machines, ACM Multimedia Systems Journal 8 (6), 2003: 482~492.

[3] Wenjuan Pan,Yong Yao, Audio classification in a weighted SVM, Communications and Information Technologies, 2007. ISCIT '07. International Symposium on17-19 Oct. 2007 Page(s):468 – 472.

[4] Shantha Selva Kumari, R., Sugumar, D. Audio Signal Classification Based on Optimal Wavelet and Support Vector Machine, Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on Volume 2, 13-15 Dec. 2007 Page(s):544 – 548.

[5] Xiao-Li Li,Zhen-Long Du, Kernel-Based Audio Classification, Machine Learning and Cybernetics, 2006 International Conference on Aug. 2006 Page(s):3313 - 3316 Kernel-Based Audio Classification.

[6] Guodong Guo, Li, S.Z. Content-based audio classification and retrieval by support vector machines, Neural Networks, IEEE Transactions on Volume 14, Issue 1, Jan. 2003 Page(s):209 – 215.

[7] Menaka R, Lonce W. Generic Audio Classification Using a Hybrid Model Based on GMMs and HMMs, In: Proceedings of the 11th International Conference of Multimedia Modelling, 2005. 12-14 Jan, 2005: 53~58..

[8] G. Martens, Hans De Meyer, Bernard De Baets, Marc Leman, Tree-based versus distance-based key recognition in musical audio. Soft Compute. 2005, 9(8): 565~574.

[9] Bai Liang, HuYali, Feature analysis and extraction for audio automatic classification. IEEE International Conference on Volume 1, 10-12 Oct. 2005:767~772.

[10] Lu Jian, Chen Yisong, Sun Zhengxing, Audio classification based on HMM Model, Journal of Software, 2002, 13(8):1593~1597.

[11] Zhao Li, Voice signal processing, the press of Machinery Industry, 2003.

[12] Bai Liang Audio Classification and Segmentation technology, National Defense University of Science and Technology, 2004.12.

RESEARCH OF UNCERTAIN REASONING MODEL BASED ON BAYESIAN ONTOLOGY

Wenhui Xiao, Huifu Zhang, Xuelan Di
School of Computer Science and Engineering, Hunan University of Science and Technology
Xiangtan, Hunan 411201, China
Email: xwh-0223@163.com

ABSTRACT

In the current, ontology is widely applied in information systems, including fault diagnosis area. It has a strong advantage to solve semantic interoperability, knowledge share, knowledge reuse. In the large and complex fault diagnosis system, there are complex relations in the inner components. Between the uncertain factors and information congest, the complex forms may be multi-fault, relevant fault and so on. But most of ontology forms are based on classic logic that is not to represent and reason uncertain and incomplete data. Bayesian Network based on probability and graph theory is put forward for resolving the uncertain problems. By combining ontology with Bayesian Network, The thesis researches uncertain knowledge's representation and reasoning process based on Bayesian Ontology, and constructs an uncertain reasoning model of rotary mechanical fault diagnosis.

Keywords: Ontology, Bayesian Network, Bayesian Ontology, Uncertainty, Fault Diagnosis

1. INTRODUCTION

In the field of equipment fault diagnosis, there are the majority of uncertain problems, because most of them are decided by the complexity of diagnosis object, the limitation of test methods and the inaccuracy of knowledge. For such uncertain information, we need to select an appropriate way to express them, in order to find out fault cause more quickly and exactly.

Ontology, as a model tool, can describe knowledge model on semantic and knowledge level, provide a formal description of concept, and lay a strong foundation for knowledge share [1]. But the major shortcoming of ontology is its inability to represent and reason uncertain and incomplete data. In other words, ontology could neither represent overlaps or intersection degree among concepts, nor support the reasoning of concepts or individual partial information. Bayesian Network based on probability reasoning is put forward for solving uncertainty and imperfection problem. It has a strong advantage about solving fault caused by complex equipment's uncertainty and relevance [2]. This paper adopts OWL to make probability extend so that it can include probability information to ontology. By combining Bayesian Network with ontology, it can meet the need for knowledge representation and reasoning of uncertainty.

2. KNOWLEDGE REPREAENTATION AND REASONING OF UNCERTAINTY

Initial attempts to represent uncertainty in ontology languages tend to begin with constructs for attaching probabilities as attributes of entities. This approach is clearly inadequate, in that it fails to account for structural features such as conditional dependence (or independence), double counting of influence on multiply connected graphs, and context-specific independence

2.1 Probabilistic Ontology

Probabilistic ontology expands standard ontology by providing logically sound constructs for representing statistical regularities and probabilistic interrelationships in a domain of application [3, 4]. Costa gave a formal definition of probabilistic ontology [5]. A probabilistic ontology is an explicit, formal knowledge representation that expresses knowledge about a domain of application. This includes: (i) Types of entities that exist in the domain; (ii) Properties of those entities; (iii) Relationships among entities; (iv) Processes and events that happen with those entities; (v) Statistical regularities that characterize the domain; (vi) Inconclusive, ambiguous, incomplete, unreliable, and dissonant knowledge related to entities of the domain; and (vii) Uncertainty about all the above forms of knowledge. In this definition, the term entity refers to any concept (real or fictitious, concrete or abstract) that can be described and reasoned about within the domain.

2.2 Bayesian Ontology in PR-OWL

PR-OWL is a probabilistic extension to the OWL Web ontology language. Its logical basis is MEBN (Multi-Entity Bayesian Networks) theory. In the initial definition of PR-OWL as an upper ontology, concepts are related to representing uncertainty in a principled way using OWL syntax. PR-OWL is composed of a set of classes, subclasses and properties that collectively form a framework for building probabilistic ontology. The first step toward building a probabilistic ontology in compliance with our definition is to import into any OWL editor an OWL file containing the PR-OWL classes, subclasses, and properties. A probabilistic ontology must have at least one individual of class *MTheory*, which is a label linking a group of *MFrags* that collectively form a valid MEBN Theory. Individuals of class *MFrag* are comprise of nodes, and each node is a random variable, thus has a mutually exclusive and collectively exhaustive set of possible states. In PR-OWL, the object property *hasPossibleValues* links each node with its possible states, which are individuals of class Entity. Finally, random variables (represented by the class Nodes in PR-OWL) have unconditional or conditional probability distributions, which are represented by class Probability Distribution and linked to its respective nodes via the object property *hasProbDist* [5, 6].

Using the PR-OWL upper ontology can achieve a principled representation of uncertainty, and this ontology structure is translated to a Bayesian network (BN) by a set of translation rules that would rely on the probabilistic information attached to individual concepts and properties within the probabilistic ontology. Finally, uncertain reasoning is realized by BN reasoning tools.

3. ROTARY MECHINERY FAULT DIAGNOSIS MEDEL

Based on fault feature, Rotary Machinery fault diagnosis can find out fault cause. According to the current state, it does real-time monitoring and fault prevention. In view of fault diagnosis, it is described with a number of attributes. The various values of these attributes compose fault's state space. In the uncertain environment, fault cause and fault symptom contain several states. For example, a fault symptom maybe has three states: extremely high, slightly high, normal. Then the uncertain problems can be resolved by means of fault's probability distribution in some state. The uncertain relationship of the fault states or states and fault causes in course of diagnosis is represented in a rotary machinery fault diagnosis system. Based on this, it builds the basic framework of fault diagnosis, as follows in figure 1.

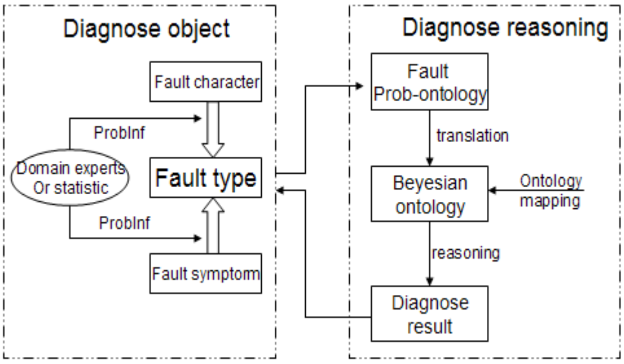


Figure 1. The Basic Framework of Fault Diagnosis Based on Bayesian Ontology

As it is shown in figure 1, fault type includes fault cause and diagnosis result. For instance, rotor fault diagnosis process goes as follows. First, it constructs rotor fault ontology. The rotor fault type (e.g. rotor unbalancing, rotor non-centering and axial cleft etc.), fault symptom and fault feature have a causal relationship. Fault symptom is used to describe various states initiated by fault, such as the high amplitude, the middle and the low, and fault feature describes their characters, for instance, rotor fault features mainly contain axis locus, stability of vibration, variation trend of vibration and so on. Second, it constructs rotor fault probabilistic ontology by augmenting OWL semantics to allow probabilistic information to be represented via additional markups. The result would be a probabilistic annotated ontology that could be translated to a Bayesian network (BN).Third, such a translation would be based on a set of translation rules that would rely on the probabilistic information attached to individual concepts and properties within the annotated ontology[7]. Each node composed of ontology class or entity corresponds to Bayesian network node with conditional probability distribution table (CPT). Fourth, it builds the ontology mapping based on Bayesian network method. Fifth, it uses Bayesian network as the underlying reasoning mechanism which can diagnose the correct result on the basis of partial information.

3.1 Constructing Rotor Fault Probabilistic Ontology

Rotor fault probabilistic ontology is built by using ontology tool that is protégé, a free, open source ontology editor and knowledge-base framework [8]. In PR-OWL, a rotor fault class is composed of rotor fault type class, fault symptom class and fault feature class. Each individual of class node is a random variable, so it has a mutually exclusive and collectively exhaustive set of possible states. The object property has *PossibleValues* links each node with its possible states, which are individuals of class entity. Finally, random variables (represented by the class nodes in PR-OWL) have unconditional or conditional probability distributions, which are represented by class probability distribution and linked to its respective nodes via the object property has *ProDist*. Figure 2 shows the partial major concepts and their relations of rotor fault class in PR-OWL.

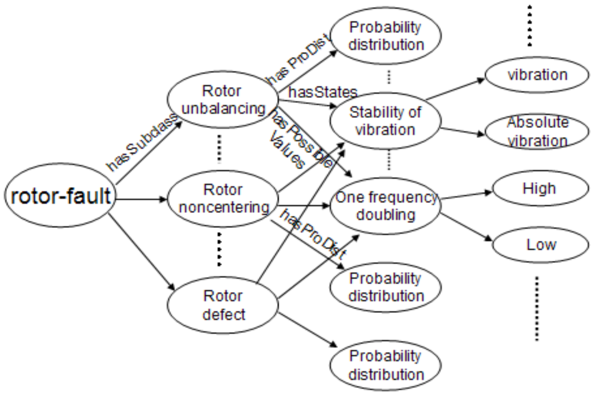


Figure 2. The Partial Major Concepts and Their Relations of Rotor Fault Class in PR-OWL

Probabilistic ontology achieves a principled representation of uncertainty and allow for the use of different probabilistic reasoning systems as a means to perform plausible reasoning and learning from data on the MEBN Theories represented in PR-OWL format. The advances presented here in terms of writing PR-OWL Bayesian Ontology can be directly applied to AI systems, increasing their potential for knowledge sharing, reusability, and interoperability.

3.2 Research of Reasoning Mechanism Based on Bayesian Ontology

In the fault diagnosis, the realization of the uncertain knowledge's representation and reasoning not only constructs probabilistic ontology but also considers to structure characteristics. Bayesian Ontology is used to describe knowledge about a domain with its associated uncertainty in a principled, structured, sharable and machine-understandable way. It must translate probabilistic ontology into Bayesian network from structure with a set of rules and as a sub-layer's reasoning method, Bayesian network can reason the fault diagnosis result on the basis of partial uncertain information. The fault diagnosis system is based on Bayesian ontology, which combines the logic reasoning of ontology with the probabilistic reasoning of Bayesian network and designs a rotor fault reasoning engine framework, as follows in figure 3.

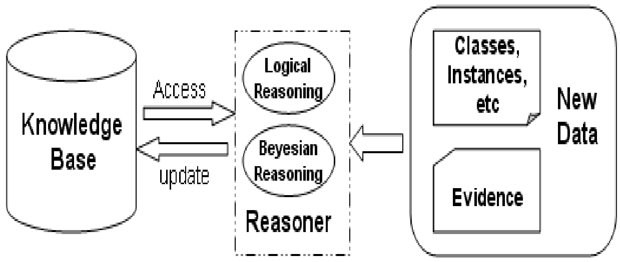


Figure 3. Reasoning Engine Framework of the Rotor fault

Knowledge base includes the probabilistic ontology and all concepts represented by the extended ontology language PR-OWL, and it expands the description to fault diagnosis area. According to the new data input by the user, such as a rotor fault feature, fault symptoms which are extracted from sensors, signal detection, inference engine finds out the generated fault node, extracts all the information in accord with the evidences, which include each conditional probabilistic distribution table of the Bayesian network node, and then analyzes the probability to meet the evidence, which can reason more accurate diagnosis result. At the same time, the current WEB agent makes use of probabilistic reasoning to update knowledge base via the used evidence. By adopting Ontology development tool to support the Bayesian inference engine's plug-in, the thesis accomplishes the uncertain reasoning process in the field of the fault diagnosis.

4. CONCLUSIONS

Bayesian Ontology is used to describe knowledge about a domain with its associated uncertainty in a principled, structured, sharable, and machine-understandable way. The thesis uses Bayesian Network to represent and reasoning uncertain knowledge used in the field of Rotary Machinery fault diagnosis, and constructs a fault diagnosis model based on Bayesian Ontology. In resolving the practical problems, it can instance the model. According to probabilistic technology, it makes reasoning from the uncertain and incomplete information. Reasoning’s granularity depends on the granularity of constructing ontology, as a result, the reasoning on the variable level is more accurate.

ACKNOWLEDGMENTS

This work was supported by Project of Hunan Provincial Key Lab under grant no: 2008TP4039-2 and by Project 50775070 supported by NSFC, by Scientific Research Fund of Hunan Provincial Education Department under grant no: 07C272.

REFERENCES

[1] HENDLER J. Ontologies on the Semantic Web [J]. IEEE Intelligent Systems, 2002, 17(2): 73-74.

[2] Jensen F V. Bayesian Network Basics. AISB Quarterly, 1996, 94:9-22.

[3] Paulo C. G. da Costa, Kathryn B. Laskey, Ghazi AlGhamdi. Bayesian Ontologies in AI Systems. Proceedings of the Fourth Bayesian Modeling Applications Workshop: Bayesian Models meet Cognition, held at the Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI 2006). July 13, 2006, Cambridge, MA, USA.

[4] Paulo Costa, Kathryn Laskey, Kenneth Laskey. PR-OWL: A Bayesian Ontology Language for the Semantic Web. In Proceedings of URSW, 2005.

[5] Costa, P. C. G. (2005). Bayesian Semantics for the Semantic Web. Doctoral Dissertation. Department of Systems Engineering and Operations Research, George Mason University: Fairfax, VA, USA. p.312. http://hdl.handle.net/1920/455.

[6] Laskey, K.B. (2006) MEBN: A Logic for Open-World Probabilistic Reasoning. http://mars.gmu.edu:8080/dspace/handle/1920/461.

[7] DING ZHONG-LI, PENG YUN. A probabilistic extension to ontology language OWL. Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS 04).Washington, DC: IEEE Computer Society, 2004, 4: 111-120.

[8] http://protege.stanford.edu/

AUTHOR CURRICULUM VITAE

Wenhui Xiao, female, is an assistant in School of Computer Science and Technology, Hunan University of Science and Technology, and her graduate teacher is Huifu Zhang. She graduates from Hunan Normal University and works on school teaching since 2003. The main directions of research are Knowledge Engineering (KE) in AI.

Huifu Zhang, male, is an associate professor in School of Computer Science and Technology, Hunan University of Science and Technology. He always works on school teaching and scientific research and his main directions of research are distributed system, computer grid and so on. In recent years, he has published over 10 papers on a number of journals and international conference, and some are received by EI and ISTP.

DISCUSSION OF FCM ALGORITHM WITH FLAWED SAMPLES

Xiao-Jun Tong, Hai-Tao Gan, Qin Jiang, Nong Sang, Wei Li, Xin Chen
Wuhan Polytechnic University, Wuhan, Hubei, 430023, China
Email: tongxiaojun1998@yahoo.com.cn

ABSTRACT

Fuzzy c-means clustering algorithm (FCM) is an important method in unsupervised classification, and takes an important role in pattern recognition. In the traditional FCM algorithm, it does not take account of the impact of samples' distribution density on cluster analysis. In this paper, a weighted FCM algorithm based on absorption is proposed, which makes the overall density of the sample more evenly distributed through appropriate homogenization. Through the IRIS and other data sets experiment, classification results show that this algorithm is better able to reduce the impact of the uneven distribution density. The clustering result is superior to the traditional FCM algorithm and the weighted FCM algorithm.

Keywords: FCM algorithm, Distribution density, Flawed samples

1. INTRODUCTION

Cluster analysis is to use mathematical methods to study and deal with the classification of the given object. It clusters the similar samples to the same class and the dissimilar samples to the different classes under a certain measure criteria. Cluster analysis is an important data analysis technique, which have a wide rang of application in pattern recognition, image processing, computer vision and other fields.

Fuzzy clustering can obtain the uncertainty degree of a sample which belongs to a class and express the intermediary of the samples, and reflect the real world objectively. Thus it has become the mainstream of cluster Analysis. In fact, the most popular fuzzy clustering method based on the objective function has been put forward by Ruspini [1]. This method is designed simply and can solve problem broadly. Furthermore, it can be transformed into an optimization problem and use the non-linear programming theory to solve it. And it is achieved easily on the computer. But the most effective algorithm-fuzzy c-means algorithm (FCM) is given by Dunn [2]. Bezdek [3, 4] developed the algorithm, and established the fuzzy clustering theory. Thence, FCM algorithm is developed vigorously and has formed an enormous system. Thus, with the use and development of computer, the fuzzy clustering algorithm based on the objective function has become a new hotspot.

In the traditional FCM algorithm, each feature of the samples plays a uniform contribution for clustering. But in fact, because the feature selection are not perfect, and their scalarization are eyeless, each feature of the feature vector is not uniform for clustering contribution. So we have to take account of the different effect of each feature. Based on the above, we introduce a method of calculating the feature weighted [5].

FCM algorithm has a strong dependence of the prototype, because the definition of the objective function depends on a prototype. The performance of FCM algorithm will be changed with the data structure. When data has defects such as uneven distribution density, that is, the distribution density of a certain part of data sets is larger than the other part, the impact on clustering of data sets is different. The part which has larger distribution density will give more impact on the cluster center, and it will result in the offset cluster centers, and affect the effect of FCM algorithm. For this issue, this paper presents an algorithm, that is the feature weighted FCM based on absorption, which is better able to resolve the problem than the traditional FCM algorithm.

2. THE TRADITIONAL FCM ALGORITHM

Assuming the sample set to be classified denotes as $X = \{X_1, X_2, \cdots, X_n\} \subset R^{n\times q}$, and the number of the samples is n, q is the feature space dimension. If we want to divide the sample set of X into class of c, then the membership degree of the samples which belong to the respective class is recorded as $U = \left[u_{ij}\right]_{c\times n}$ (fuzzy partition matrix). here, $u_{ij}(1 \le i \le c, 1 \le j \le n)$ shows the membership degree of sample X_j which belongs to class i, and u_{ij} should be meet the following two conditions:

$$\sum_{i=1}^{c} u_{ij} = 1, 1 \le j \le n \quad (1)$$

$$0 \le u_{ij} \le 1, 1 \le i \le c, 1 \le j \le n \quad (2)$$

Bezdek defined the general description of the fuzzy c-means:

$$\min J_m(X,U,P) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \quad (3)$$

In formula (3), $P = (p_1, p_2, \cdots, p_c)$ is a matrix of $s \times c$, $p_j(j = 1,2,\cdots,c) \in R^s$ is the center of class of j, $(d_{ij})^2 = \left\|x_j - p_i\right\|_A = (x_j - p_i)^T A(x_j - p_i)$ shows the general distance of data point x_j and p_i. If A is identify matrix, then the corresponding distance is Euclidean distance. Euclidean distance criteria are suitable for the ultra-spherical data sets. Using the different distances definition makes the clustering algorithm suitable for different types of clustering problem; m is the fuzzy index which control fuzzy degree of matrix U [6].

3. THE FEATURE WEIGHTED FCM ALGORITHM

In the traditional FCM algorithm, each feature of the samples plays a uniform contribution for clustering. But in fact, due to the feature selection are not perfect, and their scalarization have some eyeless, each feature of the feature vector is not uniform for clustering contribution [7]. So we have to take the different effect of each feature into account.

Suppose the clustering center of cursory classification be: $p_1, p_2, \ldots, p_c$, $p_i = \left\{p_{i1}, p_{i2}, \ldots, p_{is}\right\}$ .Next we will obtain the weight of feature by the two principles:

The principle of feature contribution balance: for ordinary classification methods, each feature contribution for clustering is important equally; but when a feature's contribution is great than the others, we have to process the original data and change the imbalance in order to make the feature balance.

Contribution balance principle, namely each feature contribution for clustering is important equally. Suppose the balance coefficient is r_j, we write it down as follows:

$$r_j = \frac{\max\left\{\sum_{i=1}^{c} p_{il}, l = 1,2,\cdots,s\right\}}{\sum_{i=1}^{c} p_{ij}} \quad j = 1,2,\cdots,s \tag{4}$$

The principle of most inter-cluster separate degree: the size of the separate degree shows that each feature has otherness, namely: the contribution of the separate degree is great, the feature weight is larger, so the new separate degree of all sorts is larger, and furthermore, the separate degree is stronger.

As said by the above, we define: the weight caused by the separate degree of different sorts is equal to the separate degree. We also know that standard deviation expresses how the data points concentrate and how they separate, so we can use the clustering prototype's standard deviation to scale the separate degree of different sorts. The corresponding expression is expressed as:

$$d_j = \sqrt{\sum_{i=1}^{c}\left(p_{ij} - \bar{p}_j\right)^2} \quad j = 1,2,\cdots,s \tag{5}$$

As the above principle, we get the following process: in order to process the clustering center, firstly we can use the traditional FCM algorithm to calculate the original clustering prototype, then, we normalize each feature of every clustering prototype. Thus we can get the balance coefficient of r, and the normalized separate degree ：

$$d_j = \sqrt{\sum_{i=1}^{c}(r_j p_{ij} - r_j \bar{p}_j)^2} \quad j = 1,2,\cdots,s \tag{6}$$

Lastly, we can get the feature weight of w.

From the above analysis, the feature weight w is expressed by the following form:

$$w_j = d_j \times r_j \quad j = 1,2,\cdots,s \tag{7}$$

The purpose of feature weight is that the feature with larger otherness should work more contribution for classification, i.e. the feature with more separability should make its weight larger. On the contrary, if a feature's separability is smaller, we can ignore it. Sum up, we may give the separate degree of d_j a power of m, thus we have the following formula:

$$w_j = (d_j)^{m'} \times r_j = \left(\sqrt{\sum_{i=1}^{c}(p_{ij} - \bar{p}_j)^2}\right)^{m'} \times r_j^{1+m'} \quad j = 1,2,\cdots,s \tag{8}$$

From the above formula, we know: there is a relationship between the separate degree of the original clustering prototype and the balance coefficient, i.e.: the power of the balance coefficient = the power of the original clustering prototype's separate degree + 1.

Following the above analysis, we know the feature weight can be divided into two parts: the balance coefficient and the normalized separate degree. Next, we try to divide the latter into two parts: the balance and the separate degree of the original prototype, namely:

$$d_j = \sqrt{\sum_{i=1}^{c}(p_{ij} - \bar{p}_j)^2} \times r_j \quad j = 1,2,\cdots,s \tag{9}$$

And

$$w_j = \sqrt{\sum_{i=1}^{c}(p_{ij} - \bar{p}_j)^2} \times r_j^2 \quad j = 1,2,\cdots,s \tag{10}$$

Now if we don't take into account the relationship between the separate degree of the original clustering prototype and the balance coefficient, we can give two powers (m_1, m_2) to the parts, thus the feature weight w becomes:

$$w_j = \left(\sqrt{\sum_{i=1}^{c}(p_{ij} - \bar{p}_j)^2}\right)^{m_2} \times r_j^{m_1} \quad j = 1,2,\cdots,s \tag{11}$$

Based on the above characteristics of the weighted theory and fuzzy clustering theory, theory of the feature weighted FCM algorithm is as follows:

First of all, the original data is rough classified to obtain the initial clusters center by the traditional FCM. Then the above formulas to calculate weights, and then to deal with original data as follows:

$$x' = w \bullet x_k, w = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_s \end{bmatrix}, k = 1,2,\cdots,n \tag{12}$$

After pretreatment, Euclidean distance of any two feature vectors can be formulated as

$$D(x_k', x_j') = \left\|x_k' - x_j'\right\| = \left(x_k - x_j\right)^T W \left(x_k - x_j\right) \tag{13}$$

here $W = w^T w$. Weighted matrix W will be optimized in the clustering process, and finally achive the optimal weighted value under a certain criteria.

4. A WEIGHTED FCM ALGORITHM BASED ON ABSORPTION

In the real world, the problem of uneven distribution of density is often encountered. In the process of data collection there will always be deviations, as shown in Figure 1:

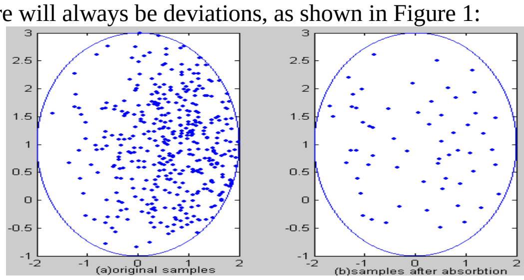


Figure 1. Samples' uneven distribution density

Figure 1(a) shows a phenomenon that a data set has uneven distribution density. When it occurs, the distribution density for a certain part of the data set is clearly larger than the other part. In this case, the impact of every data on clustering is different. The impact of data which has larger distribution density is stronger that the other. It will result in the offset cluster center, thus it will affect the effect of FCM clustering algorithm. If so, we need an algorithm which can make a larger distribution density reduce the density. Then density distribution of the entire data set is more uniform, as showed in Figure 1(b). Based on the above considerations, a weighted FCM algorithm based on absorption is proposed in this paper. The algorithm first absorbs the data set by using the traditional FCM algorithm so that density of the samples after absorption is more evenly distributed, and then obtain optimal value by using the feature weighted FCM algorithm.

The following describes the algorithm:

Step 1: initial class number c and absorption $c^{'}$, here $2 \le c \le n, c \le c^{'} \le n$;

Step 2: absorb the data set by using the traditional FCM algorithm;

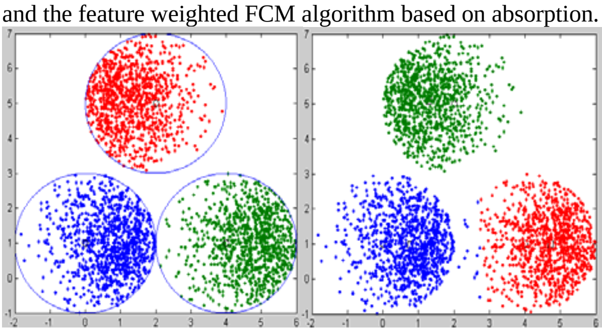
Step 3: cluster the data set after absorption by using the feature weighted FCM algorithm, and achieve the optimal value.

5. EXPERIMENTS

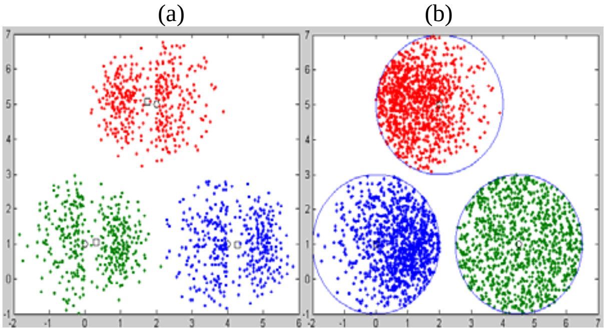
To test validity of the proposed algorithm, we give two experiments through man-made data set and IRIS set, and obtain some preliminary experimental results.

5.1 Test of Man-made Data Set

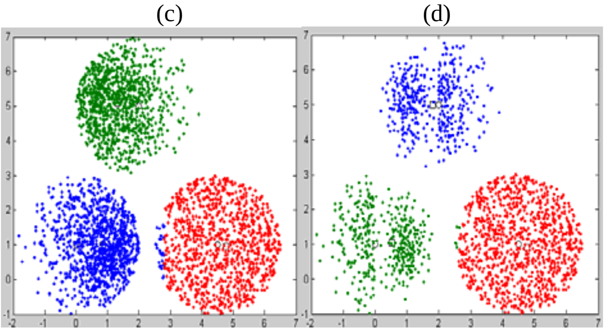
Samples used in this experiment are taken from two-dimensional sample set, composed by 3000 samples and divided into three classes clearly. There are 1000 samples in each class. We tested three experiments. Radius is the same in every class of three experiments, and is 2. In the First experiment three classes have unevenly distributed density, and in the second experiment there are two classes with uneven distribution density, the third experiment there is only one class with uneven distribution density. We compare the effect of their clustering by using traditional FCM algorithm and the feature weighted FCM algorithm based on absorption.



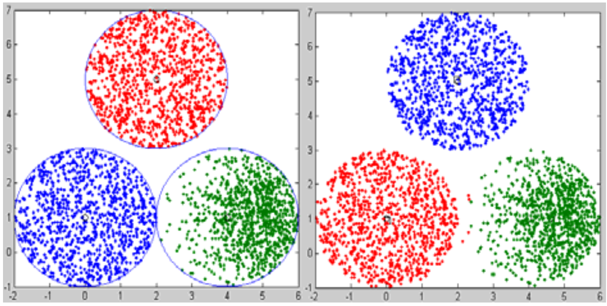
(a) (b)



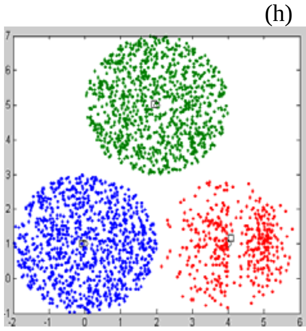
(c) (d)



(e) (f)



(g) (h)



(i)

Figure 2. Experimental results of the two algorithms

Separately, Figure (a), (d) and (g) shows original samples of three experiments. Figure (b), (e) and (h) shows the results of the traditional FCM algorithm. Figure (c), (f) and (i) shows the results of the feature weighted FCM algorithm.

From figure 2(a), (d) and (g) it can be seen that data sets can be clearly divided into three classes, at the same time every set has its characteristic. Figure 2(b), (e) and (h) show that the uneven distribution density makes cluster centers offset to the direction of the larger distribution density, and it results in many relatively misclassified numbers. Compared to the Figure 2 (b), from Figure 2(c) we can clearly see that the cluster centers are closer to the actual centers and the clustering result is relatively better. From Figure 2(e) and (h), the cluster center of data set which has uniform distribution density does not deviate. However, because the cluster center of data set which has uneven distribution density deviate, there are misclassified samples. From Figure 2(f) and (i), because of the sample homogenization which dilute the data set, clustering results are greatly improved.

From the above analysis, the uneven distribution density impacts on clustering. And the algorithm proposed in this paper can better solve this problem, and it is proved that the weighted FCM algorithm based on absorption is effective.

5.2 Test of IRIS

In order to test the effectiveness of the new algorithm, we have adopted a well-known IRIS data as the actual test data sets. IRIS is composed by the 150 sample points in the four-dimensional space. Four components of every sample separately IRIS express Petal Length, Petal Width, Sepal Length and Sepal Width. The set includes three types Setosa, Versicolor and Virginica, and each category has 50 samples, Setosa is complete separated from the other two categories, while there are sample crossed between Versicolor and virginica. IRIS data are often used to test the performance of clustering algorithm. Hathaway [8] in 1995, give the set the actual cluster center: pl=(5.00,3.42,1.46,0.24),p2=(5.93,2.77,4.26,1.32),p3= (6.58,2.97,5.55,2.02). IRIS is classified by separately using traditional FCM algorithm, the feature weighted FCM algorithm and the feature weighted FCM algorithm based on absorption in order to test the performance of the three algorithms.

As can be seen from the figure, the number of misclassified points is varied by the degree of absorption, and it is overall increased by increasing the number of absorption. In other words, the number of misclassified data points can be seen as a monotonically increasing function of the number of absorption. But there are also outliers, such as the number of absorption 29. The optimal number of absorption will be discussed in another paper.

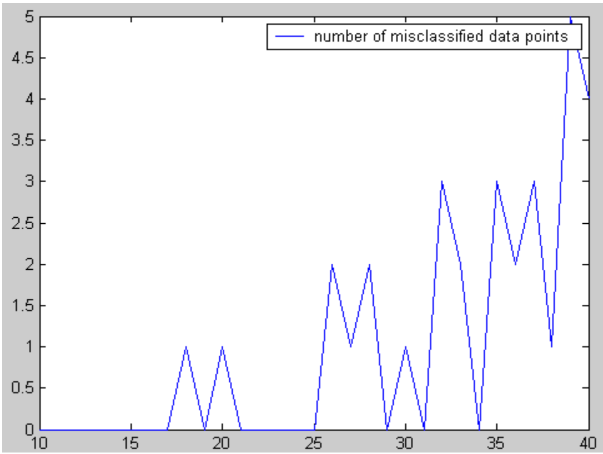


Figure 3. Curve of the numbers of misclassified data points

From the table 1, we can see that with respect to the number of misclassified points, the new algorithm is superior to the traditional FCM algorithm and the feature weighted FCM algorithm. Cluster center obtained by the feature weighted FCM algorithm is closer to the true cluster center, but the number of misclassified points is more. This shows that the feature weighted FCM algorithm is powerless to handle uneven distribution density data. Two FCM algorithms based on weighted feature separately obtain the final weight as following:

w_1=[1.0934,2.4063,4.3593,26.551] and

w_2=[1.1645,2.4482,4.4378,27.5064] .

Table 1. Clustering effect of three algorithms

algorithm	Misclassified numbers	Misclassified rate	Cluster center	Error sum of squares
Traditional FCM algorithm	16	10.67%	$p_1=(5.0062,3.4242,1.4684,0.2492)$ $p_2=(5.8946,2.7460,4.4154,1.4273)$ $p_3=(6.8484,3.0750,5.7283,2.0741)$	0.1554
The feature weighted FCM	6	4%	$p_1=(5.0059,3.4211,1.4696,0.2458)$ $p_2=(5.9820,2.7679,4.3616,1.3494)$ $p_3=(6.6605,3.0281,5.6017,2.1017)$	0.0333
The feature weighted FCM based on absorption	0	0	$p_1=(5.0604,3.4883,1.4758,0.2407)$ $p_2=(5.9704,2.7791,4.3089,1.3348)$ $p_3=(6.7080,3.0372,5.6402,2.0868)$	0.0464

6. CONCLUSIONS

The experimental results show that the algorithm proposed in this paper can be better reduced the impact of the uneven distribution density, so that the numbers of misclassified points decrease. And its effect is superior to the traditional FCM algorithm and the feature weighted FCM algorithm. But there are also some problems. It is the future direction to be explored that How to determine the optimal number of absorption and how to identify whether the distribution density is uniform or not.

ACKNOWLEDGEMENTS

This paper is supported by National Natural Science Foundation under Grant 79970025, 60403002 and 30370356 of China, and the plan of Science and Technological Innovation Team of the Outstanding Young and Middle-aged Scholars of Hubei Provincial Department of Education, and Hubei Provincial Department of Education under Grant D20081802 and Hubei provincial Natural Science Foundation under Grant 2004ABA031, 2005ABA233and2007ABB030, and National Postdoctoral Science Foundation of china (Grant 2004036016), and Foundation of Hubei Provincial Department of Education Grant 2003X130 and Scientific Research of Wuhan Polytechnic University Grant 06Q15.

REFERENCES

[1] Ruspini HE, A new approach to clustering, Inf. Cont, 15: 22~32, 1969.

[2] Dunn JC, Well. Separated clusters and the optimal fuzzy partitions. J Cybernet, 4: 95~104, 1974.

[3] Bezdek JC, Pattern recognition with fuzzy objective function algorithms. New York: Plenum Pres, 1981.

[4] Bezdek JC, Hathaway R, et al. Convergence and theory for fuzzy C- means clustering: counterexamples and repairs, IEEE Trans PAMI, 17(5): 873~877, 1987.

[5] Xiaojun Tong, Qin Jiang, Haitao Gan, Shan Zeng, Kai Zhao, Research On the Calculation Mothod for Weight of the Feature Weight Fuzzy Clustering Algorithm. CDABES, July, 27-31, 2008.

[6] Gao xinbo, Fuzzy Cluster Analysis and its Applications. The Publishing Company of Xian Electron Technology University, Xi'an, 2004.

[7] Li Jie, GAO Xin-bo, JIAO Li-cheng, A New Feature Weighted Fuzzy Clustering Algorithm, ACTA ELECTRONICA SINICA.Jan, 34(1), 2006.

[8] Hathaway R J, Bezdek J C, Nerf C-means: Non-Euclidean relation fuzzy clustering [J], Pattern recognition, 27(3):429-437, 1994.

A CLASS OF ACCELERATED M-STEP NEWTON METHOD*

Wang Xiuhua[1,2], Kou Jisheng[2]
[1]Department of Mathematics, Xiaogan University
Xiaogan 432100, Hubei, China
[2]Department of Mathematics, Shanghai University
Shanghai 200444, China
E-mail: koujisheng@yahoo.com.cn.

ABSTRACT

Solving systems of nonlinear algebraic equations is a common and important problem in science and engineering; Newton method and its variant m-step Newton method are very important and frequently used for solving these problems. In this paper, a class of new accelerated m-step Newton methods is proposed. Compared with m-step Newton method, the new methods can be more stable and efficient. Also, the parallel algorithm is given. Numerical results show that the new method can be of practical interest. These accelerating methods can be extended to many other known methods, such as Newton-like methods, m-step discrete Newton method and Brent method, and can improve their convergence and stability.

Keywords: Parallel Algorithm, Systems of Nonlinear Algebraic Equations, Stability, M-Step Newton Method

1. INTRODUCTION

Solving the systems of nonlinear equations is a very important problem in scientific and engineering computing areas. In this paper, we consider the following nonlinear systems

$$F(x)=0 \tag{1}$$

where $F: R^n \to R^n$ is a continuously differentiable nonlinear mapping.

Here, $F=(F_1,F_2,\cdots F_n)^T$, $F_i=F_i(x)$, $x=(x_1,x_2,\cdots x_n)^T$. (2)

Many robust and efficient methods for solving (1) are brought forward [1]. Newton method for (1) is an important and basic method, which converges quadratically. Some improvements of Newton method are given in [2-6]. However, it is expensive to compute the Jacobian matrix at each iteration, especially when the size of the problem is very large. The m-step Newton method is given to reduce the computational cost of Jacobian matrix, so it works better than Newton method.

Per iteration the computational error can not be avoided for any method. Compared with Newton method; per iteration the computational error produced by m-step Newton method often makes the computed approximate solution far away from the exact solution. In some cases, the iteration sequences produced by m-step Newton method are not robust and even can not converge because of the perturbation of the computational error. This problem which makes less practical interest will be shown in our numerical tests.

In order to resolve this problem, a modified method (WHM1) is given in [7], based on WHM1, a family of improved m-step Newton Method is proposed in [8] and it also notes that different methods can be derived due to different accelerating schemes. An improved m-step Newton Method (WHM2) is also given as a special case. So it may be very interesting to choose some efficient accelerating schemes. In this paper, we present a class of new accelerated m-step Newton methods based on the method given in [8]. Numerical results show that the new methods work better than m-step Newton method, WHM1 and WHM2.

2. THE IMPROVED SCHEME OF M-STEP NEWTON METHOD

2.1 Structure of the Accelerating Scheme

The fixed point iterative scheme for nonlinear systems (1) is $x^{k+1}=G(x^k)$. Suppose that this iterative scheme converges linearly, and then construct the following iterative scheme

$$\begin{cases} y^k = G(x^k) \\ z^k = G(y^k) \\ x^{k+1} = y^k - H(x^k,y^k,z^k), k=0,1,2,\cdots \end{cases} \tag{3}$$

where

$$H(x^k,y^k,z^k) = \frac{(z^k-y^k)^T\{P[\alpha(z^k-y^k)+\beta(y^k-x^k)]\}}{(z^k-2y^k+x^k)^T\{P[\alpha(z^k-y^k)+\beta(y^k-x^k)]\}}(y^k-x^k). \tag{4}$$

If we take different parameters α,β and matrix P, we can get different iterative schemes, The last substep is called accelerating scheme.

2.2 M-step Newton Method and the Structure of Its Improved Scheme

The m-step Newton method is an efficient modification of Newton method. It converges fast and can reduce the computational cost. In m-step Newton method, after computing Jacobian matrix once, it will be used m times. Generally, the convergent order of this method is m+1 and the efficiency of this method is higher than Newton method. The iterative scheme is given by

$$\begin{cases} x^{k,0} = x^k \\ x^{k,i} = x^{k,i-1} - [F'(x^k)]^{-1}F(x^{k,i-1}), \\ x^{k+1} = x^{k,m} \\ i=1,2,\cdots,m; k=0,1,2,\cdots \end{cases} \tag{5}$$

From (5), we know that m-step Newton method has double iteration. Its inner iteration is simplified Newton iteration which converges linearly. The inner iteration of m-step Newton method keeps the Jacobian matrix invariable. It reduces the computational cost, but the computational error produced by m-step Newton method often makes the computed approximate solution far away from the exact solution. In some cases, the iteration sequences produced by m-step Newton method are not robust and even can not converge because of the perturbation of the computational error. So we can use the scheme (3) to accelerate the inner iteration and a class of new accelerating methods will be derived, which efficiency is higher than m-step

* **Kou Jisheng** is corresponding author.
The work is supported by National Natural Science Foundation of China (No. 10826082), Key Disciplines of Shanghai Municipality (No. S30104) and Shanghai Leading Academic Discipline Project (No. J50101).

Newton method. Our new method is given by

$$\begin{cases} x^{k,0} = x^k \\ x^{k,1} = x^{k,0} - [F'(x^k)]^{-1} F(x^{k,0}) \\ y^{k,i} = x^{k,i} - [F'(x^k)]^{-1} F(x^{k,i}) \\ z^{k,i} = y^{k,i} - [F'(x^k)]^{-1} F(y^{k,i}), \\ x^{k,i+1} = y^{k,i} - H(x^{k,i}, y^{k,i}, z^{k,i}) \\ x^{k+1} = x^{k,m+1} \\ i = 1,2,\cdots,m; k = 0,1,2,\cdots \end{cases} \quad (6)$$

where $H(x^{k,i}, y^{k,i}, z^{k,i})$ is the same to the scheme (3). We call the scheme (6) as accelerating m-step Newton method.

3. THE PARALLEL ALGORITHM

The parallel algorithm for the scheme (6) is given in the following.

Step 1: Give the initial approximation x^0 ,the error ambit ε ,the maximal iterative times $N_{\max}$ and positive integer m.

Step 2: Suppose the iteration has carried out k times, x^k has been computed. When computing $F(x^k)$, simultaneously compute the Jacobian matrix and decompose $F'(x^k)$ according to the nonlinear systems at the same time. We take the QR decomposability. Let $F'(x^k) = QR$, where Q is orthogonal matrix, R is upper triangular matrix;

Step 3: Let $x^{k,0} = x^k$, by $R\Delta x^{k,0} = -Q^T F(x^k)$,compute $\Delta x^{k,0}$ and $x^{k,1} = x^{k,0} + \Delta x^{k,0}$,then compute $F(x^{k,1})$;if $\| F(x^{k,1}) \| < \varepsilon$,print $x^* = x^{k,1}$ and $\| F(x^{k,1}) \|$,stop;

Step 4: Suppose the accelerating iteration has done $i-1$ times, $x^{k,i}$ and $F(x^{k,i})$ have been figured out. If $\| F(x^{k,i}) \| < \varepsilon$, print $x^* = x^{k,i}$ and $\| F(x^{k,i}) \|$, stop; otherwise, carry out the following item in turn.

Step4.1: By $R\Delta x^{k,i} = -Q^T F(x^{k,i})$, compute $\Delta x^{k,i}$ and $y^{k,i} = x^{k,i} + \Delta x^{k,i}$, then compute $F(y^{k,i})$.

Step4.2: By $R\Delta y^{k,i} = -Q^T F(y^{k,i})$, compute $\Delta y^{k,i}$ and $z^{k,i} = y^{k,i} + \Delta y^{k,i}$.

Step4.3: By $x^{k,i+1} = y^{k,i} - H(x^{k,i}, y^{k,i}, z^{k,i})$, compute $x^{k,i+1}$ and $F(x^{k,i+1})$.

Step 5: If $i < m$, then $i \leftarrow i+1$,go to step 4;

Step 6: If $\| F(x^{k,m+1}) \| < \varepsilon$ or $k > N_{\max}$,then print $x^* = x^{k,m+1}$ and $\| F(x^{k,m+1}) \|$,stop; otherwise, $k \leftarrow k+1$,go to step 2.

4. NUMERICAL TESTS

In this section, we present some numerical results for the new methods and compare them with m-step Newton method, WHM1 and WHM2 on their numerical behavior. In the new methods, WHM1 and WHM2, we take m=1, while taking m=3 in m-step Newton method. So the computational cost about the outer iteration of the four methods is almost equivalent. In Tables 1-4 we show the results about the outer iteration of the four methods ($\| F(x^{(k)}) \|_2$). All computations are carried out with double arithmetic precision.

Example 1.

$$\begin{cases} f_1(x) = x_1 - \cos(x_2) = 0 \\ f_2(x) = \sin(x_1) + 0.5x_2 = 0 \end{cases}$$

with the exact solution $x^* = (0.530\cdots, -1.01\cdots)^T$

In example 1, we take $x^{(0)} = (4,1)^T$ for the four methods and in the new method, take $\alpha = 3.6, \beta = -1,\ P = F'[x^k]$. The results of the main outer iteration are shown in Table 1.

Table 1. Results of the main outer iteration with $x^{(0)} = (4,1)^T$, $\alpha = 3.6, \beta = -1,\ P = F'[x^k]$

n	New method	m-step Newton method	WHM1	WHM2
1	1.489	1.923	60.171	1.967
2	0.794	298.351	15.896	1.070
3	9.175e-3	34.734	1.690	0.019
4	9.702e-11	7.559	8.045e-4	1.442e-7

Table 2. Results of the main outer iteration with $x^{(0)} = (3,2.2)^T$, $\alpha = 1.5, \beta = -2, P = E_2$

n	New method	m-step Newton method	WHM1	WHM2
1	5.83	2.18	7.88	4.64
2	1.07	1.01e+2	11.26	3.73
3	1.12	1.64e+3	1.48	0.77
4	0.91	5.96e+2	0.62	0.87
5	9.01e-6	2.12e+1	0.90	0.45
6	1.11e-16	5.85	0.37	1.06e-3

In example 1, we take $x^{(0)} = (3,2.2)^T$ for the four methods, while in the new method, take $\alpha = 1.5, \beta = -2, P = E_2$. The results of the main outer iteration are shown in Table 2.

Example 2. $\begin{cases} f_1(x) = x_1 x_2 + 1 = 0 \\ f_2(x) = x_1 + \exp(x_1 + x_2) = 0 \end{cases}$

with the exact solution $x^* = (-1,1)^T$.

In example 2, we take $x^{(0)} = (3.5,5.7)^T$ for the four methods and in the new method, choose $\alpha = 1, \beta = 0, P = E_2$. The results of the main outer iteration are shown in Table 3.

Table 3. Results of the main outer iteration with $x^{(0)} = (3.5,5.7)^T$, $\alpha = 1, \beta = 0, P = E_2$

n	New method	m-step Newton method	WHM1	WHM2
1	3992.57	21247.74	4410.35	3996.94
2	656.69	2001.68	559.93	644.12
3	93.30	196.97	71.32	89.31
4	10.40	22.30	6.85	9.98
5	1.04	2.99	0.61	1.09
6	1.02e-3	0.58	8.66e-3	1.30e-3
7	6.70e-13	0.18	1.03e-10	1.40e-11

In example 2, we take $x^{(0)} = (5.2, 3.8)^T$ for the four methods and in the new method, choose $\alpha = 3, \beta = -1.9$, $P = F^{'}[x^{k}]$. The results of the main outer iteration are shown in Table 4.

Table 4. Results of the main outer iteration with $x^{(0)} = (5.2, 3.8)^{T}$, $\alpha = 3, \beta = -1.9$, $P = F^{'}[x^{k}]$

n	New method	m-step Newton method	WHM1	WHM2
1	3019.38	27265805.77	3040.84	3019.40
2	516.76	2531202.39	385.92	512.10
3	114.64	234979.76	49.98	109.98
4	3.38	21812.19	5.02	8.64
5	0.52	2023.73	275.28	1.31
6	1.89e-3	187.30	9.18	1.43
7	1.88e-10	17.21	0.91	7.19e-3

From Tables 1-4, we know that if the proper parameters α, β and matrix P in the new method are chosen according to different problems, after the iteration has done the same times, the computed approximate solution by the new method is closer to the exact solution than m-step Newton method, WHM1 and WHM2. So the new methods can reduce the accumulation of computational error and the efficiency is enhanced.

5. CONCLUSIONS

We propose a class of accelerating scheme bases on literature [8] and choose a proper accelerating scheme according to the different problems. The scheme is applied on m-step Newton method, therefore a new method is developed, which is faster and stable. Moreover, the parallel algorithm is given. Numerical results show that the new method is very efficient. This accelerating method can be extended to many other known methods, such as Newton-like methods, m-step discrete Newton method and Brent method, and can improve their convergence and stability.

REFERENCES

[1] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equation in Several Variables, Academic Press, New York, 1970.

[2] Verant. M. A Wavelet Adaptive Newton Method for the Solution of Nonlinear Equations [J]. Applied Mathematics Letter, 2003, 16：1301-1306.

[3] Chen Jinhai , Li Weiguo. Convergence Behaviour of Inexact Newton Methods under Weak Lipschitz Condition [J]. Journal of Computational and Applied Mathematics,2006,191: 143-164.

[4] Elias R,N, Coutinho A,L,G,A,Martins M,A,D. Inexact Newton-type Methods for the Solution of Steady Incompressible Viscoplastic Flows with the SUPG/PSPG Finite Element Formulation [J]. Computer Methods in Applied Mechanics and Engreering, 2006, 195: 3145-3167.

[5] Kou Jisheng, Li Yitian, Wang Xiuhua. A Modification of Newton Method with Third-order Convergence [J]. Applied Mathematics and Computation, 2006,181: 1106-1111.

[6] Kou Jisheng, Li Yitian, Wang Xiuhua. Third-order Modification of Newton's Method [J], Journal of Computational and Applied Mathematics, 2007, 205:1-5.

[7] Wang Xiuhua ,Huang Benwen. Improvement on Modified Newton′s Method, Journal of Wuhan University of Technology (Transportation Science & Engineering) ,2007 ,31(2):,345-347. (in Chinese)

[8] WANG Xiu-hua , HUANG Ben-wen. Family of Improved m-step Newton Method, JOURNAL OF WUHAN UNIVERSITY OF TECHNOLOGY, 2008, 30(12):192-195. (in Chinese)

Wang Xiuhua is a lecturer in department of mathematics, Xiaogan University.She is also a Ph.d. candidate in department of mathematics, Shanghai University. Her research interests are in numerical methods for nonlinear problems and parallel numerical algorithms.

Kou Jisheng is a lecturer in department of mathematics, Shanghai University. His research interests are in numerical methods for nonlinear problems and parallel numerical algorithms.

THE DECOMPOSITION THEOREM AND THE REPRESENTATION THEOREM UNDER THE TRIANGULAR NORMS OPERATORS

Qin Jiang[1,2], Xiaojun Tong[2], Lingde Xie[2], Shan Zeng[2], Nong Sang[2]
[1]Yunyang Teachers' College, Danjiangkou, Hubei, China
[2]Wuhan Polytechnic University, Wuhan, Hubei, China
Email: jqwing910@163.com

ABSTRACT

The decomposition theorem of the fuzzy sets reveals the relation of the fuzzy sets and the classical sets; and the representation theorem shows how to use nested sets to construct a fuzzy set, but these results are all gotten under Zadeh operators. In this paper, at first, we discuss the decomposition theorem under the general triangular norms operators, and gain the necessary and sufficient condition of the theorem, namely: the "or" operator must be max, but the "and" operator may be an arbitrary T operator of the triangular norms; secondly, we discuss the representation theorem, and get the following result: the necessary and sufficient conditions of the two equations of the representation theorem under S norm are all that the "or" operator must be max.

Keywords: T Norm, S Norm,The Decomposition Theorem, The Representation Theorem.

1. INTRODUCTION

Every concept has its connotation and extension. From the point of set theory, a concept's extension is a set. Some concepts have an explicit extension in particular occasion, but some do not have one. We call the latter ambiguity. The concept with ambiguity extension is called fuzzy concept.

Because classical sets can only express the concepts with explicit extension, can't express fuzzy concept, so the classical sets theory is powerless in the face of fuzzy concepts. In order to portray quantitatively fuzzy concepts and fuzzy phenomenon, L.A.Zadeh professor put forward the important concept of fuzzy sets in 1965 [1]. The basic idea is: expanding the affiliation of classical so that the membership of the elements only taken by both 0 and 1 values for the classical sets can be extended to take an arbitrary value in unit interval [0,1] , thus a fuzzy set is described. The elements of fuzzy set are vogue so that it can only expressed by its membership function.

The theory based on the classical sets provides a solid foundation for the application and development of science and technology of classical sets, but the fuzzy sets can study more and broader fields. If we are able to construct a relation between the fuzzy set and the classical set, it is possible to use the classical set theory to study the fuzzy set. For example, in literature [2,3], the fuzzy set A of U is expressed by the union of some classical sets of U through the decomposition theorem and the representation theorem, thus we can translate the study about the fuzzy sets into the research of correlative classical sets; in literature [4], the author represents the fuzzy set A of U as the intersection of some classical sets on the basis of redefining the scalar product λA from the real number λ and the fuzzy set A .These studies afford a lot of convenience for the research of fuzzy system.

The operation of fuzzy set has long been to expand the triangular norms operator, and the different operation of fuzzy set has different importance for fuzzy inference. But the decomposition theorem and the representation theorem of fuzzy set are confined under Zadeh operator, how to study the two theorems under the general triangular norms operator is particularly important.

2. THE DECOMPOSITION THEOREM UNDER ZADEH OPERATOR

Define2.1
Suppose $A \in F(U)$ be a fuzzy set, $\lambda \in [0,1]$, if

$$A_\lambda = \{u \mid u \in U, A(u) \ge \lambda\} \tag{1}$$

then A_λ is called a λ-cut set of A, and λ is called threshold or confidence level.If

$$A_{S\lambda} = \{u \mid u \in U, A(u) > \lambda\} \tag{2}$$

then $A_{S\lambda}$ is called λ-strong cut set of A [5].

Define 2.2
Suppose $A \in F(U)$, if

$$(\lambda A)(u) = \lambda \wedge A(u) \tag{3}$$

then λA is called scalar product of λ and A.Where

$$\lambda \wedge A(u) = \begin{cases} A(u), \lambda \ge A(u) \\ \lambda, \lambda < A(u) \end{cases} \tag{4}$$

Obviously λA is still a fuzzy set of U .

Define2.3
Suppose $A \in F(U)$ be a fuzzy set, if

$$C_A(u) = \begin{cases} 1, u \in A \\ 0, u \notin A \end{cases} \tag{5}$$

then $C_A(u)$ is called an eigenfunction of A [6].

Theorem2.1
(The Decomposition Theorem Ⅰ) Suppose $A \in F(U)$,we have

$$A = \bigcup_{\lambda \in [0,1]} (\lambda A_\lambda) \tag{6}$$

Theorem2.2
(The Decomposition Theorem Ⅱ) suppose $A \in F(U)$,we have

$$A = \bigcup_{\lambda \in [0,1]} (\lambda A_{S\lambda}) \tag{7}$$

3. THE REPRESENTATION THEOREM UNDER ZADEH OPERATOR

Define3.1
Suppose the mapping of sets value:

$H:[0,1]\to P(U)$

$$\lambda \to H(\lambda) \quad (8)$$

If H has the following property:

$\lambda_1 \geq \lambda_2 \Rightarrow H(\lambda_1) \supseteq H(\lambda_2)$,we call H nested sets of U [7].

Theorem3.1

(the Representation Theorem) Suppose H be a arbitrary nested sets of U , then

$$A = \bigcup_{\lambda\in[0,1]} \lambda H(\lambda) \quad (9)$$

is a fuzzy set of U , and $\forall \lambda \in [0,1]$,we have

(1) $A_{S\lambda} = \bigcup_{\alpha>\lambda} H(\alpha)$

(2) $A_{\lambda} = \bigcup_{\alpha<\lambda} H(\alpha)$

4. THE DECOMPOSITION THEOREM UNDER TRIANGULAR NORMS OPERATORS

Firstly, we will give T norm and S norm [8, 9].

Define4.1

Mapping T : $[0,1]^2 \to [0,1]$ is called a triangle norm, when $\forall a,b,c \in [0,1]$, the following conditions are satisfied:

(1) $T(a,b) = T(b,a)$

(2) $T(T(a,b),c) = T(a,T(b,c))$

(3) if $a \leq c, b \leq d$,then $T(a,b) \leq T(c,d)$

(4) $T(1,a) = a$

Define4.2

Mapping S : $[0,1]^2 \to [0,1]$ is called a triangle norm, when $\forall a,b,c \in [0,1]$, the following conditions are satisfied:

(1) $S(a,b) = S(b,a)$

(2) $S(S(a,b),c) = S(a,S(b,c))$

(3) if $a \leq c, b \leq d$,then $S(a,b) \leq S(c,d)$

(4) $S(0,a) = a$

Property 1: Suppose T be T norm, $\forall a,b \in [0,1]$, we can get:

$T(a,0) = 0$

Property 2: Suppose S be S norm, $\forall a,b \in [0,1]$,we can get:

$S(a,1) = 1$ [10]

As following, we will give the decomposition theorem under the triangle norms.

Theorem4.1

(the Decomposition Theorem Ⅰ) Fuzzy set A can be decomposed the following formula under T norm and S norm :

$$\mu_A(u_i) = \underset{\lambda\in[0,1]}{S}(T(\lambda, C_{A_\lambda}(u_i))) \quad (10)$$

Its necessary and sufficient condition is: $S = \vee$.

Proof

(Sufficient Condition) Suppose: $u_A(u_i) = \lambda_0$, then

$$\underset{\lambda\in[0,1]}{S}(T(\lambda, C_{A_\lambda}(u_i)))$$
$$= \bigvee_{\lambda\in[0,1]}(\underset{\mu_A(u_i)\geq\lambda}{T}(\lambda,1), \underset{\mu_A(u_i)<\lambda}{T}(\lambda,0))$$
$$= \bigvee_{\lambda\in[0,\mu_A(u_i))} \lambda$$
$$= \lambda_0 = \mu_A(u_i)$$

(Necessary Condition) Suppose the decomposition theorem is true under T norm and S norm.

Now we will prove the following equation is true for all fuzzy sets:

$\underset{\lambda\leq\lambda_0}{S}(\lambda_0,\lambda) = \lambda_0$,

Namely: $S(\lambda_0,\lambda) = \max\{\lambda,\lambda_0\}$.

Construct a fuzzy set: $\underset{\sim}{A} = \{x | x \in [0,1]\}$.

Suppose: $\mu_{\underset{\sim}{A}}(u_i) = \lambda_0$, $\forall\lambda$, $\lambda \in [0,1]$.

When $\lambda < \lambda_0$, we have:

$S(\lambda_0,\lambda_0) \geq \underset{\lambda<\lambda_0}{S}(\lambda_0,\lambda) \geq S(\lambda_0,0) = \lambda_0$

So we just need to prove: $S(\lambda_0,\lambda_0) = \lambda_0$;

(Reduction to Absurdity) Suppose:

$S(\lambda_0,\lambda_0) \neq \lambda_0$,

Because $S(\lambda_0,\lambda_0) \geq S(\lambda_0,0) = \lambda_0$,

So we can suppose: $S(\lambda_0,\lambda_0) = \lambda_1 > \lambda_0$.

Choose λ_2 , so as to: $\lambda_0 < \lambda_2 < \lambda_1$, so we can gain:

$$S\cdots S(\underset{\lambda\in[0,1]}{T}(\lambda, C_{\underset{\sim}{A}_\lambda}(\lambda_2))) = \underset{\lambda\in[0,\lambda_2)}{S\cdots S}(\lambda_2,\lambda)$$
$$\geq S(\lambda_2,\lambda_0) \geq S(\lambda_2,0) = \lambda_2$$

On the other hand:

$$S\cdots S(\underset{\lambda\in[0,1]}{T}(\lambda, C_{\underset{\sim}{A}_\lambda}(\lambda_2))) = \mu_{\underset{\sim}{A}}(\lambda_2) = \lambda_2$$

So $\lambda_2 \geq S(\lambda_2,\lambda_0) \geq \lambda_2$,

Namely: $S(\lambda_2,\lambda_0) = \lambda_2$;

Last, there is the following formula:

$\lambda_2 = S(\lambda_2,\lambda_0) \geq S(\lambda_0,\lambda_0) = \lambda_1$

Namely: $\lambda_2 \geq \lambda_1$.

The inequation is contradictory with $\lambda_2 < \lambda_1$, so the fore supposition is false. Thus, we get: $S(\lambda_0,\lambda_0) = \lambda_0$.

Therefore, we get: $\lambda_0 \geq \underset{\lambda<\lambda_0}{S}(\lambda_0,\lambda) \geq \lambda_0$,

Namely: $\underset{\lambda<\lambda_0}{S}(\lambda_0,\lambda) = \lambda_0$.

To sum up, the theorem is true. By the same means, we can get the following theorem.

Theorem4.2

(the Decomposition Theorem Ⅱ) fuzzy set A can be decomposed the following formula under T norm and S norm:

$$\mu_A(u_i) = \underset{\lambda\in[0,1]}{S}(T(\lambda, C_{A_{S\lambda}}(u_i))) \quad (9)$$

Its necessary and sufficient condition is: $S = \vee$.

5. THE REPRESENTATION THEOREM UNDER S NORM

Theorem5.1

(the Representation Theorem) Suppose H be an arbitrary nested sets of U , then

$$A = \underset{\lambda\in[0,1]}{S} \lambda H(\lambda) \quad (10)$$

is a fuzzy set of U , and $\forall \lambda \in [0,1]$, we have:

(1) $A_{S\lambda} = \underset{\alpha>\lambda}{S} H(\alpha)$

(2) $A_{\lambda} = \bigcap_{\alpha<\lambda} H(\alpha)$

Its necessary and sufficient condition is: $S = \vee$.

Proof

The sufficient condition is obvious**(Necessary Condition)**

Define $A = \underset{\lambda\in[0,1]}{S} \lambda H(\lambda)$,

then $\mu_{S(\lambda H(\lambda))}(u_i) = \underset{\lambda\in[0,1]}{S}(\mu_{\lambda H(\lambda)}(u_i))$,

Suppose $0 \le \lambda_1 < \lambda_2$,now we only need to proof $S(\lambda_1,\lambda_2) = \lambda_2$ for the equation (1) and (2) of theorem 5.1.

Owing to $S(\lambda_1,\lambda_2) \ge S(0,\lambda_2) = \lambda_2$, so we only need to prove the following equation is wrong: $S(\lambda_1,\lambda_2) > \lambda_2$.

(1)**(Reduction to Absurdity)**

Suppose: $S(\lambda_1,\lambda_2) > \lambda_2$ is true.

Choose λ' , so as to $S(\lambda_1,\lambda_2) > \lambda' > \lambda_2$;

choose H , so as to $u_i \in H(\lambda_2)$, $u_i \notin H(\lambda')$,

then $\mu_{\lambda_2 H(\lambda_2))}(u_0) = \lambda_2$.

From equation (1) of theorem5.1, we get:

$A_{S\lambda'} = \underset{\alpha>\lambda'}{S} H(\alpha)(u_0) = 0$

On the other hand, from

$\mu_A(u_0) = \mu_{\underset{\lambda\in[0,1]}{S} \lambda H(\lambda)}(u_0) \ge S(\lambda_1,\lambda_2) > \lambda'$

We gain: $A_{S\lambda'}(u_0) = 1$, so we get a contradictory equation: $0 = 1$, and it indicates that the fore assumption is false, so the necessary condition of equation (1) of theorem 5.1 is $S = \vee$.

(2) **(Reduction to Absurdity)**

Suppose: $S(\lambda_1,\lambda_2) > \lambda_2$ is true.

Choose λ' , so as to $S(\lambda_1,\lambda_2) > \lambda' > \lambda_2$

Choose H , so as to $u_i \in H(\lambda_2)$, $u_i \notin H(\lambda')$, then $\mu_{\lambda_2 H(\lambda_2))}(u_0) = \lambda_2$.

From equation (2) of theorem5.1, we get:

$C_{A_{\lambda'}}(u_0) = \underset{\alpha<\lambda'}{\cap} H(\alpha)(u_0) = 0$

On the other hand, from

$\mu_A(u_0) = \mu_{\underset{\lambda\in[0,1]}{S} \lambda H(\lambda)}(u_0) \ge S(\lambda_1,\lambda_2) > \lambda'$

We gain: $C_{A_{\lambda'}}(u_0) = 1$, so we get a contradictory equation: $0 = 1$, and it indicates that the fore assumption is false, so the necessary condition of equation (2) of theorem 5.1 is $S = \vee$.

To sum up, theorem 5.1 is proved.

6. CONCLUSIONS

The necessary and sufficient condition of the decomposition theorem under the general triangular norms operators is that the “or” operator must be max, but the “and” operator may be an arbitrary operator of the triangular norms; on the other hand, for the representation theorem under S norm, we get the following result: the necessary and sufficient condition of the two equations from the representation theorem under norm are all that the “or” operator must be max.

REFERENCES

[1] Zadeh L A. Fuzzy Sets. Information and Control. 1965, 8: 338~353

[2] Chen shuili, LI jinggong, Wang xianggong. Fuzzy Sets Theory and its Application. Beijing: Science Press, 2005

[3] Luo chengzhi. Fuzzy Sets Theory. Beijing: Beijing Normal University Press, 2003

[4] Gu wenxiang, Zhou jun. A New Resolution Theory of Fuzzy Sets. Journal of Northeast Normal University, 1995(2):6~7

[5] Xie jijian, Liu chengping. Fuzzy Mathematics’s Methods and Its Applications (Second edition). Wuhan: the Publishing Company of Huazhong University of Science and Technology, 2005.04

[6] Hu baoqing. The foundation of Fuzzy Theory. Wuhan: Wuhan University Press, 2004.10

[7] Gao xinbo. Fuzzy Cluster Analysis and its Applications. Xi an: the Publishing Company of Xian Electron Technology University, 2004

[8] Yang lunbiao, Gao yingyi. The Fuzzy Mathematics’s Theory and its Application (Fourth edition). Guangzhou: the Publishing Company of South China University of Technology, 2005.06

[9] Wang peizhuang, Han liyan. The Application of Fuzzy Mathmatics. Beijing: Beijing Economy College Press, 1989

[10] Zhang wenxiu. Fuzzy Mathematics Theory. Xi an: Xi’an Jiaotong University Press, 1991

NUMERICAL STUDY OF THE INCREMENTAL UNKNOWNS METHODS ON FLOATED GRID

Zhu lijun
School of Information and Calculation North University for Ethnice
Yinchuan Ningxia China 750021
Email: zhulijun1995@sohu.com

ABSTRACT

Through numerical experiments of two-dimensional second-order Dirichlet boundary-value, The discretization of the article is more simply than the classical grid. the incremental unknowns is linear stationary of linear problem in this paper and [1], as well as the matricial framework of article is more simply than [2].

Keywords: Incremental Unknowns, Floated Grid, Dirichlet oundary-Value Problem, Central Differences.

1. INTRODUCTION

Incremental Unknowns (IUs) Methods is introduced in paper[3] at earliest, in which inertial manifolds equations have been solved by multigrid methods. And more detailed of introduction is in paper [4].Our objective in this article is to bring forward the numerical study suitable to examine two-dimensional second-order Dirichlet boundary-value problems, by means of the incremental unknowns methods when two levels of discretization are considered. This kind of discretization grid is called Floated Grid ,and incremental unknowns methods is linear stationary on floated grid. Moreover textual methods is more effective than the classical method .

The method is set up in three steps: First, the two-dimensional second-order Dirichlet boundary-value problems are discretized with central differences on the finest grid. Second, linear interpolation is used to recursively define the incremental unknowns(other than that of coarsest level). Third, the node value is solved by iterative schemes. We get the linear syste $(A)(X)=(b)$ with central differences on finest grid like in paper[2], which becomes $[A][X]=[b]$ where $[A]=S^t(A)S,\ [b]=S^t(b)$ by injecting incremental unknowns methods. Here S stands for transfer matrix from the incremental unknowns $[X]$ to primitive (nodal) unknown (X), ie.. $(X)=S[X]$. And the structure of the matrix S is lower triangular matrix.

In this paper we solve the linear system $[A][X]=[b]$ instead of the linear system $(A)(X)=(b)$ by the following iterative methods: the conjugate method, when $[A]$ is a symmetric positive definite matrix. If $[A]$ is a nonsymmetric positive definite and non-singular matrix, $[A]^t[A]$ is a symmetric positive definite matrix. Then the solution of linear systems $[A][X]=[b]$ is equivalent to $[A]^t[A][X]=[A]^t[b]$ which is solved by the conjugate method.

2. INCREMENTAL UNKNOWNS METHODS

Make use of the incremental unknowns methods solve partial differential equations is establishment foundation in multigrid methods, the different of divide grid and layering way caused dissimilarity of methods , the floated grid and the incremental unknowns are defined as fellows (see Figure2.1).

$$
\begin{array}{cccccccc}
\Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond \\
\times & \circ & \times & \circ & \times & \circ & \times & \circ \\
\Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond \\
\times & \circ & \times & \circ & \times & \circ & \times & \circ \\
\Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond \\
\times & \circ & \times & \circ & \times & \circ & \times & \circ \\
\Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond & \Delta & \Diamond \\
\times & \circ & \times & \circ & \times & \circ & \times & \circ
\end{array}
$$

Figure 2.1. Floated Grid, Coarse Girdpoint ($\Diamond$) and Fine Girdpoint ($\Diamond,\Delta,\times,\circ$) For N=4

Where $\Diamond \in G_H$ (coarse node), $\Delta,\times,\circ \in G_h / G_H$ (rest node), $H=2h$, $\Omega=(0,1)^2$.

Here ,we consider the two-dimensional second-order Dirichlet boundary-value problem

$$\delta\frac{\partial^2 u}{\partial^2 x}+\rho\frac{\partial^2 u}{\partial^2 y}+\theta\frac{\partial^2 u}{\partial x\partial y}+\alpha\frac{\partial u}{\partial x}+\beta\frac{\partial u}{\partial y}+\gamma u=f,\quad in\,\Omega=(0,1)^2 \tag{1}$$

$$u\,|_{\Gamma}=0, on\ \Gamma=\partial\Omega \tag{2}$$

Where $\delta,\rho,\theta,\alpha,\beta,\lambda$ are given constants.Furthermore, $\Delta_{xx},\Delta_{yy},\Delta_{xy},\Delta_x,\Delta_y$ are the standard centered finite difference operators approximating the derivatives $\frac{\partial^2}{\partial^2 x},\frac{\partial^2}{\partial^2 y},\frac{\partial^2}{\partial x\partial y},\frac{\partial}{\partial x},\frac{\partial}{\partial y}$,

Respectively, assume that $u_{m,n}(m,n=1,2,3,\cdots,2N)$ is the approximate value of function u at the nodes of the grid (m,n), where we designate by $f_{m,n}=f(mh,nh)$ the exact value of f and by $u_{m,n}\approx u(mh,nh)$,we set $h=\frac{1}{2N+1}$,then the equations (1)and(2) is discretized with central difference on the finest grid so that its finite difference approximate is

$$(\delta\Delta_{xx}+\rho\Delta_{yy}+\theta\Delta_{xy}+\alpha\Delta_x+\beta\Delta_y+\gamma I)u_{m,n}=f_{m,n},\quad m,n=1,2,3\cdots 2N \tag{3}$$

We show according to Fig2.1, assume that $u_{i,j},(i,j=1,2,3,\cdots,2N)$ is unknown of nodes,

where $u_{2i,2j}\in G_H$ $u_{2i,2j-1}u_{2i-1,2j}u_{2i-1,2j-1}\in G_h/G_H, i,j=1,2,\cdots,2N$

we define transparent

$$Z_f=U_f-R\bullet Y \tag{4}$$

Respectively, where $R:G_H\to G_h/G_H$ is operator of quadratic interpolation,

$U_f\in G_h/G_H$, $Y\in G_H$, we define that Z_f is incremental unknown, then R is

$$z_{2i,2j-1}=u_{2i,2j-1}-\frac{1}{2}(u_{2i,2j}+u_{2i,2j-2}),\quad u_{0,2j}=0, i,j=1,2,\cdots,N \tag{5}$$

$$z_{2i-1,2j} = u_{2i-1,2j} - \frac{1}{2}(u_{2i-2,2j} + u_{2i,2j}),\quad u_{2i,0} = 0, i, j = 1,2,\cdots;N \tag{6}$$

$$z_{2i-1,2j-1} = u_{2i-1,2j-1} - \frac{1}{4}(u_{2i,2j} + u_{2i,2j-2} + u_{2i-2,2j} + u_{2i,2j}),\quad u_{0,2j} = u_{2i,0}0, i, j = 1,2,\cdots,N \tag{7}$$

We defined linear system by (5)-(7),we have $[A][X]=[b]$ (8)

Where $[A]=S^t(A)S$, $[b]=S^t(b)$, $[X]$ is incremental unknown, S is lower triangular matrix, it is

$$S = \begin{pmatrix} I_{NN} & & & \\ S_{21} & I_{NN} & & \\ S_{31} & & I_{NN} & \\ S_{41} & & & I_{NN} \end{pmatrix}$$

Respectively,where $S_{21} = \frac{1}{2}(I_N \otimes B), S_{31} = \frac{1}{2}(B \otimes I_N), S_{41} = \frac{1}{4}(B \otimes B)$ we will denote by $A \otimes B$ the tensor product of the matrix A and B ,by I_N the identity matrix of order N,by B the matrix [12].

3. NUMERICAL EXPERIMENT

Taking two-dimensional second-order Dirichlet boundary-value problem for example, when the boundary-value is non-zero, the non-singular problem can be transferred to the singular broblem. Thereby the singular problem is considered here only. First, the method in this paper is to be testified effective and stable for the solving of

$$-\Delta u(x,y) + \alpha \frac{\partial u(x,y)}{\partial x} + \beta u(x,y) = f, (x,y) \in \Omega$$

$$u(x,y)|_{\Gamma} = 0, on\ \Gamma = \partial\Omega ,$$

whose three-dimensional comparative graph(Figure3.2)of exact solution and numerical solution is also shown. Second, we give the comparative graph of the numerical solution obtained through the method in this paper and classical method respectively, which illustrates that the method in this paper is more stable than the classical method for the solution of linear problem. Finally, the comparative figure (Figure3.3) of iterative error and iteration number of incremental unknown method which is obtained on the two different discrete grids demonstrates the feasibility of the method in this paper.

The linear system $[A][X]=[b]$ which are discretized from the problem above are solved by conjugate gradient method.

For convenience of comparison, in the numerical experiment suppose $u(x,y) = \sin(x(1-x)y(1-y))$

The linear equations as follows are considered:

$$-\Delta u(x,y) + \alpha \frac{\partial u(x,y)}{\partial x} + \beta u(x,y) = f,\quad n\ \Omega = (0,1)^2 \tag{9}$$

$$u(x,y)|_{\Gamma} = 0, on\ \Gamma = \partial\Omega \tag{10}$$

where Δ is Laplace operator.Equ.(9) and Equ.(10) are central differential discretized by dividing grid and hierarchical division showed by Fig 2.1. We get the linear system as follows:

$$(A)(X) = (b) \tag{11}$$

In the linear system (11), $(A) = (\Delta) + \alpha(\Delta_x) + \beta I$, where I is identity matrix of $4 \times N \times N$, and (Δ_x) and (Δ) are given by:

$$(\Delta) = \frac{1}{h^2}\begin{pmatrix} -4I_{NN} & I_N \otimes B & I_N \otimes B & 0 \\ I_N \otimes B^t & -4I_{NN} & 0 & B \otimes I_N \\ I_N \otimes B^t & 0 & -4I_{NN} & B \otimes I_N \\ 0 & B^t \otimes I_N & B^t \otimes I_N & -4I_{NN} \end{pmatrix}$$

$$(\Delta_x) = \frac{1}{2h}\begin{pmatrix} 0 & I_N \otimes B_1 & 0 & 0 \\ I_N \otimes (-(B_1)^t) & 0 & 0 & 0 \\ 0 & 0 & 0 & B_1 \otimes I_N \\ 0 & 0 & -(B_1)^t \otimes I_N & 0 \end{pmatrix}$$

Where I_N is identity matrix of N , I_{NN} is identity matrix of $N \times N$,and B and B_1 are given by:

$$B = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}_{NN}$$

$$B_1 = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{NN} \tag{12}$$

The linear system(8) are obtained by introducing incremental unknown from linear system(11).Figure3.1 shows the comparison of numerical solution of the linear system(8) which is get by conjugate gradient method and exact solution of Equ.(9) and Equ.(10)

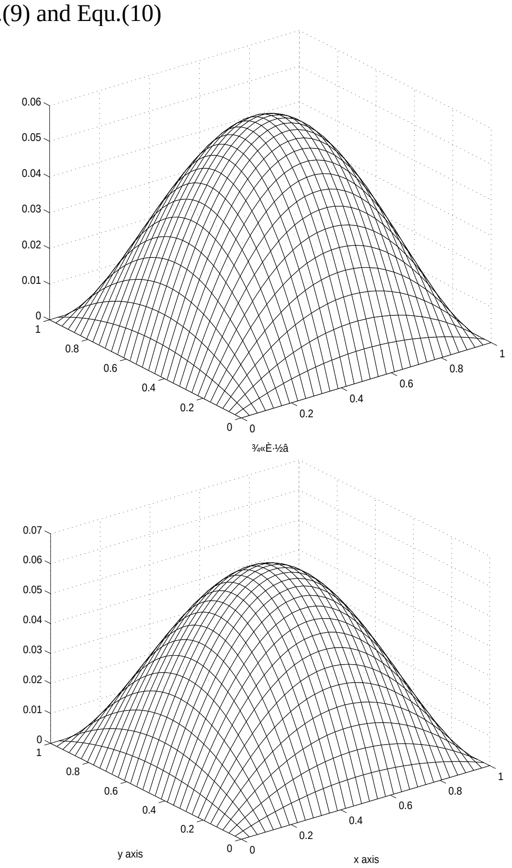


Figure 3.1. Numerical Solution and Exact Solution(N=15)

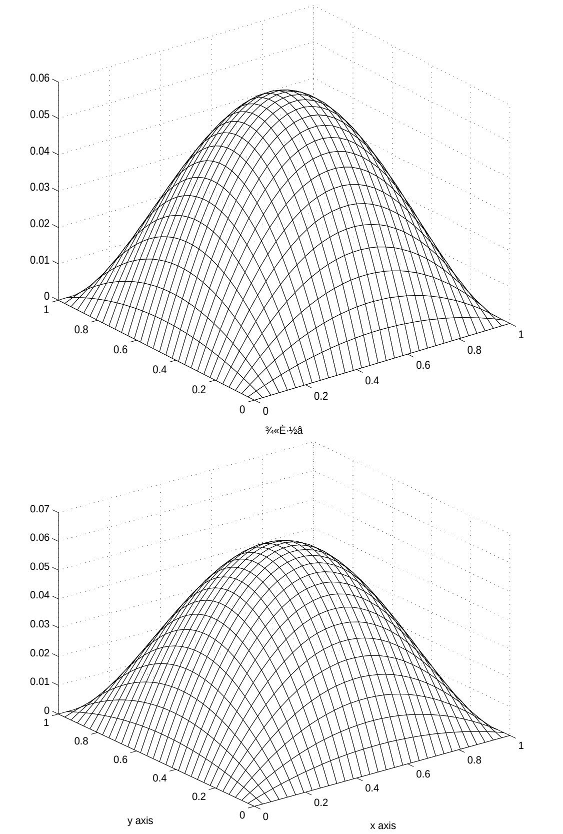


Figure 3.2. Comparison Chart of Numerical Solutions on Method in This Article and Paper (1)

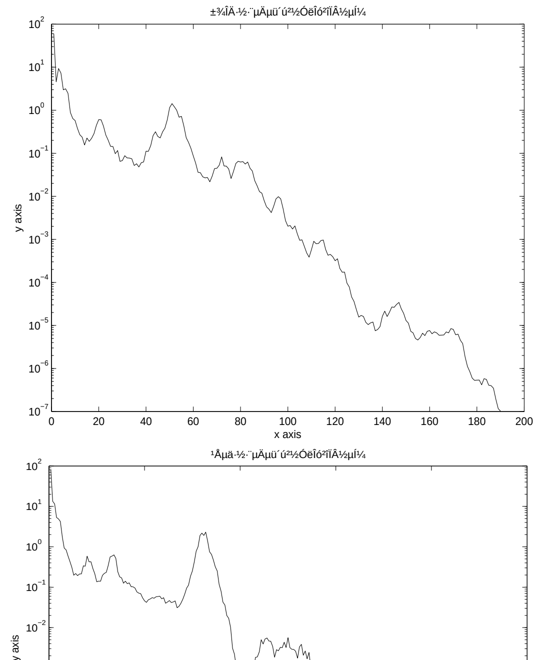


Figure 3.3 . Comparison Chart of Iterative Error and Iteration Number on Method in This Article and Paper(1)

4. CONCLUSIONS

It is illustrated that both the incremental unknowns methods which are defined on discrete grid and classic grid respectively are linear stationary and feasibility for the equation

$$-\Delta u(x,y)+\alpha\frac{\partial u(x,y)}{\partial x}+\beta u(x,y)=f,(x,y)\in\Omega$$

$$u(x,y)\,|_{\Gamma}=0, on\ \Gamma=\partial\Omega\ .$$

However, the numerical solutions near the maximum point is not smooth enough when solving the same problem by the classical method. The method in this article avoids the problem by contrast, which indicates this method is more stable. The Fig 3.3 showed that the method in this paper is superior over the classical method on iteration speed by numerical computing in the case of the same error of the same problem.

In conclusion, the method in this paper is linear stationary and computational feasibility for the two-dimensional second-order Dirichlet boundary-value problem. And it is superior over the classical method by comparison.

REFERENCES

[1] S.Garcia, Numerical study of the incremental unknowns method, Numer.Methods. Partial Differential Equa.10 (1994)103-127.

[2] S.Garcia, The matricial framework for the incremental unknowns method, Numer. Functional. Optimiz.14, (1993)25-56.

[3] R.Temam,Inertial manifords and multigrid methods, SIAM J.Math.Anal.21(1)(1990)154-178.

[4] M.Chen, R.Temam,Incremental unknowns for solving partial di®erential equations, Numer.Math.59(3)(1991) 255-271.

[5] W.Hackbush, Multigrid Methods and Applications, Springer-Verlag,Berlin,1983.

THIRD-ORDER DIRECTIONAL NEWTON METHOD FOR MULTIVARIABLE EQUATIONS*

Kou Jisheng [1], Wang Xiuhua[2]
[1]Department of Mathematics, Shanghai University
Shanghai 200444, China
[2]Department of Mathematics, Xiaogan University
Xiaogan 432100, Hubei, China
E-mail: koujisheng@yahoo.com.cn

ABSTRACT

We present a variant of the directional Newton method for solving a single nonlinear equation in several variables. Under suitable assumptions, we prove the cubic convergence speed of this new method. This method is suitable for parallel implements. The related parallel algorithms are discussed. Numerical examples show that the new method is feasible and efficient, and has better numerical behavior than the directional Newton method.

Keywords: Parallel numerical algorithm, Nonlinear equations, Directional Newton method, Iterative method

1. INTRODUCTION

In this paper, we consider the iterative methods for finding the solution of nonlinear equation in *n* variables:

$$f(x)=0 \tag{1.1}$$

where $f : R^n \to R^1$ is a continuously differentiable nonlinear mapping.

Solving such multivariable equations is a very important problem in scientific and engineering computing areas [2-4]. In computer graphics, we often need to compute and display the intersection $\mathcal{C} = \mathcal{M} \cap \mathcal{N}$ for two surfaces $\mathcal{M}$ and $\mathcal{N} \in R^3$. Specifically, if the two surfaces are explicitly given by

$$\mathcal{M} = \{(x_1,x_2,x_3)^T \mid x_3 = f_1(x_1,x_2)\}, \tag{1.2}$$

$$\mathcal{N} = \{(x_1,x_2,x_3)^T \mid x_3 = f_2(x_1,x_2)\}, \tag{1.3}$$

then the point $x^* = (x_1^*, x_2^*, x_3^*)^T \in \mathcal{C}$ must satisfy the nonlinear equation

$$f_1(x_1^*, x_2^*) = f_2(x_1^*, x_2^*), \tag{1.4}$$

and $x_3^* = f_1(x_1^*, x_2^*)$. This results in the solution of a nonlinear equation in two variables of the form

$$f(x_1,x_2) = f_1(x_1,x_2) - f_1(x_1,x_2) = 0\,. \tag{1.5}$$

In mathematical programming, for an equality-constraint optimization problem, e.g.,

$$\min \varphi(x), s.t. f(x)=0 \tag{1.6}$$

where $\varphi, f : R^n \to R^1$ are nonlinear functions, we may need to seek a feasible point to start a numerical algorithm. This requires us to compute a solution of the nonlinear equation $f(x) = 0$ [5].

* **Kou Jisheng** is corresponding author.
This work is supported by National Natural Science Foundation of China (No. 10826082), Key Disciplines of Shanghai Municipality (No. S30104) and Shanghai Leading Academic Discipline Project (No. J50101).

In order to solve a system of nonlinear equatins $F(x) = 0$, with $F : R^n \to R^n$, we may turn to solve the reduced nonlinear equation

$$\|F(x)\|_2 = 0\,, \tag{1.7}$$

if the zero point of the nonlinear function *F* is isolated or locally isolated and if the rounding error is neglected.

In [2], the authors studied the following directional Newton (DN) method for solving such nonlinear equations. The iterative formula of the DN method is given by

$$x^{k+1} = x^k - \frac{f(x^k)}{\nabla f(x^k)\cdot d^k} d^k \tag{1.8}$$

where

$$\nabla f(x^k) = \left(\tfrac{\partial f(x^k)}{\partial x_1}, \tfrac{\partial f(x^k)}{\partial x_2}, \cdots, \tfrac{\partial f(x^k)}{\partial x_n}\right)^T \tag{1.9}$$

is the gradient of *f* and d^k is a direction at x^k. Under certain assumptions, they proved the convergence and demonstrated the quadratic convergence speed of this DN-method for the certain choices of the direction vector d^k.

However, it is expensive to compute the gradient of *f* at each iteration, especially when the size of the problem is very large. We all know that the third-order methods for the case of nonlinear equations have the higher computational efficiency, for example, the methods proposed in [6-9].

In this paper, we present a variant of the directional Newton method for multivariable equations. The new method requires two evaluations of the functions and one of the gradient of *f*, but the order of convergence attains three. This method is suitable for parallel implements. The related parallel algorithms are discussed. Numerical examples show that the new method is feasible and efficient, and has better numerical behavior than the directional Newton method.

2. THE NEW METHOD

The third-order method for systems of nonlinear equations proposed in [6] can reduce the cost of the Jacobian matrix, and therefore has the higher computational efficiency. In order to reduce the times of computing the gradient of *f*, by making use of the ideas in [6], we introduce a new method

$$\begin{cases} y^k = x^k - \dfrac{f(x^k)}{\nabla f(x^k)\cdot d^k} d^k \\ x^{k+1} = y^k - \dfrac{f(y^k)}{\nabla f(x^k)\cdot d^k} d^k \end{cases} \tag{2.1}$$

For the new method, corresponding to the two choices of the directions d^k in the DN-method [2], we have the following strategies:

(i) d^k is a direction sufficiently close to the gradient $\nabla f(x^k)$ of $f(x)$ at the current iterate x^k,

(ii) $d^k = e^{m(k)}$, where $e^{m(k)}$ is the $m(k)$th unit vector in R^n, and $m(k)$ is chosen such that the $m(k)$th component of $\nabla f(x^k)$ has the maximal modulus.

For the new method given by Eq.(2.1), we have the following theorem.

Theorem 1. Let the function $f: D \subset R^n \to R^1$ be continuously differentiable, x^* be the root of $f(x)$ and $\nabla f(x^*) \neq 0$. Define $\lambda = \|\nabla f(x^*)\|$. Furthermore, assume that there exists a positive number L such that for any $x \in D$

$$\|\nabla f(x) - \nabla f(x^*)\| \leq L\|x - x^*\| \tag{2.2}$$

and d^k is chosen to satisfy

$$\left|\nabla f(x^k) \cdot (y^k - x^*)\right| \geq \alpha \|\nabla f(x^k)\| \|y^k - x^*\| \tag{2.3}$$

$$\left|\nabla f(x^k) \cdot (x^{k+1} - x^*)\right| \geq \alpha \|\nabla f(x^k)\| \|x^{k+1} - x^*\| \tag{2.4}$$

where $0 < \alpha \leq 1$, then there exists a set S such that for any $x^0 \in S$, the sequence $\{x^k\}$ produced by (2.1) is well defined in S and converges to x^*. Moreover, the order of the method (2.1) is three.

Proof. Without loss of generality, we use the Euclidean norm as the norm in the following. Let $S = \{x \mid \|x - x^*\| \leq \delta\}$, where $\delta = \frac{\alpha\lambda}{2L}$.

For $x \in S$, by (2.2) we have

$$\|\nabla f(x) - \nabla f(x^*)\| \leq L\delta \leq \tfrac{1}{2}\lambda. \tag{2.5}$$

So we obtain that

$$\|\nabla f(x)\| \geq \|\nabla f(x^*)\| - \|\nabla f(x) - \nabla f(x^*)\| \geq \tfrac{1}{2}\lambda.$$

It can be concluded that for $x \in S, \nabla f(x) \neq 0$.

From Eq. (2.2), for $x^0 \in S$, we can get [1]

$$\|f(x^*) - f(x^0) - \nabla f(x^0) \cdot (x^0 - x^*)\| \leq \tfrac{1}{2}L\|x^0 - x^*\|^2. \tag{2.6}$$

Furthermore, by Eq. (2.2) we have

$$\begin{aligned} &\|\nabla f(x^0) \cdot (y^0 - x^*)\| \\ &= \|f(x^*) - f(x^0) - \nabla f(x^0) \cdot (x^0 - x^*)\| \\ &\leq \tfrac{1}{2}L\|x^0 - x^*\|^2. \end{aligned} \tag{2.7}$$

So we obtain from Eq. (2.3) that

$$\|y^0 - x^*\| \leq \tfrac{L}{\alpha\lambda}\|x^0 - x^*\|^2 \leq \tfrac{1}{2}\|x^0 - x^*\| \tag{2.8}$$

This implies that $y^0 \in S$.

Now we consider x^1. It is obtained that

$$\begin{aligned} &\|\nabla f(x^0) \cdot (x^1 - x^*)\| \\ &\leq \|f(y^0) - f(x^*) - f(x^0) - \nabla f(x^*) \cdot (y^0 - x^*)\| \\ &+ \|\nabla f(x^0) - \nabla f(x^*)\| \|y^0 - x^*\| \\ &\leq \tfrac{1}{2}L\|y^0 - x^*\|^2 + L\|y^0 - x^*\| \|x^0 - x^*\|. \end{aligned} \tag{2.9}$$

Thus we get by Eq. (2.4) that

$$\begin{aligned} &\|x^1 - x^*\| \\ &\leq \tfrac{L}{\alpha\lambda}\|y^0 - x^*\|^2 + \tfrac{2L}{\alpha\lambda}\|y^0 - x^*\| \|x^0 - x^*\| \\ &\leq \tfrac{5}{2}\left(\tfrac{L}{\alpha\lambda}\right)^2 \|x^0 - x^*\|^3 \\ &\leq \tfrac{5}{8}\|x^0 - x^*\| \end{aligned} \tag{2.10}$$

It holds that $x^1 \in S$.

By induction we obtain that for any $x^0 \in S$, the sequence $\{x^k\}$ produced by Eq. (2.1) is well defined in S and satisfies

$$\|x^{k+1} - x^*\| \leq \tfrac{5}{2}\left(\tfrac{L}{\alpha\lambda}\right)^2 \|x^k - x^*\|^3 \tag{2.11}$$

$$\|x^{k+1} - x^*\| \leq \tfrac{5}{8}\|x^k - x^*\| \tag{2.12}$$

By making use of

$$\|x^k - x^*\| \leq \left(\tfrac{5}{8}\right)^k \|x^0 - x^*\| \tag{2.13}$$

we obtain that when $k \to \infty, x^k \to x^*$. Moreover, we know by Eq. (2.11) that the order of the method (2.1) is three. This ends the proof.

Remark. Let $\angle(\cdot,\cdot)$ be the angle between two vectors. The conditions (2.3) and (2.4) are equivalent to say that the direction d^k is chosen such that

$$\angle(\nabla f(x^k), y^k - x^*) = \frac{|\nabla f(x^k) \cdot (y^k - x^*)|}{\|\nabla f(x^k)\| \|y^k - x^*\|} \geq \alpha \tag{2.14}$$

$$\angle(\nabla f(x^k), x^{k+1} - x^*) = \frac{|\nabla f(x^k) \cdot (x^{k+1} - x^*)|}{\|\nabla f(x^k)\| \|x^{k+1} - x^*\|} \geq \alpha \tag{2.15}$$

3. PARALLEL IMPLEMENT

In this section, we consider the implement of the new method in parallel computing systems. Assume that there exist $p+1$ computer processors. We choose one processor, which number is signed as P_0, to compute the values of $f(x^k)$ and $f(y^k)$, and distribute n variables on the rest p processors which numbers are signed from P_1 to P_p. At this moment, x, d and $\nabla f(x^k)$ can be divided into p parts. Each part contains at most m components where $m = [n/p]$. The ith part of x, d and $\nabla f(x^k)$ are defined by

$$x_{Pi} = (x_{im-m+1}, \cdots, x_{im})^T, \tag{3.1}$$

$$d_{Pi} = (d_{im-m+1}, \cdots, d_{im})^T, \tag{3.2}$$

and

$$\nabla f(x^k)_{Pi} = \left(\frac{\partial f(x^k)}{\partial x_{im-m+1}}, \cdots, \frac{\partial f(x^k)}{\partial x_{im}}\right)^T.$$

For the ith processor, the iterative formula given by Eq. (2.1) becomes

$$y_{Pi}^k = x_{Pi}^k - \frac{f(x^k)}{\nabla f(x^k) \cdot d} d_{Pi}^k \tag{3.3}$$

$$x_{Pi}^{k+1} = y_{Pi}^k - \frac{f(y^k)}{\nabla f(x^k) \cdot d} d_{Pi}^k \tag{3.4}$$

The parallel algorithm for the method (2.1) can be described as following:

Step 1: Give the initial approximation $x^{(0)}$.

Step 2: For $k = 0,1,2,\cdots$ until convergence

Step 2.1: The P_0 processor compute $f(x^k)$, and then the rest processors parallelly compute $\nabla f(x^k)_{Pi}$, d_{Pi}^k and y_{Pi}^k by (3.3) to get y^k.

Step 2.2: The P_0 processor compute $f(y^k)$, and then the rest processors parallelly compute x_{Pi}^{k+1} by (3.4) to get x^{k+1}.

4. NUMERICAL TESTS

In this section, we present some numerical results for the new method and compare it with DN method on their numerical behaviors. For two methods, we take $d^k = \nabla f(x^k)$. All computations are carried out with double arithmetic precision.

Example 1. $f(x) = \sum_{i=1}^{n} |\sin(x_i)|$ with the exact solution $x^* = 0$.

In example 1, we take n = 100, $x^{(0)} = (1, L, 1)^T$ for two methods. The absolute values of f at each iteration computed by two methods are shown in Table 1.

Table 1. Results of Example 1

k	DN	New
1	0.5290	0.4093
2	0.0659	0.0026
3	9.5722e-5	1.9015e-14
4	2.9236e-13	

Example 2. $f(x) = \sum_{i=1}^{n} |x_i| e^{1+x_i}$ with the exact solution $x^* = 0$.

In example 2, we take n = 100. Let $s = (1, L, 1)^T$. We choose the two initial values $x^{(0)} = s$ and $x^{(0)} = 1.5s$ respectively for two methods. The absolute values of f at each iteration computed by two methods are shown in Tables 2 and 3.

From the numerical results displayed in Tables 1-3, we can see that the new method converges faster than the DN method and therefore the computational efficiency is improved. This means that the new method is feasible and efficient, and has better numerical behavior than the directional Newton method.

Table 2. Results of Example 2 with $x^{(0)} = s$

k	DN	New
1	2.2408	1.3416
2	0.5352	0.1086
3	0.0663	2.7882e-4
4	0.0015	5.8636e-12
5	8.3295e-7	5.4572e-35
6	2.5524e-13	

Table 3. Results of Example 2 with $x^{(0)} = 1.5s$

k	DN	New
1	6.0173	3.8544
2	1.7749	0.5748
3	0.3934	0.0196
4	0.0397	1.9689e-6
5	5.5589e-4	2.0659e-18
6	1.1361e-7	
7	4.7483e-15	

Time performance is also important for the methods. The computational time of two methods is compared for examples 1 and 2 where n = 100000. The time consumption of the two methods stopped when $|f(x^k)| < 10^{-13}$ is listed in Table 4. Second is used as the time unit.

From Table 4, we can see that the new method costs less than the DN method.

Table 4. Computational time of two methods

Example	$x^{(0)}$	DN	New
1	$(1, \cdots, 1)^T$	0.165118	0.150641
2	s	0.350914	0.338974
2	$1.5s$	0.355041	0.333102

5. CONCLUSIONS

In order to reduce the cost of computing the gradient of f per iteration, we present a variant of the directional Newton method for solving a single nonlinear equation in several variables. Under suitable assumptions, we prove the cubic convergence of the new method.

We discuss the parallel implements of new method. The new method can run on the parallel computing systems. Numerical examples show that the new method is feasible and efficient, and has better numerical behavior than the directional Newton method.

REFERENCES

[1] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equation in Several Variables,Academic Press, New York, 1970.

[2] Y. Levin, A. Ben-Israel, “Directional Newton method in n variables”, Math. Comput. Vol.71, No. 237, 2001, pp. 251-262.

[3] G. Lukt´acs, The generalized inverse matrix and the surface-surface intersection problem, in: W. Strasser, H.-P. Seidel (Eds.), Theory and Practice of Geometric Modeling, Springer, Berlin, 1989, pp. 167-185.

[4] H. An, Z. Bai, “Directional secant method for nonlinear equations”, J. Comput. Appl. Math., Vol.175 , 2005, pp. 291-304.

[5] B.T. Polyak, Introduction to Optimization, Optimization Software, Inc., Publication Division, NewYork, 1987.

[6] F.A. Potra, V. Pt´ak, Nondiscrete induction and iterative processes, in: Research Notes in Mathematics, Vol.103, Pitman, Boston, 1984.

[7] J.R. Sharma, A composite third order Newton-Steffensen method for solving nonlinear equations, Appl. Math. Comput., Vol.169, 2005, pp. 242-246.

[8] H.H.H. Homeier, On Newton-type methods with cubic convergence, J. Comput. Appl. Math., Vol.176, 2005, pp. 425-432.

[9] J. Kou, Y. Li, X. Wang, Third-order modification of Newton’s method, J. Comput. Appl. Math., Vol. 205, 2007, pp. 1-5.

Kou Jisheng is a lecturer in department of mathematics, Shanghai University. His research interests are in numerical methods for nonlinear problems and parallel numerical algorithms.

Wang Xiuhua is a lecturer in department of mathematics, Xiaogan University. Her research interests are in numerical methods and parallel numerical algorithms.

AN IMPROVED ELITIST STRATEGY FIGHTING PREMATURE CONVERGENCE

Ming Zhou, Qingping Guo
Department of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei 430063.P.R.China
Email: zhouming186@yahoo.cn; qpguo@mail.whut.edu.cn

ABSTRACT

With the analysis of the premature convergence problem in Genetic Algorithm and the Elitist Strategy, an Improved Elitist Strategy is proposed in the paper. When the proportion in number of some individual is too high in the population, the Improved Elitist Strategy, which stores the rejected individual in the other population, can injects new individual into the population to fighting premature convergence. But the new strategy may slow down the speed of convergence, when the initial population size is too large. And the paper gives the future directions to the research about the Improved Elitist Strategy.

Keywords: Genetic algorithm, Elitist strategy, Select, premature convergence, Improved elitist strategy

1. INTRODUCTION

Genetic Algorithm (GA), which is based on Darwin' organic evolution theory "survival of the fitness" and Mendel's genetic theory [1], is first proposed by Holland, a professor of Michigan university in American, in the 1970s and developed it. GA is a good method to deal with complex and nonlinear problems, such as combinatorial optimization, pattern identification and image processing. Now GA is well developed and used in many fields, such as combinatorial optimization and artificial intelligence. GA consists of four main parts: coding mechanism, controlling parameters, fitness function, and genetic operators. AS the developing of GA, there are many concrete methods to carry out the four parts above. The central part, genetic operators, has three operators: select crossover and mutation. The paper is mainly discussing the select operator. And the Improved Elitist Strategy is used with a concrete select method.

2. THE FUNDAMENTAL PRINCIPLE OF GA

The fundamental principle of GA is Darwin's organic evolution theory "survival of the fitness" and Mendel's genetic theory. "Every species appears better adaptive to the nature in its evolutionary process. The basic characteristics of the individual in some species are inherited by its next generations. But, the next generations are not the simple copy of the individual. They have their own characteristics." Described in Darwin' organic evolution theory [2], [3]. Accordingly, Mendel's genetic theory [2], [3], "Heredity, as a genetic instructions, is packaged in each cell, and it is included in the chromosome by the form of genes. Every gene has many loci, and each gene locus controls a particular character. Each gene from the individual results a certain degree of adaptability to the environment. The crossover and mutation of gene may give birth to next generations with stronger adaptability to the environment. Through the survival of the fittest natural selection, genetic structure with high adaptive value is preserved."

Holland combined two theories above, and created the Genetic Algorithm [1,2,3]. With the help of computer programming, GA describe the unsolved problems as strings (or chromosome), which is binary digital code or integer string. In this way, the algorithm creates a group of string, and put the strings in the problem solving environment. According to the survival of the fittest theory, stings with high adaptive value are selected, and create next generation of higher adaptive value through two genetic operators: crossover and mutation. As the evolution goes on, the algorithm converges to a string with the highest adaptive value, and the string is the best solution to the unsolved problem.

3. THE SELECTION STRATEGY IN GA

The Simple Genetic Algorithm (SGA) contains three main operators: selection, crossover, mutation. And the selection operator in GA is the basis of the other two operators: the selection operator provides regenerate chromosomes for the followed operators (crossover and mutation), and provides a basis for simulating of the biological evolution survival principle, "survival of the fitness ". So, since GA was proposed, many different selection strategies is put forward: for example, classic research results: Brindle's study of selection and disposal issues in GA in1981, and Baker's experiment basis on sorting selection method EST. Sexual selection and dipole selection [4], which reflect the trend of more intelligent.

3.1 Some Common Selection Methods

3.1.1 Roulette Wheel Selection [4],[5]

Roulette wheel selection is a common allocation method based on the fitness ratio; it is also called Monte Carlo method. The method uses the ratio of individual fitness to decide the possibility of its generation to retain. If there is a individual i, and its fitness is f_i, then its probability of being selected is: $pro_i = f_i \Big/ \sum_{i=1}^{popsize} f_i$.

Roulette wheel selection's concept is: Divide the wheel into n pieces. The size of the fan-shaped area i is in proportion to individual i's fitness value. The wheel has one pointer. Rotate the wheel one time, we can select one individual. So we need to rotate the wheel n times (the population size is n).

3.1.2 Stochastic Universal Sampling Selection [4],[5]

Stochastic universal sampling selection is improved on the roulette wheel method, it is also an allocation method based on the fitness ratio. Like roulette wheel selection, divide the wheel into n pieces. The size of the fan-shaped area i is in proportion to individual i's fitness value. The difference is that this wheel has n pointers, so rotate the wheel one time, we can select n individuals.

This approach appears the same with roulette wheel selection, but they are two totally different types of selection operator. Stochastic universal sampling selection overcomes some disadvantages of roulette wheel selection, for example, its controlling of selection pressure and biodiversity loss is better than roulette wheel selection. The method also provides zero bias and the best individual expansion.

3.1.3 Tournament Selection [6],[7]

In the tournament selection method, select a certain number (tour) of individuals from the population randomly, and then choose the individual with highest fitness value to be a parent. Repeat the process until we get n individuals (the population size is n). Tournament selection method has a parameter called the size of Tour, and the Tour's value range from 0 to n. Generally, we set the tour equal to 2.

3.1.4 Expectation Selection [5],[7]:

Expectation selection is also called non-playback randomly selection. Random selection operator is computing according to each individual's expectation in the next generation. Its procedure is as follows:

(1) Compute the survival expectation value of each individual: $N_i = M \bullet F_i \Big/ \sum_{i=1}^{M} F_i$

(2) If a individual has been selected to participate in the crossover operator, minus the survival expectation value of the individual with 0.5. If the individual hasn’t been selected to participate in the crossover operator, minus the survival expectation value of the individual with 1.0.

(3) As the selection process goes on, if one individual’s survival expectation value is less than 0, the individual has no chance to be selected.

The approach can increase the opportunity, which the best individual is selected, but it is very complex to be carried out.

3.2 Analysis of the Common Selection Operator

If we use the random method (roulette wheel selection) GA will degenerate to random search. The individual with the optimal fitness value in the population can not be retained, so the process of looking for the optimal solution is not monotonic decreasing, and then the algorithm may not find the optimal solution. If we use the method that the fitness is in proportion to the individual (expectation selection): Sorting the individuals according to their fitness value, and select the individual sorted in the front of the queue. In this way, the fitness value of the individuals in the population are very close, so the searching range is very limited, and the optimal solution found by the algorithm may be the local optimal solution.

In order to solve the problem, the Elitist Strategy is proposed.

3.3 Elitist Strategy [8]

In the process of SGA, the best current individual may be broken, because of the randomly selection, crossover and mutation. That will significantly reduce the efficiency for SGA to search the optimal solution. Moreover, it also reduces the average fitness of the population, and has negative impact on the operating efficiency and the convergence of the algorithm.

So we use Elitist Strategy to solve the problem above. Its concept is: Store the optimal individual, and let it avert crossover and mutation, so we can prevent the optimal individual's fitness from being broken.

3.4 Reasons for Premature Convergence [8]

Although the elitist strategy is an important condition to ensure the convergence of genetic algorithm, but the strategy do not effectively solve the problem of premature convergence.

After some generations, because the selecting opportunity in selection strategy is in proportion to the adaptive value, and the role of crossover operator and mutation operator, it is easy to lead to the following questions:

(1) The super-individual problem: In the initial stage of evolution, some individual with high adaptive value is generated. After several generations, many locus in the population may be taken by the individual and its generations.

(2) The closed competition problem: If individuals in the population have similar adaptive value, and the values are all great, the crossover operator will reduce the effect of the search. The algorithm can not effectively guide the search, so it is very difficult to find the optimal value. These questions will make the population with limited range, and the search may stop at the immature stage.

So, the introduction of mutation operator in GA will make the search process out of the closed competition. In fact, if the individuals in the current population have high adaptive value, then the opportunity to get individuals outside of the population through mutation operator, which generates individual with higher adaptive value, is very small. Under the selection mechanism basis on the sorting according to fitness, the individual almost has no chance to be a parent and to participate in the crossover operator. In this way, crossover operation is still fulfilled by the individuals in the closed population, and algorithm is equivalent to use individuals generated by mutation operator to conduct ineffective random search.

4. IMPROVED ELITIST STRATEGY

Ultimately, the only way to solve the problem of premature convergence is to extend individuals in the closed population. But once the population is initialized, a closed space is formed; the selection operator and the mutation operator are fulfilled in that closed space. The reason for the premature convergence is that the closed space is seized by individual with high adaptive value. In order to solve the problem, we need to inject new individuals into the closed space, or to store rejected individuals as the other choice to extend the closed space.

The concept of the improved elitist strategy: We can improve the elitist strategy in this way: We can store the rejected individuals as new individuals outside of the closed space as well. After some generations (for example 10 generations), detect retained individuals in the population. If it is found that a individual with high fitness is mass reproduction, or fitness of the individuals retained in the population has little difference, we can use new individuals stored outside of the closed space to replace some of the individuals in the closed space, to increase the diversity of individual. In this way, we can avoid the premature convergence problem effectively. And its procedure is:

(1) In the Improved Elitist Strategy, we create another population called reserved population(r-population). The size of the reserved population can be the same as the population. And we store the rejected individuals in the r-population. In the r-population, individuals are sorted by their fitness value, and the individuals have different fitness values.

(2) After every select operation, store the rejected individuals in the r-population. But the size of r-population is limited; it can not store all of the rejected individuals. When there is no room in the r-population., we need a replacement mechanism: ①The individuals in the r-population have different fitness values. If the fitness value of the rejected individual is equal to one individual stored in the r-population, the rejected individual will not be stored. ②If the fitness value of the rejected individual is smaller than one individual stored in the r-population, the rejected individual will not be stored, either. ③When the fitness value of the rejected individual is larger than one individual stored in the r-population; the rejected individual will replace the last individual sorted in the r-population. Then sort the individuals in the r-population again.

(3) On the end of the select operation, we will add a detect operator before the crossover operator. The detect operator is used to detect retained individuals in the population. If it is found that a individual with high fitness is mass reproduction, or fitness of the individuals retained in the population has little difference, we can use new individual stored in the r-population to replace some of the individuals in the population, to increase the diversity of individual.

As the Improved Elitist Strategy is not a concrete selection operator, so it must be used with some selection operators. And the paper gives the procedure figure1:
GA with Improved Elitist Strategy:

5. FRUTURE TO THE RESERCH

Because the time is limited, the paper does not prove the improved elitist strategy. So here are the future works the research:

(1) There is a parameter in the strategy to determine whether the percentage of one individual is to large (in figure1, the parameter is equal to 30%). I think we need to do some experiment to make the parameter more precision.

(2) We should use experiment data to prove the advantage of the new strategy.

(3) Figure out more effective replacement mechanism, we can also use selection operator in this filed.

(4) The new strategy will use more space for r-population, and it will slow down the convergence speed. So, we should find a best size for r-population.

6. CONCLUSIONS

The paper gives a new strategy: Improved Elitist Strategy. In theory, it can fight premature convergence effective. But the paper does not prove it. And, I give the future directions of the research. I will continue the research to consummate the new strategy.

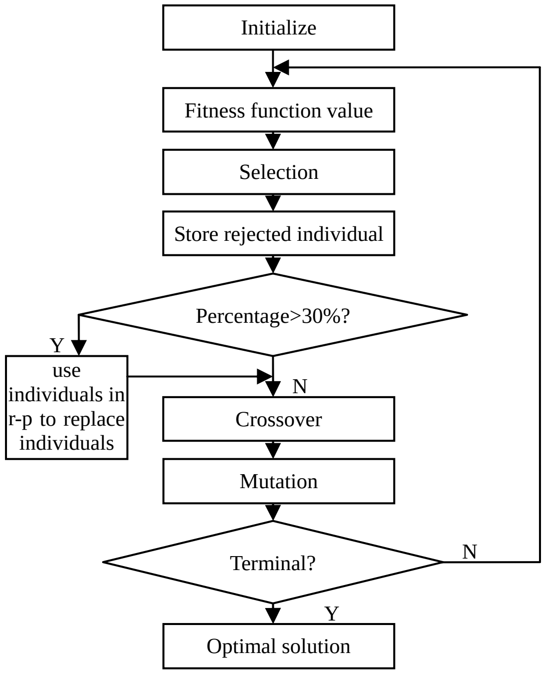


Figure1. GA with Improved Elitist Strategy

REFERENCES

[1] Wang XiaoPing, Cao LiMing "Genetic AlgorithmTheory And Application to Software".[M],Xi'an Jiao Tong University press.2004
[2] Dai XiaoHui, Li MinQiang, Kou JiSong, "Survey on the Theory of Genetic Algorithms". Journal of Control And Decision. Vol.15, No.3. May 2000. pp.263~266.
[3] Caselles V, Kimmel R,G. "Geodesic active contours", International Journal of Computer Vision, Vol.22, No.1, 1997. pp.61~79.
[4] Yang Ping, Zheng JinHua, "Comparison and research over genetic selection operators". Journal of Computer Engineering and Applications, Vol.43, No.15. 2007. pp.59~62.
[5] Ji Genlin, "Survey on Genetic Algorithm". Journal of Computer Application and Software, Vol. 21, No.2. pp.69~73.
[6] Ren Qingshen, Ye ZhongXing, "Analysis on Genetic Operators". Journal of Shanghai Jiaotong University, Vol.34, No4, 2000, pp.564~566.
[7] Hu MiaoJuan, Hu Chun, Qian Feng, "Genetic Analysis of Selection Strategy in Genetic Algorithm". Journal of Computing and Digital Engineering, Vol.34, No.3, 2006, pp.1~3.
[8] Xie XiaoFeng, Zhang WenJun, Yang ZhiJian, " A Parents selection strategy fighting premature convergence in floating genetic algorithm" . Journal of Control and Decision, Vol.17, No.5, Sept, 2002, pp.626~628.

MAINTAIN GENE DIVERSITY WITH SET PAIR ANALYSIS IN PARALLEL GENETIC PROGRAMMING

Xiaotao Wei[1], Xiaohui Yang[2], Yilong Deng[1]
Software school[1], School of Computer and Information Technology[2], Beijing Jiaotong University
Beijing 100044, China
Email: weixt@bjtu.edu.cn

ABSTRACT

This paper proposed a set pair analysis(SPA) based migration model for parallel genetic programming (PGP) algorithm. In order to keeping the gene variety in one subpopulation, only individuals that different enough can migrate to this subpopulation. A novel method for evaluating the similarity of two tiny programs is presented. The algorithm is developed in the MapReduce framework. The experimental results proved that our SPA based PGP can maintain the gene diversity and find the best program. Also the performance is efficient.

Keywords: Parallel Genetic Programming, Set Pair Analysis, MapReduce

1. INTRODUCTION

Genetic programming (GP) is a variation of genetic algorithms in which the evolving individuals are themselves computer programs instead of fixed length strings from a limited alphabet of symbols [1]. It is a machine learning technique used to optimize a population of computer programs according to a fitness function determined by a program's ability to perform a given computational task. The programs are traditionally represented in memory as tree structures. Every tree node has an operator function and every terminal node has an operand, making mathematical expressions easy to evolve and evaluate. The GP approach evolves a population of trees by using the genetic operators of reproduction, recombination and mutation.

The sequential GPs may get trapped in a sub-optimal region of the search space thus becoming unable to find better quality solutions. So the parallel genetic programming (PGP) is taken into account. PGP concurrently evolves a number of separate subpopulations which once in a while exchange individuals, thus overcomes the problem in population diversity.

PGP is computationally very intensive. Executing a complex PGP program requires high performance computer clusters which are not always available to the common researchers. But recently, the idea of cloud computing provides a solution for these computational intensive applications. In this paper, we implement our parallel genetic programming algorithm on the Google MapReduce framework.

In section 2, we introduce some related works. In section 3, we describe our MapReduce architecture based parallel genetic programming algorithm. Followed by the experimental results. Finally, some conclusions are given in Section 5.

2. RELATED WORKS

2.1 Parallel Genetic Algorithm

Comparing to sequential genetic algorithm, parallel implementations allow to tackle even larger and harder problems since they not only provide solutions faster but also very often converge to better solutions. The three best known models for parallel genetic algorithms are the farming model, the migration model, and the diffusion model [6]. The migration models are most widely used. In this model, the whole population is divided into a number of subpopulations which exchange individuals from time to time, this is called migration.

A parallel genetic algorithm following the migration model consists of the following steps:

define number of subpopulations
define connection among subpopulations
randomly generate initial subpopulations
evaluate each individual in all subpopulations
DO (for each subpopulation in parallel)
    select new parents
    create offsprings from parents with the crossover operator
    mutate offsprings
    evaluate offsprings
    migration:(might not be executed in every generation)
        send individuals to neighboring subpopulations and receive individuals from neighboring subpopulations
LOOP UNTIL termination criterion satisfied

In this model, the migration strategy plays an important role in keeping gene diversity within every subpopulation. One subpopulation should only accept different enough individuals from other subpopulations. In genetic programming, to evaluate the difference of two programs is very hard. Here we employ the set pair analysis to solve this problem.

2.2 The Set Pair Analysis Method

Set Pair Analysis (SPA) [7] is a method that integrates both qualitative and quantitative analysis for studying information systems. With the ‘contact number’, SPA can analyze and process various uncertain problems such as those caused by fuzzy, stochastic and incomplete information.

Let X and Y be two given sets, and denote a set pair made up with the two sets H=(X, Y). Under some specific background W, set pair H have N features on X and Y, in which S features are identical ones of X and Y, P features are on the contrary in X and Y, F features are neither identical nor opposite ones of X and Y. We define the ratio under background W as follows:

$\frac{S}{N}$: The identity degree of X and Y;

$\frac{F}{N}$: The difference degree of X and Y;

$\frac{P}{N}$: The opposite degree of X and Y.

Let

$$\mu(X,Y) = \frac{S}{N} + \frac{F}{N} i + \frac{P}{N} j \qquad (1)$$

represent the relationship of X and Y, *i* is the coefficient of the difference degree, $i \in [-1,1]$; *j* is the coefficient of the opposite degree, and is specified as -1. μ is called the connection degree of X and Y under the background W, namely the contact number of H.

In the next section we describe a new migration method based on SPA.

2.3 Cloud Computing and the MapReduce Framework

Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Users need not have knowledge of, expertise in, or control over the technology infrastructure "in the cloud" that supports them. The concept incorporates infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS) as well as other recent technology trends that have the common theme of reliance on the Internet for satisfying the computing needs of the users. [2]

MapReduce is a software framework introduced by Google to support distributed computing on clusters of computers. The framework is inspired by “map” and “reduce” functions. [3] In Google AppEngine, programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

The map function goes over a conceptual list of independent sub-problems and performs a specified operation on each of them independently. The reduce operation takes the answers to all the sub-problems and combines them appropriately to get the output - the answer to the problem it was originally trying to solve.

The advantage of MapReduce is that it allows for distributed processing of the map and reduction operations. Provided each mapping operation is independent of the other, all maps can be performed in parallel.

3. THE SPA BASED PGP

3.1 The Migration Strategy

In our algorithm, after evolving several generations independently within each subpopulations, the best individual is selected and transferred to another subpopulation whose individuals are mostly different from this one. The similarity of two tiny programs is calculated by computing the contact number μ of the two tiny programs A and B.

Let X = all elements(operators and operands) in program A;
Y = all elements in program B;
N = number of elements in X∪Y
S = number of elements both in X and Y
F = number of operand elements only in X or Y
P = number of operator elements only in X or Y.

We set *i* = - 0.2, *j* = - 1, then according to (1):

$\mu(X,Y) = \frac{S}{N} - 0.2 * \frac{F}{N} - \frac{P}{N}$ will represents the similarity of two tiny programs A and B. The higher the contact number μ is, the similar the two programs are.

For example, given that:

$$A = 3x^2 + 2y - 5$$
$$B = 6x^2 + 3x - 5y$$
$$C = x + 2y + 3$$

we can evaluate the similarity of A, B and C by the following process: The elements in A, B and C are:

E(A)={3,*,x,^,2,+,y,-,5}
E(B)={6,*,x,^,2,+,3,-,5,y}
E(C)={x,+,2,*,y,3}

First we calculate the similarity of A and B:

N(A,B)=E(A)∪E(B)= {3,*,x,^,2,+,y,-,5,6}
S(A,B)= {3,*,x,^,2,+,y,-,5}
F(A,B)= {6}
P(A,B)= Ø

So N=10, S=9, F=1, P=0. then the contact number:

μ (A,B)=9/10 - 0.2*1/10 – 0/10 = 0.88

Now the similarity of A and C:

N(A,C)=E(A)∪E(C)= {3,*,x,^,2,+,y,-,5}
S(A,C)= {x,+,2,*,y,3}
F(A,C)= {5}
P(A,C)= {^,-}
N=9, S=6, F=1, P=2
μ (A,C)=6/9 - 0.2*1/9 – 2/9 = 0.42

In like manner, we can get:

μ (B,C)=0.36

From the results we know that A and B are quit similar whereas B and C are quit different. Now we can present our algorithm.

3.2 The SPA Based PGP Algorithm

We employed the MapReduce framework to implement our PGP algorithm. In the map step, we assign each subpopulation to a work node to evolve independently. In the reduce step, we collect all the subpopulations and select the best individual. If the best score is not good enough, we apply migration algorithm and repeat the evolving process. The algorithm is described as follows:

Algorithm1: SPA based PGP

Input:
maxgen: the max generation of evolution;
reservingrate: the proportion of reserving best individuals;
mutationrate: the probability of mutating;
breedingrate: the probability of breeding;
pnew: the probability of introducing new individuals

Output:
the best individual

Initialization:
Randomly create a population with *m***n* individuals
Randomly divide the *m***n* individuals to *n* subpopulations with *m* individuals each.

Loop:

Map:
For each of the *n* subpopulations:

Loop:

```
        Rank all individuals in this subpopulation with
            fitness function,
        Sort the individuals with their scores.
        if best_score=0 break;

        Reserve the best reservingrate individual to the next
            generation

        While population not full:
            with probability pnew do:
                add a new random individual
            or else
                Select two individuals and breed a new
                individual, and mutate it with probability
                mutationrate.
        End While
    Until maxgen reached

  Reduce:
    Select the best individuals.
    If best_score=0 then
        Return this individual and End.
    Migration:
        Migrating individuals according to 3.2

Until the maximum iteration number I is reached.
```

4. EXPERIMENT AND RESULTS

Our PGP algorithm is coded in Python. We test our algorithm on a Hadoop cluster with 4 PCs.

The task is to guess the following function:

$$z(x, y) = x^2 + 2xy + 7 \tag{2}$$

We construct a dataset with capacity of 50 from this function. The fitness function to evaluate the evolved programs is given below.

$$score(f) = \sum \left| f(x, y), z(x, y) \right| \tag{3}$$

The initial parameters are set as follows:

population size: *m***n*=1000,
subpopulation number: *n*=5,
m=200,
maxgen=5,
reservingrate=0.01,
mutaionrate=0.1,
breedingrate=0.1,
pnew=0.05,
the maximum iteration number: *I*=50.

The evolution process ends if there is an individual whose score(f)=0 or the maximum iteration number is reached.

We repeated the experiment for 15 times. The results are shown in Table 1.

Table 1. Experiment result of SPA based PGP

	Evolved generations	Evolving time (seconds)
1	6	58.500
2	5	23.954
3	6	71.719
4	4	8.375
5	2	29.984
6	7	279.094
7	5	4.797
8	6	12.390
9	10	38.327
10	3	18.734
11	3	21.828
12	8	17.671
13	8	71.906
14	13	11.641
15	13	31.359
average	5.25	46.685

We also run the PGP with a random migration strategy for the same task with the same parameters. The results are shown in Table 2.

From the experimental results we can figure out that the random migration model based PGP algorithm may not converge, while our SPA based PGP has no such instance. That means the SPA based PGP maintains the gene diversity well.

Table 2. Experiment result of random PGP

	Evolved generations	Evolving time (seconds)
1	6	13.327
2	5	20.000
3	6	13.140
4	Not converge	
5	4	19.906
6	16	22.703
7	1	13.345
8	2	24.016
9	Not converge	
10	4	7.015
11	4	8.641
12	4	33.593
13	6	34.999
14	3	67.952
15	Not converge	
average	5.08	23.220

These two approaches perform pretty much the same on average evolved generation, but our PGP approach has a higher computational time because of the extra calculation work of comparing the programs in migration stage. But consider the significant advantage in finding the best individual with 100% probability, it worth that.

5. CONCLUSIONS

This paper introduces an implementation of parallel genetic programming algorithm. In order to maintain gene diversity we proposed a novel method called set pair analysis to calculate the similarity of two tiny programs. We also modified the traditional parallel genetic algorithm to smoothly translate it to the MapReduce framework. The experiment result verifies the idea and brings a satisfactory result.

ACKNOWLEDGEMENT

This research was supported by the cooperative project of Beijing Municipal Commission of Education and Beijing Jiaotong University number 353011535.

REFERENCES

[1] J. R. Koza. Genetic Programming: On Programming Computers by Means of Natural Selection and Genetics, MIT Press, 1992.

[2] http://en.wikipedia.org/wiki/Cloud_computing
[3] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters, http://labs.google.com/papers/mapreduce.html
[4] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia, Above the Clouds: A Berkeley View of Cloud Computing http://d1smfj0g31qzek.cloudfront.net/abovetheclouds.pdf
[5] Michael G. Noll, Writing An Hadoop MapReduce Program In Python, http://www.michael-noll.com/wiki/Writing_An_Hadoop_MapReduce_Program_In_Python.
[6] Udo Kohlmorgen, Parallel Genetic Algorithms. Proceedings of the 1NWGA Vaasa. Finland, January,1995.
[7] Dong Liang, Li Guangjun, He Zhongxiong, Pattern Recognition Based on All Set Theory and SPA in Complex System, Journal of Communication and Computer, Vol3(6),2006

USING ALGEBRAIC EXPRESSION FOR CONSTRUCTION OF PROGRAM DEPENDENCE GRAPH *

Duanzhi Chen
School of computer, Wuhan University
Wuhan, Hubei, 430072, China
Email: cdz830526@126.com

ABSTRACT

Program slicing provides a new way to software maintenance, program understanding. It can give the programmer the statements only relevant to the computation of a given function. It can transform a large program into a smaller one. In this paper, we present a new algorithm to produce Program Dependence Graph which is used in Program slicing. The algorithm uses an input called F(p) which models each program in the form of an algebraic expression.

Keywords: Program Slicing, Control Flow Graph, Control Dependence Subgraph, Data Dependence Subgraph, Algebraic Expression

1. INTRODUCTION

Program slicing is first brought forward by Mark Weiser in his Doctoral dissertation in 1979. In the 80's K.J.Ottenstein and L.M. Ottenstein create Graphics-up algorithm based on Program Dependence Graph to compute Intraprocedural Backward Slice[1,2,3].

Program slicing is a technology to compute the program slice. The program slice consists of all statements in a program that may affect the value of variable v at some point you have chosen. The Static Slice Criteria is a binary group<V,p> which the "V" is represented the group of variables and the "p" is represented the interested point in the program. All the statements in the program which can affect the value of the variables in the group "V" in the interested point "p" without considering the input is the Static slice[1]. The Dynamic Slice Criteria is a three group<V,p,x>which the "x" represented the program's input. All the statements in the program which can affect the value of the variables in the group "V" in the interested point "p" with the input represented by "x" is the Dynamic slice[4].

Program slicing is used in debugging; understanding programs[5].It can give you a smaller size of program you are dealing with. The program slice only contains the statements which affect the value of variables in the Program Slice Criteria so you can make use of it in debugging.

Mark Weiser uses the Control Flow Graph to compute the program slice[1]. Than K.J.Ottenstein and L.M. Ottenstein makes Program Dependence Graph to be the internal program representation[2]. The Program Dependence Graph(PDG) consists of Control Flow Graph(CFG), Data Dependence Subgraph(DDS) and Control Dependence Subgraph(CDS)[3].

The DDS describes the data dependence between the different statements. If there is a path P from the A statement to B statement in the program's Control Flaw Graph (1) ; the value of the variable v is defined or redefined in A (2) ; the statements in the path P which contacts A and B don't redefine the value of the variable v (3) ; the B refers the v (4), we call the B has the data dependence to A by the variable v. We can compute the reachable definitions of the statements in the program to produce the DDS[5].

The CDS describes the control dependence between the different statements. B has the control dependence to A if the following three conditions is satisfied;(1)there is a path exits in the program's Control Flaw Graph which contacts A and B;(2)all the statements in the path p except A and B is post-dominanced by B;(3)A is not post-dominanced by B.
This paper presents a novel algorithm to produce the PDG. It makes the program's internal representation called F(p) to the algorithm's input. Then we will give the brief introduction of F(p) in the next section.

2. INTERNAL REPRESENTATION

F(p) is the internal representation of the program which models the program by algebraic expression[6]. It is language independent and it is suitable for the procedural languages. A procedural-language program can be viewed as a set of P of test predicates and a set of B of instruction (procedure) blocks embedded in a set S of one-in/one-out control environments. The statements which have one-in/one-out control flaw in the program's control flaw graph belong to set B. For example, assignment statements, procedure call statements, "return" statements and so on. The statements like "if " or "while" which get together with the test predicates belong to set P. For example, there is an F(p) algebraic expression shows in the Figure 2 represents the program shows in the Figure1.

The items of the algebraic expression have two attributes---one is the keywords and the other one is the line number. The range of the keywords is the set K= {b, if, w, c, br}. Among the set K "c" represents "continue" and "br" represents "break", then "w" represents "while" and "if" represents "if", final the symbol "b" represents all of the statements who belong to simple statements. The number in the back with the keywords represents the line number of the statement in the program.

The Structured Programming Languages only have sequence structure, branch structure, loop structure. As noted above that the F(p) models the program by the algebraic expression we give the expressions respectively to model these three structures as follows:

- sequence: seq= X_i / X_j/ . .. / X_n
- if-else: $if_i \bullet (X_i + X_j)$
- while: $w_i = w_i \bullet (X_i + \lambda)$

* Supported by the National High Technology Research and Development Program of China (863) (2002AA111010)

```
b1:      int i=0;
b2:      int j=10;
w3:      while(j>0)
         {
if5:         if(i<j)
             {
b7:              i=i+j/10;
             }
b9:          j--;
if10:        if((i%2)==1)
             {
b12:             i++;
             }
+:           else
             {
b16:             i=i+2;
             }
b18:         printf(i);
}
```

Figure 1. The Program Source Code

b1/b2/w3●(if5●(b7+λ)/b9/if10●(b12+b16)/b18+λ)

Figure 2. The F(p)

The subscript "i" follows the "X" or the "if" and the "w" describes the line number of the statements in the program source code. The symbol "/" denotes the serialization, it describes the symbol "X" which is in front of it has the control flow to the one which next to "/"in the control flow graph and they are the implementation of the order. The symbol "+" means choice. For example, "if_i●(X_i+ X_j)", "X_i" and "X_j" are the alternative operators , when the "X_i" is executed then the "X_j" isn't executed at the same time and vice versa. The condition which decides who will be executed is the predicates embedded in "if "structure. The symbol "●" denotes the inclusion . After it there must be a left bracket and a right bracket which corresponds to the left one. The part of the algebraic expression which is enclosed in the above-mentioned brackets is inclusive operations of the "if" or "while" structure which is in front of the symbol "●". The symbol "λ" means no operation.

Finally the procedural program can be represented by these functions. Then we will give three algorithms respectively to construct CFG, DDS and CDS.

3. PRODUCING CONTROL FLOW GRAPH

Control flow graph describes the transmission of the control flow between the statements which is the basis of the analysis on data dependence and control dependence. Structured programming Languages in general consist of sequence structure, branch structure and loop structure. Among them, sequence structure transfers the control flow directly to the next statement; branch structure needs to divide the control flow into two parts and transfer them respectively to the corresponding branches, at the exit of the branch structure it needs to re-gather these flow and transfer them to the next statement; loop structure needs to transfer the control flow back to the beginning of the loop structure when control flow arrive at the end of loop structure.

The algorithm reads each item of the algebraic expression and takes the different operations base on the active item as well as the reading item. Its pseudo code is shown in the Figure 3.

```
CFGdrawing{
  active node= "Start node";
  while( F(p) still has item){
    read next item;
    switch the item's keywords{
    case "b"|| "c"|| "br"
    relation (active, reading); active node= reading item;
    case "w"|| "if"
    relation (active, reading); push into stack "predicate";
      active node= the reading item;
    case "+"
       if(top of "predicate" is "if"){
         create virtual exit node; relation (active, virtual);
         push virtual to "exit"; active node="if";}
    case ")"
      if(top "predicate" is "if"){ pop the stack "exit";
         relation (active, exit); active node= exit;}
      else if(top "predicate" is "w"){
         relation (active, exit); active node= "w";}}
```

Figure 3. Algorithm to Construct CFG

We develop a tool to draw the CFG base on the algorithm above-mentioned. We use it to draw the CFG of the program shown in figure 1 and the result is shown in Figure 4.

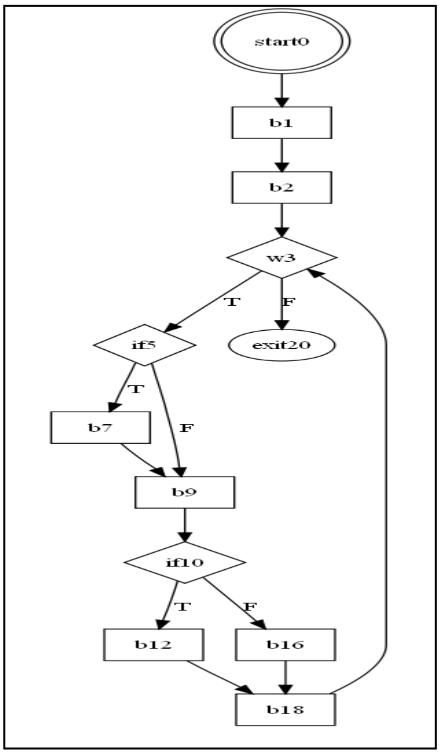


Figure 4. The CFG of Program

4. PRODUCING DATA DEPENDENCE GRAPH

After construction of the CFG we can construct the DDS base on it. We assume that the KILL set and the GEN set of every statements are created. KILL set makes up of the statements which re-assignment variables. GEN set makes up of the statements which assignment variables firstly. In order to compute data dependence the algorithm needs to create the IN set and the OUT set of each statements. The IN set of a node consists of those definitions that reach the point immediately before the statement; the OUT set consists of those definitions that reach the point immediately after the statement.

Definition 1(Reach-definition): If there exist a path in the CFG from node a to the node b, a define a variable v and in this path no other one re-define v then we call <a, v> is a Reach-definition of b.

Definition 2(USE set): USE set consists of the variable that used in the statement.

All the Reach-definition of a form IN(a). IN(a) ∩USE(a) is the data dependence of node a.

The input of the algorithm is the CFG, KILL set, GEN set and USE set. The output of the algorithm is IN set and OUT set. Algorithm is shown in Figure 5.

```
Definition-reachablity(node X)
  IN(X)= ∪(OUT(Y),Y in pred(X) in CFG);
  OUT(X)=IN(X)∪GEN(X)-KILL(X);
  if(IN(X) or OUT(X) changes from before){
     for all Y in succeed(X) in CFG{
        call Definition-reachablity(node Y);}}
            End Definition-reachablity
```

Figure 5. Algorithm to Compute IN Set and OUT Set

The DDS of the program is shown in Figure 6.

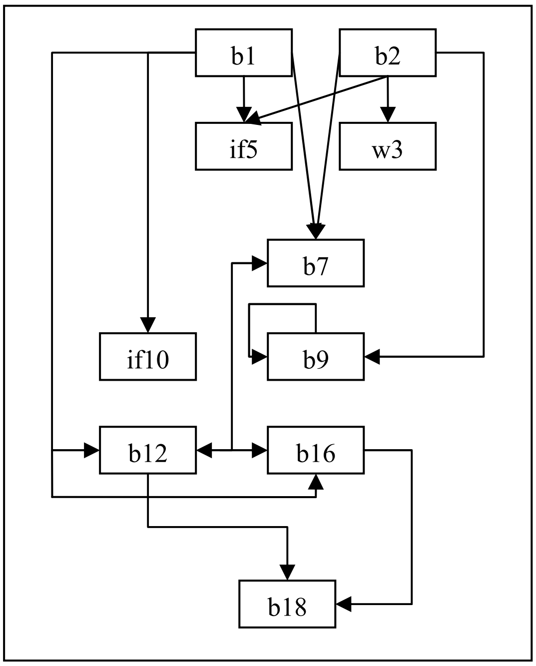


Figure 6. The DDS of the Program

5. PRODUCING CONTROL DEPENDENCE GRAPH

The main idea of the algorithm is to read the items in the F(p) from the first one to the last one according to their original order. It reads one item at a time. Algorithm decides what will do bases on the item it is reading. The particular description of the algorithm showed in the Figure 7.

In addition to the block nodes and the predicate nodes there are the region nodes created by the algorithm. The region nodes aren't relevant to the source code like the other nodes. It is used to summaries the set of control dependences

The variable named "now" storages the current active node. It is used to be the pre-node of the element reading by the algorithm. The stack named "active" storages predicate node and region node. We use stack to maintained the nesting structure's control dependences.

```
DDSdrawing{
  now node= "Start node";
  while( F(p) still has item){
    read next item;
    switch the item's keywords{
    case "b"|| "c"|| "br"
    relation (now, reading);
    case "w"|| "if"
      relation (now, reading);create region node;
      push reading node and region node into "active";
    now= region node;
    case "+"
     pop "active"; read next item;
     if(reading node is not "λ"){
      create region node; relation (top of "active", region);
      push region node into "active"; now= region node;}
    case ")"
      if(top of "active" is region)pop "active";
        pop "active"; now=top of "active";}
```

Figure 7. Algorithm to Produce the CDS

When the last element in the F(p) is dealt with. The CDS of the program constructed successfully. The Control Dependence Subgraph is shown in the Figure 8.

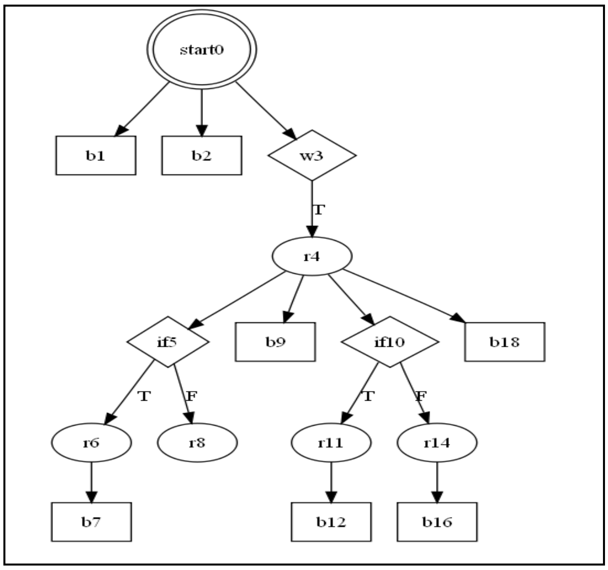


Figure 8. The CDS of the Program

6. CONCLUSION

The definition of an internal representation of the programs from which different external representations (or views) is a difficult thing. The F(p) algebraic expression is a good internal representation of the programs. The paper presents a novel algorithm to constructs the PDG that is used in program slicing. In addition to the Program Dependence Graph, several other denpendence graphs are brought up. S.Horwitz, T.Reps, and D.Binkley use the System Dependence Graph to compute the interprocedural program slice[7]. B. A. Malloy invents the Object-oriented Program Dependency Graphs to slice the Object-oriented Program[8].

The algorithm still can not deal with the interprocedural programs. The structure transfer statements such as “continue” or “break” can not be dealt with too. So the next step is to find out the solution to the structure transfer statements and finally solve the programs follows the Structured Programming.

REFERENCES

[1] Mark Weiser, “Program slicing”, IEEE Transaction on Software Engineering, SE-10, No.4, July 1984, pp. 352-357.

[2] Mark Weiser, “Programmers use slices when debugging ” , Communications of the ACM, SE-25, No.7, 1982, pp. 446-452.

[3] K.J.Ottenstein, L.M. Ottenstein “The program dependence graph in a software development environment”, ACM SIGSOFT/SIGPLAN Software Engineering Symposium on Practical Software Development Environments, April 1984, pp.177-184.

[4] Korel.B,Rilling.J,“Program slicing in understanding of large programs” , 6th International Workshop on , Ischia,06/24/1998, pp. 145-152.

[5] Mary Jean Harrold,Brian Malloy, Gregg Rothermel, “Efficient construction of program dependence graphs”, the 1993 ACM SIGSOFT international symposium Software testing and analysis , ACM Press, June 1993, pp. 160-170.

[6] Cimitile, A, and De Carlini, “Reverse Engineering Algorithms for Program Graph Production”, IEEE Transactions on Software Engineering, Zurich, Volume: 24, Issue: 9, Sept 1997, pp. 721-740.

[7] S.Horwitz, T.Reps, and D.Binkley, “Interprocedural slicing using dependence graphs”,ACM Transaction on Programming Languages and System, Volume:12, Issue: 1, Jan 1990, pp.26-60.

[8] B. A. Malloy, J. D. McGregor, A. Krishnaswamy, and M. Medikonda. “An extensible program representation for object-oriented software”, ACM Sigplan Notices, Volume:29, No.12, December 1994, pp. 38–47.

PARALLEL FINITE ELEMENT METHOD FOR 2-DIMENSIONAL WAVE EQUATION *

Qinwu Xu, Zhoushun Zheng, Shuang Wang
School of Mathematical Sciences and Computing Technology, Central South University
Changsha, Hunan 410083, China
Email: qinwu_xu1986@126.com

ABSTRACT

By using finite element discretization in space domain and 2-order central difference discretization in time domain separately, a full discretization scheme is given for 2-dimensional wave equation. Then parallelization of the formation and solution process of finite element equations is described in detail. Numerical experiments were carried out base on parallel computing environment using C language and MPI. Numerical results show that, the total time increases while the data scale increases; when data scale is fixed, speedup ascends approximately linearly and the total time descends sharply when number of processors is not too large. As the number of processors increases, parallel efficiency descends gradually. Take time used by serial part of the program into account, total time is highly identical with the theoretic value.

Keywords: Wave Equation, Finite Element Method, Parallel Computing, Speedup.

1. INTRODUCTION

The computational complexities of transient dynamic nonlinear finite element analysis of structures may become very large for many practical engineering problems. There being not enough storage space and long computing time are two difficulties in engineering computing[1]. Parallel and distributed processing, which permits the engineer to undertake the finite element analysis in a considerably shorter time, is therefore of increasing significance[2]. Recent years many researchers have done a lot of work about parallel finite element method at home and abroad [1,3-7], and now parallel finite element method has been used in a various of application fields[8-10]. In this paper a full discretization scheme is given for 2-dimensional wave equation and process of parallelization will be described in details. At last numerical experiments will be carried out based on parallel computing environment using C language and MPI.

2. FULL DISCRETIZATION SCHEME FOR 2-DIMENSIONAL WAVE EQUATION

We consider the following initial boundary value problem for 2-dimensional wave equation,

$$\begin{cases} \dfrac{\partial^2 u}{\partial^2 t} = \Delta u \\ u(x,y,0) = \varphi(x,y) \quad ,(x,y)\in\Omega \\ u_t(x,y,0) = \psi(x,y) \\ u|_\Gamma = \eta(x,y,t) \end{cases} \tag{1}$$

For simplicity, we choose $\Omega = (0,1)\times(0,1)$

$$\varphi(x,y) = \max\left\{0, 2-5\sqrt{(x-1/2)^2+(y-1/2)^2}\right\} \tag{2}$$

$$\psi(x,y) = 0 \tag{3}$$

$$\eta(x,y,t) = 0 \tag{4}$$

For $\forall v \in C^2(\Omega)\cap C(\overline{\Omega})$, multiply the two side of the first equation of Equ.(1) by v and integrate over Ω, we obtain

$$\iint_\Omega \left(\frac{\partial^2 u}{\partial^2 t} - \Delta u\right) dxdy = \frac{\partial^2}{\partial^2 t}\iint_\Omega uvdxdy - \iint_\Omega \Delta uvdxdy = 0$$

Integrating by parts and using Green's formula, we get

$$\frac{\partial^2}{\partial^2 t}\iint_\Omega uvdxdy - \iint_\Omega \Delta uvdxdy$$
$$= \frac{\partial^2}{\partial^2 t}\iint_\Omega uvdxdy + \iint_\Omega \nabla u\cdot\nabla vdxdy - \frac{\partial u}{\partial \mathrm{n}}v\Big|_\Gamma$$

Using the boundary condition (4), we get

$$\frac{\partial^2}{\partial^2 t}\iint_\Omega uvdxdy + \iint_\Omega \nabla u\cdot\nabla vdxdy = 0 \tag{5}$$

Assume that the triangulation T_h is given, we define a finite element space

$V_h = \left\{v : v\in P^1\right\}$

where P^1 denotes the piecewise 1st-order polynomials on T_h.

Let $\phi_1,\phi_2,\cdots,\phi_N$ be a group of basis functions of V_h, where $\phi_i (i=1,2,\cdots,N)$ is piecewise 1st-degree polynomial, which equals to 1 at node *i* and to 0 at elements that do not contain the node *i*. N is the number of nodes in triangulation T_h.

For a fixed time *t*, *u* can be expressed as $u=\sum_{i=1}^N u_i\phi_i$. For different time *t*, we set

$$u(t) = \sum_{i=1}^N U_i(t)\phi_i \tag{6}$$

Substituting Eq.6 into Eq.5, we obtain the following ordinary differential equations

$$\frac{\partial^2}{\partial t^2}\sum_{i=1}^N U_i(t)\iint_\Omega \phi_i\phi_j dxdy + \sum_{i=1}^N U_i(t)\iint_\Omega \nabla\phi_i\cdot\nabla\phi_j dxdy = 0, \quad j=1,2,\cdots N \tag{7}$$

Let *M, K* be the mass matrix and diffusion matrix, then Eq.7 can be expressed as

$$\frac{\partial^2}{\partial t^2}MU(t) + KU(t) = 0 \tag{8}$$

We get the semi-discretization form of equation (1). For time discretization, we use the 2nd-order central difference scheme, as follows

$$M\frac{U^{n+2}-2U^{n+1}+U^n}{\Delta t^2} + KU^{n+1} = 0 \tag{9}$$

Where $U^n = \left(U_1^n, U_2^n, \cdots, U_N^n\right)^T$.

To solve the difference equations, we need to known two time levels to start the iteration process. First, we compute the $U(t)$ at time level 2 using the initial value condition (2) and (3). In order to preserve 2nd-order accuracy, we also use the central

* Supported by the National Natural Science Foundation of China under Grant No.60773218; the National High-Tech Research and Development Plan of China under Grant No 2009AA01Z122.

difference scheme $M\dfrac{U_t^{1+1/2}-U_t^1}{\Delta t/2}+\dfrac{K}{2}\left(U^2+U^1\right)=0$, where $U_t^{1+1/2}=\left(U^2-U^1\right)/\Delta t$.

Then we get the following equation

$$M\frac{\left(U^2-U^1\right)/\Delta t-U_t^1(t)}{\Delta t/2}+\frac{K}{2}\left(U^2+U^1\right)=0$$

Above all, we get

$$\begin{cases}\left(M+\Delta t^2K/4\right)U^2=\left(M-\Delta t^2K/4\right)U^1+\Delta tMU_t^1\\ MU^{n+2}=\left(2M-\Delta t^2K\right)U^{n+1}-MU^n, n\geq 1\end{cases}\tag{10}$$

3. PARALLELIZATION AND SOLUTION PROCESS OF FINITE ELEMENT EQUATIONS

The parallelization process is partitioned into 4 steps:

(1) First processors are divided into several groups and forms the process topology for communication, for example the process topology with 8 processors is shown in Fig.1.

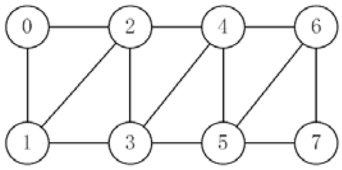


Figure 1. Processors topology

The line connecting two processors implies that there exists information exchange between the two processors.

(2) Then the elements of the triangulation are divided into several groups, each group of elements is mapped onto a processor, as shown in Fig.2. Every two groups adjacent with each other have a common part that contains a row or a column of elements, in order that the stiff matrix can be computed separately on each processor without communication with each other, as shown in Fig.3.

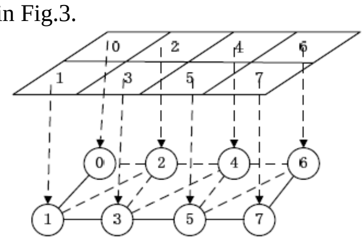


Figure 2. The mapping between mesh and processors

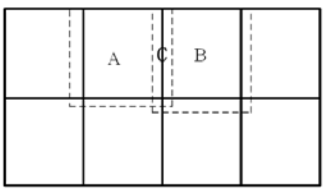


Figure 3. Elements grouping

(3) Since each element has communication only with the closest neighbors, each element stiff matrix is computed separately, then these element stiff matrices are accumulated to forms total stiff matrix for each group of elements. During the assembling process processors have no communication with each other, so the parallel degree can reach a high degree. At the same time, the right-hand side is also computed separately.

(4) The finite element equations are solved for each time step iteratively by parallel conjugated gradient method.

4. NUMERICAL EXPERIMENTS AND RESULTS

First five triangulations with different number of nodes are given and the problem is solved with the same number of processors. The time used by solving the problem at five time levels with four processors is shown in Tab.1. From Tab.1 and Fig.4, we know that the total time increases as the number of nodes increases but the change law is not obvious. By analyze the iteration times and time of each iteration step we know that different equations needs different iteration times to reach accuracy of the same order, so the total time used is not certain.

Table 1. The time used by solving the problem with four processors

No.of nodes	100×100	200×200	300×300	400×400
No.of elements	19602	79202	178802	318402
Total time	1.46875	11.04296	22.847656	23.660156
Iteration times	1096	3327	2447	1398
Time/ Iteration	0.0013116	0.003286	0.009233	0.01660

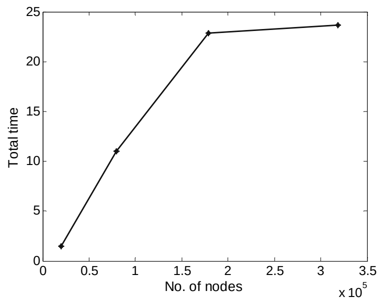


Figure 4. The relation between total time and the number of nodes

Then problems with the same number of nodes are solved using different number of processors, the result for 400×400 nodes at five time levels is shown in Tab.2 and Fig.5.

Table 2. Relation between time, speedup and No. of processes

No. of processes	1	2	3	4	6	12
Time (s)	22.89	11.5	7.50	5.82	4.2	2.78
speedup	1.0	1.99	3.05	3.93	5.45	8.23

We can see from Tab.2, Fig.5 and Fig.6 that at the beginning speedup ascends approximately linearly and the total time descends sharply and the average parallel efficiency is up to 99.8% when No. of processors is less than 4. As the number of processors increases, parallel efficiency descends gradually. When the number of processors reach a certain degree the speedup tend to stability and even begin to descend.

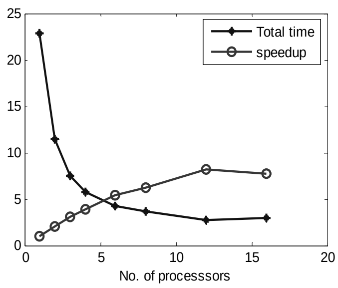


Figure 5. Relation between time, speedup and number of processors

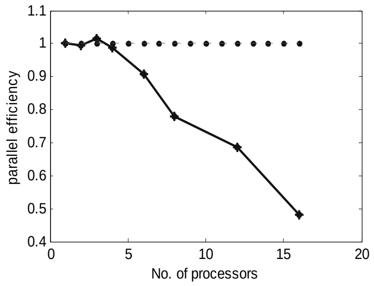


Figure 6. Relation between parallel efficiency and number of processors

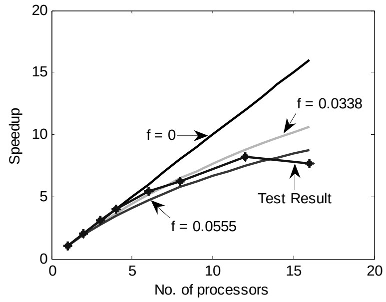


Figure 7. Speedup under different series ratio

We know under ideal condition speedup will change linearly with respect to the number of processors and its value will equals to the number of processors. Since part of the program can not be carried out in parallel and there exists communication between processors, speedup doesn't subject to linear relation with respect to the number of processors.

Let f be the ratio of serial part of the program and p be the number of processors. Let t_s be the time used when problems are solved serially and $S(p)$ be speedup when p processors are used. Analyzing the numerical results, the following formula is derived

$$S(p) = \frac{p}{1+(p-1)f} = \frac{p}{1+0.0555(p-1)} \qquad (10)$$

Consider time used by the serial part of the program, we can obtain the following equation

$$T = t_{serial} + t_{parallel} = f \cdot t_s + \frac{(1-f)t_s}{p} \qquad (11)$$

Where t_{serial} denotes time used by serial part of the program, $t_{parallel}$ denotes time used by parallel part of the program, t_s denotes time used when problem is solved serially.

Analyzing experimental result, we get $f = 0.0321$, $t_s = 22.89$, and

$$T = 0.73477 + \frac{22.15523}{p} \qquad (12)$$

From Fig.8 we can see total time is highly identical with the theoretic value.

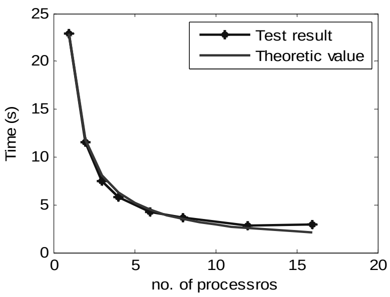


Figure 8. Relation between time and number of processors

5. CONCLUTIONS

In this paper a full discretization scheme for 2-dimensional wave equation is given. Then parallelization of the formation and solution process of finite element equations is described in detail. Numerical experiments were carried out base on parallel computing environment using C language and MPI. Numerical results show that, the total time increases while the data scale increases; when data scale is fixed, speedup ascends approximately linearly and the total time descends sharply and the average parallel efficiency is up to 99.8% when No. of processors is less than 4. As the number of processors increases, parallel efficiency descends gradually. When the number of processors reaches a certain degree the speedup tend to stability and even begin to descend. Take time used by serial part of the program into account, total time is highly identical with the theoretic value.

REFERENCES

[1] Ruan Hong-he, Yuan Yong, Liu Xian. Progress of parallel finite element method on distributed-memory parallel computer systems. Journal of Tongji University (Natural Science), Vol 33, No 1, Jan. 2005, pp.21~27.

[2] A. Rama Mohan Rao. MPI-based parallel finite element approaches for implicit nonlinear dynamic analysis employing sparse PCG solvers. Advances in Engineering Software. Vol 36,No.3, March 2005, pp.181~198.

[3] Liu Yao-Ru,Zhou Wei-Yuan, Yang Qiang. Parallel finite element analysis based on element-by-element method and its application. Chinese Journal of Rock Mechanics and Engineering, Vol 24, No. 17, September 1, 2005, pp.3023~3028.

[4] Paszynski, MacIej. Performance of multi level parallel direct solver for hp finite element method. Parallel Processing and Applied Mathematics - 7th International Conference, PPAM 2007, Lecture Notes in Computer Science, Vol 4967 LNCS, 2008, pp.1303~1312.

[5] Chaojiang Fu. Parallel computing for finite element structural analysis on workstation cluster. Proceedings - International Conference on Computer Science and Software Engineering, CSSE 2008, v 3, 2008, pp.291~294.

[6] Nakajima, K. Three-level hybrid vs. flat MPI on the Earth Simulator: parallel iterative solvers for finite-element method. Applied Numerical Mathematics, Vol 54, No. 2, July 2005, pp.237~55.

[7] Liu Yao-Ru, Zhou Wei-Yuan, Yang Qiang. Parallel 3-D finite element analysis based on EBE method. Engineering Mechanics, Vol 23, No. 3, March 2006, pp.27~31.

[8] Basermann Achim, Tsukerman Igor. Parallel generalized finite element method for magnetic multiparticle problems. Lecture Notes in Computer Science, Vol 3402, 2005, pp.325~339. High Performance Computing for Computational Science - VECPAR 2004 - 6th International Conference.

[9] Cao Yinfeng, Li Guangyao, Zhong Zhihua. Parallel explicit finite element simulation of vehicle collisions. Chinese Journal of Mechanical Engineering, Vol 41, No.2, Feb.2005, pp.153~157.

[10] Li Hong-Lin, Zhang Hai-Lei, Wang Xi-Cheng. Parallel algorithm using finite element method in quantum chemistry. Journal of Dalian University of Technology, Vol 45, No. 4, July 2005, pp.469~472.

Zheng Zhoushun is full Professor and Vice-Chairman of School of Mathematical Science and Computing Technology, Central South University. He graduated from Yunnan Normal University as a Bachelor of Science, July 1985; from Yunnan Normal University as Master of Algebra, July 1987; from State Key Lab of Powder Metallurgy, Central South University, Numerical Simulation of Powder injection moulding, PhD, February 2006.

THE DESIGN AND THEORETICAL ANALYSIS ON PARALLEL NEWTON INTERPOLATION ALGORITHM *

Aimin Yang[1], Juan Li[2], Yuhuan Cui[1] , Jingguo Qu[1]
[1]College of Science, Hebei Polytechnic University Tangshan
Hebei Province, 063009 China
[2] Academic Affairs Department, Hebei Polytechnic University
Tangshan, Hebei Province, 063009 China
Email: aimin_heut@163.com

ABSTRACT

By parallel analysis of Newton interpolation, combined with Strategy of Divide and Conquer, we get an improved interpolation method, parallel Newton interpolation, and given the speed-up ratio and efficiency in the two cases that the number of processor is enough and the number of processor is four. Theoretical analysis results show that the superiority and effectiveness of the method, so it provides a new method for functions which is not easy to processing or computing, and broadens the application areas of parallel algorithms.

Keywords: Divide and Conquer, Parallel algorithms, Newton interpolation algorithm, Speed-up.

1. INTRODUCTION

In many practical application of numerical method, we generally should give a quick estimated to a function by using the method of interpolation. The Language Interpolation Polynomials can deal with this problem which is shown in reference [1-4]. In recent years, many different parallel algorithms can solve the problems of polynomial interpolation. The cluster system has become the most popular high-performance computer platform, and the proportion in high-performance computers is becoming larger and larger. The scale of the system can range from single pc, few networked computers to the large scale parallel system with thousands of nodes. Because this system is cheap and can be easily popularized, it is widely used in our daily life.

This paper presents a parallel algorithm which works out n nodes Newton interpolation polynomial, and provides speedup and efficiency in two states.

2. NEWTON INTERPOLATION

2.1 The content of operation facilities layout

A number of functions $f(x)$ are often more complex or difficult to find its analytical expression in Practical problems. Sometimes through experiments or numerical calculation, We can only get some discrete points $x_i (i = 0,1,...,n)$ on the function value, that is $y_i = f(x_i)$. In practice, the data generally be arranged in tabular form. According to the basis of existing data table of function $f(x)$ to calculate the function value of new points, that is, the problem to be solved by interpolation. Therefore, the so-called interpolation is given by means of the function re-inserted into forms by the required function value of the new point value. The basic idea of interpolation is, First of all, according with the existing function value we trying to construct a simple function $y(x)$ as the approximate expression of the function $f(x)$, and then used $y(x)$ to calculate the new point value as the approximation of function $f(x)$, Polynomial function can usually be selected as the approximate function $y(x)$, because the polynomial has Arbitrary Rank Derivatives, and more convenient to calculated value.

Suppose that function in $y = f(x)$ is continuous in the interval $[a,b]$, and we know $n+1$ different function value of function $f(x)$ in the interval $[a,b]$, that is $y_i = f(x_i)$, among them $x_i \neq x_j (i \neq j)$. Find a polynomial which the degree is not more than n, and let it to meet the conditions $P_n(x) = f(x_i)$, the $P_n(x)$ referred to $f(x)$ as the interpolation polynomial.

In order to structure the interpolation polynomial which thorough $n+1$ different points and the degree is not more than n , $P_n(x)$ can be expressed as the following forms:

$$P_n(x) = N_n(x) = a_0 + a_1(x - x_0) + \cdots$$
$$+a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

It meet the following conditions:

$$N_n(x) = N_{n-1}(x) + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

in accordance with interpolation conditions,

$$N_n(x_0) = a_0 = f(x_0)$$
$$N_n(x_1) = a_0 + a_1(x_1 - x_0) = f(x_1)$$
$$N_n(x_n) = a_0 + a_1(x_n - x_0) + \cdots$$
$$+ a_n(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1}) = f(x_n)$$

This is the lower triangular equations on the unknown number $a_0, a_1, \cdots, a_n$, we can obtain that

$a_k = f[x_0, x_1, ..., x_k] (k = 1, 2, ..., n)$, so

$$N_n(x) = f(x_0) + f[x_0, x_1](x - x_0) +$$
$$f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots$$
$$+ f[x_0, x_1, ..., x_n](x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

The above equation is called Newton divided difference interpolation-type n-order interpolation polynomial, or n-order Newton Interpolation formula.

It should be noted that Newton interpolation formula is generally not more than 8 orders in practical applications. This is because the stability of interpolation of the higher order interpolation polynomial is poor; there will be instability, that is, Runge phenomenon. Therefore, in practical applications we generally use the sub-low interpolation.

3. THE DESIGN OF PARALLEL NEWTON INTERPOLATION ALGORITHM

Through the analysis of Newton interpolation, we can find that

* Natural Scientific Fund Project of Hebei Province (A2009000735), Scientific Fund Project of Hebei Polytechnic University (z200917).

the parallelism may be excavated in the process of calculation divided difference. In view of this the Newton interpolation method has been improved through the process of calculation carried out in parallel to improve the algorithm efficiency. In order to facilitate the introduction of methods, here we combine traditional Newton interpolation method and parallel Newton interpolation method to compare, explain, derivation.

For example, we have the following functions form

i	x_i	y_i	i	x_i	y_i
1	28	20	8	38	43
2	18. 68	65	9	39. 62	28. 44
3	188	100	10	78	35
4	17. 76	25	11	156. 6	36
5	17. 34	90	12	280	35
6	31	94	13	520	0
7	35	80			

Structure polynomial interpolation by using Newton interpolation from the 13 nodes in the above table.

Solution: considering the precision, we choose the fourth-order interpolation to construct function, so we divide the 13 groups of data into three groups:

$$\{28,\ 18.68,\ 188,\ 17.76,\ 17.34\}$$
$$\{17.34,\ 31,\ 35,\ 38,\ 39.62\}$$
$$\{39.62,\ 78,\ 156.6,\ 280,\ 520\}$$

First we calculate the divided difference table of the first group of data, as follows:

x_i	f(x_i)	First-order	Second-order	Third-order	Fourth-order
28	28				
18.68	65	-4.828326			
188	100	0.206709	0.031469		
17.76	25	0.44055	- 0.25418	0.027895	
17.34	90	-154.7619	0.909425	-0.868362	0.084077

Bring all the divided difference into the Newton interpolation formula in above table, we can obtain polynomial interpolation, and the Newton interpolation formula as follows:

$$N_n(x) = f(x_0) + f[x_0, x_1](x - x_0)$$
$$+ f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots$$
$$+ f[x_0, x_1, ..., x_n](x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

After calculating the other two groups, we will also get the polynomial formulas. Through the analysis of the above serial algorithm, it can be found that the parallelization is existence in the calculation process: (1) The calculation process of three groups of data is can be parallel. (2) There is a parallel in the calculation process of divided difference, which during the excavation of the parallel algorithm can be a high degree of improvement.

The following, we use the he idea of divide and conquer to process the parallel in (1).

The idea of divide and conquer is that we can divide a problem into several sub-problems, And then solve the subproblems at the same time, finally combined results of these sub-problems so we get the whole solution of the problem. That is, divide and conquer contains the following steps:

Step one: the problem is divided into sub-problems $P_1, P_2, ..., P_s$, solving at the same time and get the solution.

Step two: combined results of these sub-problems to be the final outcome.

In fact, the grouping of the serial algorithm has revealed parallelism very clearly. Serial algorithm must be calculated in turn in the processing of the above-mentioned three groups of data. In the parallel algorithm, using multi-processor systems, we can parallel processing, that is, three groups of data can be calculated by using parallel processing system of three-processor. Selecting different multi-processor system for the different problem, we can carry out parallel processing of different issues.

However, through careful analysis, we found that in the process of calculating the divided difference, the parallel processing can be carried out by using the binary tree model and it can greatly improves efficiency by the parallel processing algorithm. Therefore, we use the binary tree model for their improvement in the parallel of (2), the following are bad for the calculation of the main tree model using Newton interpolation method to study and improve. In the following we main use the binary tree model to research and improve the Newton interpolation for the calculation of divided difference

Next, we use the five numbers in the first group of data to parallel compute for the problem and extend it to the general situation. As shown below in figure 1, we can establish the five-level binary tree for the five numbers:

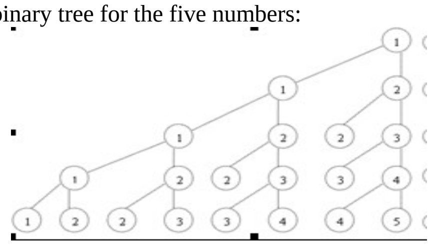


Figure 1. Five-level binary tree.

For clear statements, the real data are replaced by the number in the tree, the bottom of the tree is the given data, the other floor of the tree is the corresponding each order divided difference, the same number in each floor shows that it needs to use twice, and except for the bottom, the other on each floor the node number 1 are the corresponding diagonal divided difference on the diagonal divided table.

The assumption that there are now four available processor P_1, P_2, P_3, P_4. So the parallel computing process is as follows:

First-order divided difference:

Calculated by the processor P_1: $(F_{11})\, f[x_0, x_1]$

Calculated by the processor P_2: $(F_{21})\, f[x_1, x_2]$

Calculated by the processor P_3: $(F_{31})\, f[x_2, x_3]$

Calculated by the processor P_4: $(F_{41})\, f[x_3, x_4]$

Second-order divided difference:

Calculated by the processor P_1: $(F_{22})\, f[x_0, x_1, x_2]$

Calculated by the processor P_2: $(F_{32})\, f[x_1, x_2, x_3]$

Calculated by the processor P_3: $(F_{42})\, f[x_2, x_3, x_4]$

Third-order divided difference:

Calculated by the processor P_1: $(F_{33})\, f[x_0, x_1, x_2, x_3]$

Calculated by the processor P_2: $(F_{43})\, f[x_1, x_2, x_3, x_4]$

Fourth-order divided difference:

Calculated by the processor P_1: $(F_{44})\, f[x_0, x_1, x_2, x_3, x_4]$

It is clear to see that after calculating each-order divided difference, the number of processors reduce one, so when the divided difference of three groups of data were calculated, in accordance with the parallelism between groups, this time using three processors to parallel compute, we can obtain the Newton interpolation polynomial of each group. Finally, we combine the results of the sub-problems and get the ultimate results.

4. ANALYSIS SPEEDUP AND EFFICIENCY OF ALGORITHM

Now we analyze the speed-up ratio and efficiency of the above algorithm, in order to more clearly reflect the parallel and the advantages of the algorithm in this paper, it is not to be considered the unit time of computing a single arithmetic, and we only consider the computing of divided difference as a unit time T , the problem can be shown as good. Therefore, based on the above analysis, without considering communication costs and software costs, it can be calculated as follows.

For the serial algorithm, the calculating time of a groups of data are as follows: $1T+2T+3T+4T=10T$, The total computing time is $T_1=3\times10T=30T$.

For the parallel algorithm, each group uses 4 processors, and then the whole process will use 12 processors. the same-order divided difference in each group were calculated by parallel processing, and it is also parallel processing between group and group, so the computing time is $T_P=4T$.

In accordance with the Calculation Formula of the speed-up ratio and efficiency: $S_p=\frac{T_l}{T_P}=\frac{30T}{4T}=7.5$， $E_p=\frac{S_p}{P}=\frac{7.5}{12}=0.625$

So when there are12 processors available, the theoretical speed-up ratio of the algorithm is 7.5 , the efficiency is 0.625 .

When the dealing with data is N , we assume that the polynomial is still 4-order, so the number of processors required is $N-1$ and the data is divided into $(N-1)/4$ groups. For serial algorithm, the time required is $T_1=\frac{N-1}{4}\times10T=\frac{5}{2}(N-1)T$, for parallel algorithm, the time required still is $T_P=4T$, In accordance with the Calculation Formula of the speed-up ratio and efficiency:

$$S_p=\frac{T_l}{T_P}=\frac{\frac{5}{2}(N-1)T}{4T}=\frac{5}{8}(N-1)\text{，}\quad E_p=\frac{S_p}{P}=\frac{\frac{5}{8}(N-1)}{N-1}=0.625$$

So when the dealing with data is N and there are 12 processors available, the theoretical speed-up ratio of the algorithm is 7.5 , the efficiency is 0.625 .

However, based on the current development of science and technology, unlimited multi-processor system is not reasonable in theory, therefore, we also improve the above-mentioned parallel algorithm in the following.

In order to better apply to the actual, we limit the number of processor 4, which is entirely achievable in practice. At this time the data processed is still N , as limiting the number of processors available, so each group only can calculate the divided difference by parallel processing algorithms, and the parallel computing is not consideration between groups, that is, it still be processed by the serial computing between groups. Under the certain number of processors, in accordance with the thinking of the above, algorithm can also improve the efficiency. In the following, we will conduct a brief analysis of the algorithm in the above-mentioned conditions.

At the moment, the dealing with data is N , so the data is divided into $(N-1)/4$ groups, for parallel algorithm, the time required still is $T_P=4T$, The total parallel computing time is $T_p=T\times(N-1)$, In accordance with the Calculation Formula of the speed-up ratio and efficiency:

$$S_p=\frac{T_l}{T_P}=\frac{\frac{5}{2}(N-1)T}{T\times(N-1)}=\frac{5}{2}\quad,\quad E_p=\frac{S_p}{P}=\frac{\frac{5}{2}}{4}=0.625$$

It can be seen that, in this case the theoretical speed up ratio is 2.5 , efficiency is 0.625 . It also to achieve the same satisfactory results in the efficiency, but meanwhile the speed-up ratio is less than the former. So in practical applications, in accordance with the difference purpose, you can select different methods to meet the actual needs.

5. CONCLUSION

In accordance with the dividing-conquering strategy, we design a Parallel Newton interpolation algorithm and give a comprehensive exposition and derivation. We start from concrete examples, and extend it to the general situation. So the traditional Newton interpolation method has been improved. Compared with the serial algorithm, we prove that this algorithm has good parallelism. We also have analyzed the speed-up ratio and efficiency. The superiority of the parallel Newton interpolation algorithm can be seen from the analysis.

REFERENCES

[1] Hildebrand F B.Introduction to Numerical Anaylsis[M]. New York: McGraw-Hill, 1956
[2] Schroeder H, Murthy V K, Krishnamurthy E V. Systolic algrithm for polynomial interpolation and related problems [J]. Parallel Computing, 1991, July: 493-503
[3] Murthy V K, Krishnamurthy E V, Chen P. Systolic algorithm for rational interpolation and Pade approximation[J]. Parallel Computing,1997: 75-83
[4] Jana PK,Sinha BP. Fast parallel algorithm for polynonimal interpolation[J]. Computers Math. Applic. 1995,29:85-92
[5] Aimin YANG and Yiming CHEN, MPI Parallel Programming Environment and Programming Research, 3rd Journal in 2005 of Academy Newspaper of Heibei Polytechnic University, 2005,11:41-44
[6] Chunfeng LIU, Aimin YANG. The Parallel Algorithm of QR Decomposition of Matrix in Cluster System. World Academic Press. 2006:107-111.

Yang Aimin: lecturer, now acting as the Math teacher in Hebei Polytechnic University. He graduated from Yanshan University, majoring in calculus mathematics. Now he is the secretary of Parallel Computing Laboratory and Scientific Computing Laboratory. He has been invited to attend DCABES for 3 years (2006-2008). He has published 3 scientific computing books, over 20 papers and presided over 4 research projects. His research directions include the design and analysis of parallel computation, elastic problems numerical simulation, and etc.

MODELING OF A NEAR FIELD OPTICAL SPOT BASED ON PARALLEL COMPUTING *

Xia Youxin
School of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, 430063, China
Email: xyx@whut.edu.cn

ABSTRACT

The paper analyzes the waveguide exit beam of light field and the near field optical spot energy distribute of the hybrid medium surface. Parallel processing has become such an important and mainstream technology, and multi-core computer will become more common in the future. According to energy balance way, a new model has adopted and the size of near field optical spot has studied based on parallel computing.

Keywords: Hybrid recording, Parallel processing, Near field optical spot, Multi-core computer.

1. INTRODUCTION

The magnetic storage industry is currently on the verge of another paradigm shift as it approaches the superparamagnetic limit [1-3]. With smaller recording medium grains, the energy barrier height for the magnetization reversal of grains becomes lower and thermal fluctuation in the recorded magnetization increases. As a result, the output signals decrease over time. So the high coercivity of the recording medium must adopted, In a hybrid recording system, at normal temperature, magnetic head is not enough to reverse the magnetization bit, the thermal gradient and magnetic field gradient can be used to determine the critical dimensions of a recorded bit. Since the coactivity of the medium is dependent on temperature, an external laser may be used to locally reduce its value to a point where the field from the magnetic head is sufficient to switch the magnetization. The areal density may be decided by the dual-gradient (the thermal gradient and magnetic field gradient)[3,4], the paper suggests a new scheme of hybrid recording by high refractive index optical waveguide, and gives the size of near field optical spot study based on parallel computing. Evanescent energy can not only be utilized directly in near field optical recording, but also in hybrid recording, which will heat the recording medium so as to overcome SFE and increase recording density. Since the size of near field recording spot determines directly the potential of near field recording density, which includes near field optical recording and hybrid recording, the size of near field recording optical spot of the system to be designed should be exactly known during the system design stage.

2. OPTICAL FIELD DISTRIBUTE OF THE WAVEGUIDE EXIT

In our scheme, the track density of the hybrid recording is decided by the width of exit beam of light, now we use the Maxwell electromagnetism theory of optical wave to model and simulate the track density of the scheme.

For the length to width ratio of the rectangular waveguide of the scheme is big; it can be model as three-layer plane waveguide for simply. Shown in Fig.1, Laser beam is transmission by X axis direction, Z-axis is thickness direction of waveguide, and Y axis is width direction of waveguide. Supposing that the refraction index of the core is n1, the refraction index of the clad is n2, the thickness of the core is d, according to the Maxwell electromagnetism theory, the $\vec{E}$ and $\vec{H}$ of the TE mode in the three layers waveguide is following:

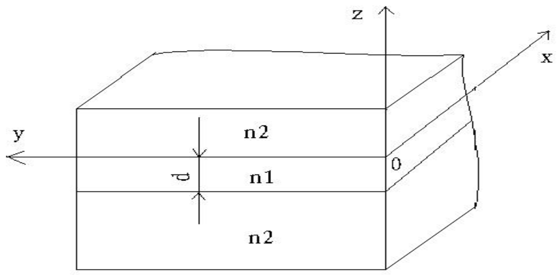


Figure 1. Coordinate of waveguide

$\vec{E} = (0, E_y, 0)$; $\vec{H} = (H_x, 0, H_z)$

When $0 < z < +\infty$:

$$E_y(x,z,t) = E_0 \exp(-\delta z)\exp[-i(\omega t - \beta x)] \tag{1}$$

$$H_x(x,z,t) = -\frac{i\delta}{\omega\mu} E_0 \exp(-\delta z) .\exp[-i(\omega t - \beta x)] \tag{2}$$

$$H_z(x,z,t) = \frac{\beta}{\omega\mu} E_0 \exp(-\delta z) .\exp[-i(\omega t - \beta x)] \tag{3}$$

When $-d < z < 0$:

$$E_y(x,z,t) = E_0\left(\cos\beta_c z - \frac{\delta}{\beta_c}\sin\beta_c z\right) .\exp[-i(\omega t - \beta x)] \tag{4}$$

$$H_x(x,z,t) = -\frac{i\beta_c}{\omega\mu} E_0\left(\sin\beta_c z + \frac{\delta}{\beta_c}\cos\beta_c z\right) .\exp[-i(\omega t - \beta x)] \tag{5}$$

$$H_z(x,z,t) = \frac{\beta}{\omega\mu} E_0\left(\cos\beta_c z - \frac{\delta}{\beta_c}\sin\beta_c z\right) \exp[-i(\omega t - \beta x)] \tag{6}$$

* this paper is sponsored by “Chinese 211 Research Fund for Fundamental Key Projects” (No: 471-38650057) and the Dr Science Foundation of China (No: 20070497042).

When $-\infty < z < -d$:

$$E_y(x,z,t) = E_0\left(\cos\beta_c d + \frac{\delta}{\beta_c}\sin\beta_c d\right)\left(\exp\left[(z+d)\delta\right]\right) \exp\left[-i(\omega t - \beta x)\right] \tag{7}$$

$$H_x(x,z,t) = \frac{i\delta}{\omega\mu}E_0\left(\cos\beta_c d + \frac{\delta}{\beta_c}\sin\beta_c d\right) \left(\exp[(z+d)\delta]\right)\exp[-i(\omega t - \beta x)] \tag{8}$$

$$H_z(x,z,t) = \frac{\beta}{\omega\mu}E_0\left(\cos\beta_c d + \frac{\delta}{\beta_c}\sin\beta_c d\right) \left(\exp[(z+d)\delta]\right)\exp[-i(\omega t - \beta x)] \tag{9}$$

where $\beta_c = \left(k^2 - \beta^2\right)^{\frac{1}{2}}$, $\delta = \left(k_1^2 - k_2^2 - \beta_c^2\right)^{\frac{1}{2}}$, β is the optical transmission constant in the media, k is the wave number for the media. β can be obtained from following equation:

$$\tan(\beta_c d) = \frac{2\beta_c\delta}{\beta_c^2 - \delta^2} \tag{10}$$

The electromagnetism power density of the TE mode in the transmit direction can be obtained by following formula:

$$\langle S\rangle = \frac{1}{2}\mathrm{Re}\left(E_y \cdot H_z^*\right) \tag{11}$$

So we can get the Optical field distribute of the waveguide exit according to the equation (11).

3. STUDY ON THE SIZE OF NEAR FIELD OPTICAL SPOT

When optical wave transmits through free space, if it is in the scope of scalar wave equation, integral methods such as Green function can be used to find the solution. But when the diameter of exit of fiber probe (or waveguide) hole in floating blocks is small than wavelength of laser, and the factor of the parts of apparatus is less than optical wavelength, the scalar equation is not applicable. The solution to vector wave equation should be applied. Now we give a way to get the optical field of the optical exit [5], The optical exit of nanometer hole is assumed as the optical diffraction orifice D, and suppose the distribution of laser wave field at D is f (x,y), this is boundary condition of calculation ,it is needed to find the solution to the distribution of laser wave field over recording medium when the distance between the medium and diffraction orifice is Z. assume that the field distribution is u(x,y,z), so u(x,y,z)=f(x,y) at hole D, f(x,y) can be expressed as Fourier integral:

$$F_0\left(k_x, k_y\right) = \iint_D f_0(x,y) \cdot \exp\left\{-i\cdot\left(xk_x + yk_y\right)\right\}dxdy \tag{12}$$

$$f(x,y,z) = \frac{1}{4\pi^2}\cdot\iint_H F_0\left(k_x, k_y\right) \cdot \exp\left\{i\cdot\left(xk_x + yk_y\right)\right\}dk_x dk_y \tag{13}$$

Now let us find the field solution f (x,y,z) at point (x,y,z). f(x,y,z) is expressed as Fourier integral：

$$f(x,y,z) = \frac{1}{4\pi^2}\cdot\iint_H F\left(k_x, k_y, z\right) \cdot \exp\left\{i\cdot\left(xk_x + yk_y\right)\right\}dk_x dk_y \tag{14}$$

And the Fourier integral formula of f(x,y,z) is used in wave equation

$$\nabla^2 f + k^2 f = 0 \tag{15}$$

Thus we get:

$$\iint\left(\frac{\partial^2 F}{\partial z^2} - \left(k^2 - k_x^2 - k_y^y\right)F\left(k_x, k_y, z\right)\right) \cdot \exp\left\{i\cdot\left(xk_x + yk_y\right)\right\}dk_x dk_y = 0 \tag{16}$$

So

$$\frac{\partial^2 F}{\partial z^2} - \left(k^2 - k_x^2 - k_y^y\right)F\left(k_x, k_y, z\right) = 0 \tag{17}$$

Suppose

$$k_z^2 = k^2 - k_x^2 - k_y^2 \tag{18}$$

Above formula can be written as:

$$\frac{\partial^2 F}{\partial z^2} - k_z^2 F = 0 \tag{19}$$

The solution to this equation is in the following form:

$$F\left(k_x, k_y, z\right) = F_0\left(k_x, k_y\right)\cdot\exp\left(\pm ik_z z\right) \tag{20}$$

Thus, we get the field solution f (x,y,z) at point (x,y,z).

$$f(x,y,z) = \frac{1}{4\pi^2}\cdot\iint_H F_0\left(k_x, k_y\right) \cdot \exp\left\{i\cdot\left(xk_x + yk_y\right)\right\}dk_x dk_y \tag{21}$$

This can be look as the hybrid recording heating optical spot energy distributed [6]. Our project team designs the waveguide with the core thickness of 60nm, longness of 2um.

4. PARALLEL COMPUTING

The differential equations can be changed into homogeneous linear equations; sufficient differential is must for getting accurate solution. In order to improve the accuracy of calculation results, we develop parallel programs and multi-threaded applications on DELL PE2950 quad-core Server, Intel multi-core processors efficiently process multiple tasks simultaneously, providing a breakthrough computing experience in business environments, Multi-core computer will become more common in the future. by Programming with Windows* Threads ,we can Write code to create and terminate threads Use synchronization objects to coordinate thread execution and memory access, Developing new levels of processing performance, multi-core technology enables improved computing experiences while paving the way for the evolution of tera-scale computing. OpenMP uses a shared-memory programming model，Most variables are shared by default. Global variables are shared among threads, But not everything is shared: Stack variables in functions called from parallel regions are private，Automatic variables within a statement block are private.

We can collect sampling data for the program and identify every function that consumes the most execution time by Intel

VTune Performance Analyzer; it evaluates applications on all sizes of systems based on Intel processors, from embedded systems through supercomputers, to help improve application performance. VTune Performance Analyzer makes application performance tuning easier and is indispensable for making our software run it's fastest on the latest multicore systems. It helps tune multi-threaded applications for optimal performance on multicore processors. VTune gives us key performance insights into how our code can scale better, and how we can make more effective use of hardware.

we get the energy distribute in thickness of the waveguide exit optical spot according to section 2, the energy distribute can be looked as the waveguide exit optical spot, the energy distribute of the near field optical spot can be got by the energy distribute of the waveguide exit optical spot according to the section 3.The simulation result of the equation (11) is the boundary condition of the calculation of the equation (12). The energy distribute in thickness of the waveguide exit optical spot can be looked as the scalar field f_0 (x,y), according to the equation (11), we can get the waveguide exit optical spot energy distribute in thickness, the simulation result is show as figure 2.

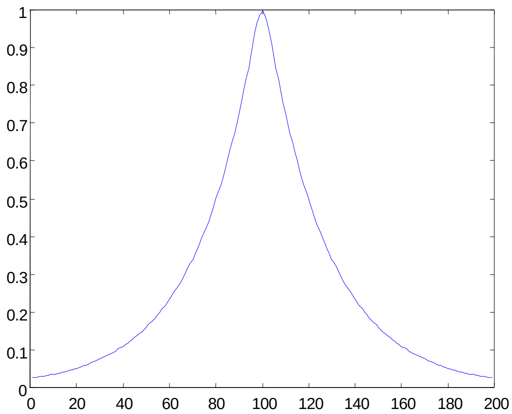


Figure 2. Waveguide exit optical spot energy distribute in thickness

According to the equation (21), we can get the energy distribute of the near field optical spot; the simulation result is show as figure 3 and figure 4. Here, the XY coordinate plane is the wavelength, Z axis is the relative light intensity (center of light intensity is 1).In Figure 3, the fiber parameters are: core diameter = 10.0 wavelength; cladding refractive index = 1.498; core refractive index = 1.5. In Figure b, the fiber parameters are: the entrance of core diameter = 10.0 wavelength; entrance clad diameter = 20.0 wavelength; export core diameter = 0.0625 wavelength; exports cladding diameter (including core) = 0.125 nm; cladding refractive index = 1.498; core refractive index = 1.5; fiber cone tip length = 397.5 wavelengths. We can modulate the waveguide parameter adopted the different application. This needs Largeness simulation computation.

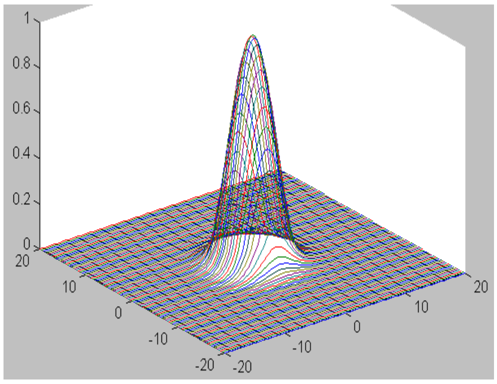


Figure3. Single mode fiber near optical spot energy distribute

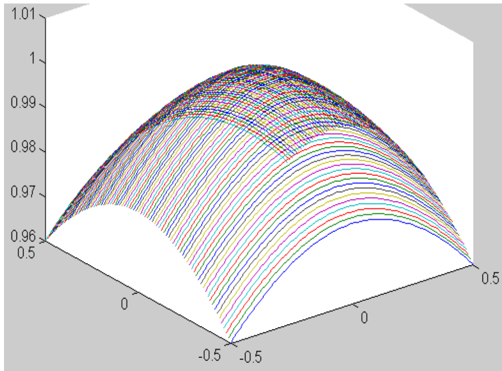


Figure4. Waveguide near optical spot energy distribute

5. CONCLUSION

We succeed to use the Maxwell electromagnetism theory to model and simulate the waveguide exit energy distribute, obtained the near field optical spot of the hybrid recording medium surface by Green function.

REFERENCE

[1] Hiroyuki Katayama, Masaki Hamamoto, Jun-ichi Sato, New Developments in Laser-Assisted Magnetic Recording. IEEE Trans Magn, 2000, 1, 36(1): 195-198

[2] J. J. M. Ruigrok, R. Coehoorn, S. R. Cumpson, H. W. van Kesteren, Disk Recording Beyond 100Gb/in2: Hybrid Recording. J. Appl. Phys. 87, 2000:5398-5401

[3] Hirofumi Sukeda, Hideki Saga, Hiroaki Nemoto. Thermally Assisted Magnetic Recording on Flux-Detectable RE-TM Media. IEEE Trans Magn, 2001, 7, 37(4): 4: 1234-1237

[4] huanghao,xiayouxin. Research on slider bearing of the hybrid recording. Optical storage (ISOS 2005),2005,4,5966OJ:1-4

[5] Huanghao. Research on some key technologies for near-field recording disc drives (D). HuaZhong University of Science and Technology: library, 1999

[6] Peixiandeng, Xiayouxin. Study on the thermal effect for hybrid record. SPIE, 5060: 300-303

RESEARCH OF EVALUATION METHOD OF TEACHING QUALITY BASED ON GRAY TREND RELATIONAL ANALYSIS

CHE Hai-ning[1], ZHAO Tian-fang[2], LI Zhi-jun[2]
[1] The modern educating center, Wuhan University of technology, Wuhan 430063, China
[2] School of Automation, Wuhan University of technology, Wuhan 430063, China
Email:chn826@163.com; xiaobianb2006@163.com; junwenlee@sohu.com

ABSTRACT

The gray trend relational system is discussed in this paper by means of gray trend relational degree and general system theory. As the application of this model, the gray trend relational clustering method is put forwarded. Taking the rating of five teachers in a teaching evaluation as data mining object, 5 teachers' teaching contents, teaching methods, teaching attitude and basic teaching skills and teaching effects were evaluated and calculated using the gray relational clustering methods, and the clustering result is in accord with fact. It shows that the gray trend clustering method is efficient, convenient and applicable.

Keywords: Teaching Evaluation, Fuzzy Clustering, Trend Relational, Data Mining

1. INTRODUCTION

The evaluation and inspection of teaching quality is the primary coverage of management of education quality, they are highly related to teachers' front-line teaching. The evaluation and inspection of teaching quality matters a lot, they could improve the quality of teaching and bring it under control, and they could also be used as a reference for the teacher's career prospect. However, the evaluation and inspection of teaching quality are intractable. Technically, teaching is complicated, specialized and creative, therefore, to evaluate and inspect it, many spheres of knowledge is needed, such as pedagogy, psychology, etc. As the bases used in specific evaluation selected by teachers, peer evaluation is the most important, then is the student response and student rating, quantification assessment and leader-check-up-on-teaching follow. In all these bases, quantification assessment is a comprehensive one. The depth of quantification plays an important part in grading the style of teaching quality management [1].

According to above opinion, this paper has investigated a gray trend relational system model, and combined gray trend relational analysis with fuzzy clustering to form a new data mining method. Instructed by this model, this paper discusses gray trend relational clustering method and its application in the data mining.

2. TREND RELATIONAL ANALYSIS

Let original time sequence aggregation is $\left\{X_i^{(0)}\right\}$. Suppose that there are h investigated objects. Then the correspond time sequence is

$$X_i^{(0)} = \left\{X_i^{(0)}(k)\right\} \qquad k = 1,2,\cdots,n \qquad i = 1,2,\cdots,h \tag{1}$$

Suppose that there are two factors: $X_r^{(0)}$ and $X_c^{(0)}$. We call $X_r^{(0)}$ reference factor, and $X_c^{(0)}$ relative factor. Correspondingly, we call $\left\{X_r^{(0)}(k)\right\}$ reference time sequence, and $\left\{X_c^{(0)}(k)\right\}$ compare time sequence, $r, c \in \{1,2,...,h\} = H$.

In order to compare the comparability and proximity between the time sequence $\left\{X_r^{(0)}(k)\right\}$ and $\left\{X_c^{(0)}(k)\right\}$, we dispose the above three time sequences as follow:

$$\Delta X_{rr}^{(0)}(k) = X_r^{(0)}(k) - X_r^{(0)}(k-1) \qquad k = 2,3,\cdots,n \tag{2}$$

$$\Delta X_{cc}^{(0)}(k) = X_c^{(0)}(k) - X_c^{(0)}(k-1) \qquad k = 2,3,\cdots,n \tag{3}$$

$$\Delta X_{rc}^{(0)}(k) = X_r^{(0)}(k) - X_c^{(0)}(k) \qquad k = 1,2,\cdots,n \tag{4}$$

Definition 1 If mapping

$$\left\{\Delta X_{rc}^{(0)}(k)\right\} \times \left\{\Delta X_{rr}^{(0)}(k)\right\} \times \left\{\Delta X_{cc}^{(0)}(k)\right\} \to \varsigma(k) \in [0,1] \tag{5}$$

Then we call $\varsigma(k)$ as trend relational function between $\left\{X_r^{(0)}(k)\right\}$ and $\left\{X_c^{(0)}(k)\right\}$, expressing the comparability and proximity.

Obviously, trend relational function $\varsigma(k)$ is one relationship among factors in the trend relational space, which has normalization, symmetry and proximity. If trend relational function $\varsigma(k)$ =const, $\left\{X_r^{(0)}(k)\right\}$ and $\left\{X_c^{(0)}(k)\right\}$ is proved that they have completely "common" behaviors, which is similar to the curves of $\left\{X_r^{(0)}(k)\right\}$ and $\left\{X_c^{(0)}(k)\right\}$ parallel. The more common and closer, the larger of $\varsigma(k)$; when they are equal, $\varsigma(k)$ =1[3].

Theorem 1[2] Suppose that $\left\{X_r^{(0)}(k)\right\} \in \left\{X^{(0)}(k)\right\}$, $\left\{X_c^{(0)}(k)\right\} \in \left\{X^{(0)}(k)\right\}$, $r, c \in H$. Let

$$\varsigma_{rc}(k) = \varsigma_{rc}\left(\left\{X_r^{(0)}(k)\right\}, \left\{X_c^{(0)}(k)\right\}\right) = \frac{1}{1+\rho\left|\Delta X_{rc}^{(0)}(k)+\Delta X_{rc}^{(0)}(k-1)\right| + \left|\Delta X_{rr}^{(0)}(k)-\Delta X_{cc}^{(0)}(k)\right|} \qquad k = 2,3,\cdots,n \tag{6}$$

Where, $\rho \in [0,1]$. Generally, let $\rho = 0.5$. Then $\varsigma_{rc}(k)$ is called a kind of trend relational function between $\left\{X_r^{(0)}(k)\right\}$ and $\left\{X_c^{(0)}(k)\right\}$.

Definition 2 If $\varsigma_{rc}(k)$ is trend relational function between

$\{X_r^{(0)}(k)\}$ and $\{X_c^{(0)}(k)\}$, then we call

$$\Xi_{rc} = \frac{1}{n-1}\sum_{k=2}^{n}\varsigma_{rc}(k) \tag{7}$$

as trend relational degree between $\{X_r^{(0)}(k)\}$ and $\{X_c^{(0)}(k)\}$.

Definition 3 Suppose that original time sequence is X, and there are h factors in X:

$$X = \{x_i(k) | , k = 1,2,...,n, \quad i = 1,2,...,h\}$$

If there is a specific mapping $\gamma \in \Gamma$:

$(x_r, x_c) \to \Xi_{rc}, \Xi_{rc} \in [0,1]$, $x_r, x_c \in X$, $r, c \in H = \{1,2,...,h\}$, where Ξ_{rc} is trend relational degree between factor x_r, x_c, factor aggregation X and trend relational mapping aggregation Γ constitute a trend relational space, marked by (X,Γ) [4].

We call

$$R = \begin{bmatrix} \Xi_{11} & \Xi_{12} & \cdots & \Xi_{1h} \\ \Xi_{21} & \Xi_{22} & \cdots & \Xi_{2h} \\ & & \cdots & \\ \Xi_{h1} & \Xi_{h2} & \cdots & \Xi_{hh} \end{bmatrix} \tag{8}$$

as trend relational matrix on the trend relational space (X,Γ).

3. ESTABLISHING INDEX SYSTEM FOR THE EVALUATION OF TEACHING QUALITY

What teaching quality evaluation does is to evaluate the course and the achievement of teaching, which graded the teaching and reviewed the advantages and disadvantages of it to make improvement. Some principles should be followed to set up a scientific index system for the evaluation of teaching quality. They are:1)The feature of teaching in colleges should be reflected in the index system, making the evaluation the guidance for teaching.2)To be a system, not only the abundant information in the course of teaching should the evaluation provide, but also the feedback on the teaching. So the evaluation could give a systematic assessment to all aspects of education, and diagnose the teaching, improving the quality of teaching.3) Well-founded, in other words, the indexes must be scientific selected. The indexes and their weights should be dynamic, reflecting the focus of teaching task and the policies of education department in different periods; Indexes should be independent of each other, being different in connotation and denotation; Quantitative indexes should combine with qualitative ones, and the index system must be capable of fault-tolerant, so the subjective factors can be wipe off. The participation should be highlight in the index designing [5,6]. Because the more the teachers or students participate in and coordinate, the more validity the index is, and the more accuracy evaluation. Establishing index system is the most basic and crucial job in the evaluation. On these principles, taken the assessment of the quality of classroom teaching for instance, an index system is set up suited to the evaluation of teachers' teaching quality by conducting a survey to some teachers in our college. As shown in Table 1 are the indexes [7].

Table 1. Index System for the Evaluation of Teaching Quality

principal element		secondary element	
name	weight	name	weight
teaching content	0.3467	clear theory explanation	0.2
		compact and abundant content	0.2
		suitable textbook	0.25
		related to research projects	0.2
		teaching to cultivate	0.15
teaching method	0.2823	brief and focused, appropriate	0.25
		explanation of detail enlightening, guide students to think actively	0.25
		multiple teaching methods combined	0.25
		communicate and interact with students	0.25
teaching attitude	0.1567	well prepared and rigorous on scholarship	0.5
		responsible and be all ears for the views of students	0.25
		have a awareness of service, treat students equally	0.25
basic skills	0.1054	accurate and rigorous in language expression	0.5
		blackboard-writing and the courseware are clear and organized	0.35
		neat and grave in appearance	0.15
teaching effect	0.1089	achieve the purpose of teaching	0.1
		have a meaning of guidance in practice	0.4
		inspiring, students are motivated and the knowledge are well absorbed	0.5

4. AN EXAMPLE ON APPLYING TREND RELATIONAL SPACE

Taking an evaluation of our school for instance, we evaluated the quality of classroom teaching by the method and model mentioned above. The assessments of five teachers' teaching quality by the education expert group are selected at random, then the analysis and calculation for these assessment are conducted. The ten-people education expert group graded the specific index by means of calculating the arithmetic means of secondary indexes. The original indexes and related information can be found in reference [1].

Five teachers' assessment sequences are gained by calculating the original data form reference [1]:

$$X_1 = [0.42, 0.60, 0.70, 0.40, 0.60]$$

$$X_2 = [0.52, 0.42, 0.74, 0.32, 0.56]$$

$$X_3 = [0.58, 0.56, 0.74, 0.52, 0.52]$$

$$X_4 = [0.52, 0.40, 0.68, 0.40, 0.38]$$

$$X_5 = [0.40, 0.52, 0.60, 0.32, 0.38]$$

Reference sequences are gained by upper limit result measure:

$$X_r = [0.58, 0.60, 0.74, 0.52, 0.60]$$

From formula (2):

$$\Delta X_{rr} = [0.02, 0.14, -0.22, 0.08]$$

From formula (3) and (4), Table 2 and Table 3 are gained.

Table 2. $\Delta X_{cc}(k)$

k	$\Delta X_{11}(k)$	$\Delta X_{22}(k)$	$\Delta X_{33}(k)$	$\Delta X_{44}(k)$	$\Delta X_{55}(k)$
2	0.18	-0.10	-0.02	-0.12	0.12
3	0.10	0.32	0.18	0.28	0.08
4	-0.30	-0.42	-0.22	-0.28	-0.28
5	0.20	0.24	0.00	-0.02	0.06

Table 3. $\Delta X_{rc}(k)$

k	$\Delta X_{r1}(k)$	$\Delta X_{r2}(k)$	$\Delta X_{r3}(k)$	$\Delta X_{r4}(k)$	$\Delta X_{r5}(k)$
1	0.14	0.06	0.00	0.06	0.18
2	0.00	0.18	0.04	0.20	0.08
3	0.04	0.00	0.00	0.14	0.14
4	0.12	0.20	0.00	0.12	0.20
5	0.00	0.04	0.08	0.22	0.22

According to $\Delta X_{rr}(k)$ 、$\Delta X_{cc}(k)$ 、$\Delta X_{cc}(k)$, utilizing formula(6) and (7), trends incidence matrix is obtained:

$$R = \begin{bmatrix} 1 & 0.865 & 0.788 & 0.943 & 0.793 & 0.820 \\ 0.865 & 1 & 0.835 & 0.850 & 0.810 & 0.870 \\ 0.788 & 0.835 & 1 & 0.803 & 0.883 & 0.805 \\ 0.943 & 0.850 & 0.803 & 1 & 0.850 & 0.815 \\ 0.793 & 0.810 & 0.883 & 0.850 & 1 & 0.857 \\ 0.820 & 0.870 & 0.805 & 0.815 & 0.857 & 1 \end{bmatrix} \quad (9)$$

Put matrix R in to compositional operation, the follow matrix R^2, R^4, R^8, R^{16} are obtained:

$$R^2 = \begin{bmatrix} 1 & 0.865 & 0.835 & 0.943 & 0.850 & 0.865 \\ 0.865 & 1 & 0.835 & 0.865 & 0.857 & 0.870 \\ 0.835 & 0.835 & 1 & 0.850 & 0.883 & 0.857 \\ 0.943 & 0.865 & 0.850 & 1 & 0.850 & 0.850 \\ 0.850 & 0.857 & 0.883 & 0.850 & 1 & 0.857 \\ 0.865 & 0.870 & 0.857 & 0.850 & 0.857 & 1 \end{bmatrix}$$

$$R^4 = \begin{bmatrix} 1 & 0.865 & 0.857 & 0.943 & 0.857 & 0.865 \\ 0.865 & 1 & 0.857 & 0.865 & 0.857 & 0.870 \\ 0.857 & 0.857 & 1 & 0.850 & 0.883 & 0.857 \\ 0.943 & 0.865 & 0.850 & 1 & 0.857 & 0.865 \\ 0.857 & 0.857 & 0.883 & 0.857 & 1 & 0.857 \\ 0.865 & 0.870 & 0.857 & 0.865 & 0.857 & 1 \end{bmatrix}$$

$$R^8 = \begin{bmatrix} 1 & 0.865 & 0.857 & 0.943 & 0.857 & 0.865 \\ 0.865 & 1 & 0.857 & 0.865 & 0.857 & 0.870 \\ 0.857 & 0.857 & 1 & 0.857 & 0.883 & 0.857 \\ 0.943 & 0.865 & 0.857 & 1 & 0.857 & 0.865 \\ 0.857 & 0.857 & 0.883 & 0.857 & 1 & 0.857 \\ 0.865 & 0.870 & 0.857 & 0.865 & 0.857 & 1 \end{bmatrix}$$

$$R^{16} = \begin{bmatrix} 1 & 0.865 & 0.857 & 0.943 & 0.857 & 0.865 \\ 0.865 & 1 & 0.857 & 0.865 & 0.857 & 0.870 \\ 0.857 & 0.857 & 1 & 0.857 & 0.883 & 0.857 \\ 0.943 & 0.865 & 0.857 & 1 & 0.857 & 0.865 \\ 0.857 & 0.857 & 0.883 & 0.857 & 1 & 0.857 \\ 0.865 & 0.870 & 0.857 & 0.865 & 0.857 & 1 \end{bmatrix}$$

Obviously, $R^{16} = R^8$, then transitive closure $\tilde{R}$ of trend relational matrix is equation (11). $\tilde{R}$ is a fuzzy equivalence relation. Let the threshold value be 0.943 and 0.865, we can obtain respective clustering matrix[4,8].

$$\tilde{R}_{0.943} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\tilde{R}_{0.865} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

According to the clustering relation matrix definitized by the given threshold value, we can classify the trend relational degree into 3 kinds: {Teacher_r, Teacher_3}, {Teacher_1, Teacher_5}, {Teacher_2, Teacher_4}. Obviously, Teacher_r, Teacher_3 belongs to class 1; Teacher_1, Teacher_5 belongs to class 2; Teacher_2, Teacher_4 belongs to class 3. If we classify the teachers into three degrees (excellence, good, common), we can know that Teacher_3 belongs to excellent; Teacher_1 and Teacher_5 belongs to good, and Teacher_2 and Teacher_4 belongs to common.

5. CONCLUSIONS

To adjust to the reform and development of teaching in college nowadays, principle of establishing an index system for the evaluation of teaching quality suit to colleges in our country is proposed in this paper, and the teaching evaluation model with feedback mechanism is built, which make the interaction between teaching and learning perfect. It is helpful to arouse enthusiasm in teachers and students both sides, and play a positive role in improving teaching quality of colleges in our country. This paper takes full account of the weights of all the indexes, which minimize the "egalitarianism" problem. Thus the assessment is more reasonable and the strategy decision is more reliable. According to the analysis of the teaching quality evaluation, it is evident that the trend relational analysis is much more effective. Fuzzy comprehensive evaluation combined with grey system theory gives full play to both strengths and a better effect is gained. The trend relational method is step-definitely, reliable and promising, it has a positive effect on the evaluation of teaching quality for colleges.

REFERENCES

[1] Ma Xing. Research of Evaluation Method of Teaching Quality Based on Hierarchy Correlation[J]. Journal of Wuhan University of Technology(Information & Management Engineering), 2007,29(5)：122~125

[2] Mianyun Chen. Trend Relational Degree and Its Application in Gray System Modeling[J]. Huazhong University of Sciences and Technology Transaction, 1994, 22(8):66~68.

[3] Mianyun Chen, Zhijun Li, Long Zhou et al. SCGM-model and grey control of “poor” information systems[J]. Kybernetes, 2004, 33(2):231~237.

[4] LI Zhi-jun, ZHANG Lei, CHEN Mian-yun, et al. Research and Application of Data Mining Method Based on Gray Trend Relational Analysis. Proceeding of PACIIA 2008, Vol. 2:98-101.

[5] Li Mingliang. A new Descriminant Byelaw for Grey Interconnect Degree and Its Calculation Formulas. System Engineering, 1998,16(1): 68~70

[6] ZHAN Qian-yong. The Grey Multi-Hierarchy Decision Model and Its Application in the Evaluation of Quality of Test Papers[J]. System Engineering-theory & Practice, 2000, (7):135~138

[7] SU Jun-min. Evaluation of the Education Management by AHP-Grey Correlation Anaysis[J]. Journal of Tianjin University of Technology and Education, 2003, 13(1)：30~33

[8] HE Qing, LI Hongxing. Fuzzy Equivalent Matrices in Fuzzy Clustering[J]. System Project Theory and Practice, 1999, (4): 8~12

Che Hai-ning is a engineer of Morden Education Technology Centre of Wuhan University of Technology. He graduated from Huazhong University of Science and Technology in 1988 with specialty of electronic automation. He has published over 16 journal papers.His research interests are in electric equipment control and data mining.

THE CALCULATION OF THE TETRAQUARK($\bar{c}\bar{c}qq$) INTERACTION ENERGY *

Zhang Ao, **Tian Yi-Min**
Science Education Department, Beijing Institute of Graphic Communication
Beijing ,China
Email: ahf200805@sina.com , tym@bigc.edu.cn

ABSTRACT

Using the color-magnetic interaction potential energy with the color symmetry, we perform a schematic study of the tetraquarks($\bar{c}\bar{c}qq$). By group theory way, we can calculate the S-wave tetraquarks ($\bar{c}\bar{c}qq$) interaction energy of a colored-quark-gluon model, which based on a semi-classical approximation to the MIT bag theory..

Keywords: Tetraquark, Interaction Energy

1. INTRODUCTION

Since LEPS collaboration announced the surprising discovery of Θ^+ pentaquark, many experimental groups have claimed the observation of evidence of its existence [1]. In particular, spin-dependence in the quark interaction may arise due to one gluon exchange, coupled channel effects, and relativistic dynamics. There is good evidence of the $J^{pc}=1^{-+}$ exotic mesons which can not be $q\bar{q}$ mesons.

Although one may hope that short range quark interaction are described by one gluon exchange ,this is far from obvious in the intermediate regime relevant to hadrons[2]. We are at variance with previous potential mode studies because they have failed to properly take account an important long-range color-mixing effect [3]. In this paper we calculate the S-wave tetraquarks ($\bar{c}\bar{c}qq$) interaction energy of a colored-quark-gluon model based on a semi-classical approximation to the MIT bag theory.

2. A MEDOL FOR THE $\bar{c}\bar{c}qq$ SYSTEM

2.1 The spin is one between $\bar{q}$ and q

The constituent quark model (CQM) is quite successful in the description of the meson and baryon spectrum. Firstly, we discuss that the spin is one between $\bar{q}$ and q . For the multi-quark system, the model interaction energy [4], [5] reads:

$$V=V_{conf}+V_{SD}$$
$$=-\frac{4}{3}\frac{\alpha_s}{r}+br+\frac{1}{3m_1m_2}\overline{S_1}\cdot\overline{S_2}V_4 \tag{1}$$

Where

$$V_4=\frac{32\pi\alpha_s}{3}\left(\frac{\sigma}{\sqrt{\pi}}\right)^3e^{-\sigma^2r^2}$$
$$=\frac{32\pi\alpha_s}{3}\tilde{\delta}_\sigma(r) \tag{2}$$

V_{conf} is the harmonic confinement potential and V_{SD} is the color hyperfine interaction. By Eq.(1)and Eq.(2), we can get the interaction energy:

$$V=-\frac{4}{3}\frac{\alpha_s}{r}+br+\frac{32\pi\alpha_s}{9m_1m_2}\left(\frac{\sigma}{\sqrt{\pi}}\right)^3e^{-\sigma^2r^2}\overline{S_1}\cdot\overline{S_1} \tag{3}$$

The ground state wave function is

$$\mathrm{R}_{00}(\mathrm{r})=\frac{2\beta^{\frac{3}{2}}}{\pi^{\frac{1}{4}}}\exp\left(-\frac{\beta^2r^2}{2}\right) \tag{4}$$

Then we can get

$$<\mathrm{R}_{00}(\mathrm{r})\left|\frac{1}{\mathrm{r}}\right|\mathrm{R}_{00}(\mathrm{r})>=\frac{2\beta}{\pi^{\frac{1}{2}}} \tag{5}$$

$$<\mathrm{R}_{00}(\mathrm{r})\left|\mathrm{r}\right|\mathrm{R}_{00}(\mathrm{r})>=\frac{2}{\pi^{\frac{1}{2}}\beta} \tag{6}$$

$$<\mathrm{R}_{00}(\mathrm{r})\left|\tilde{\delta}_\sigma(r)\right|\mathrm{R}_{00}(\mathrm{r})>=\frac{1}{\pi^{\frac{3}{2}}}\left[\frac{\beta^2\sigma^2}{\beta^2+\sigma^2}\right]^{\frac{3}{2}} \tag{7}$$

$$<\overline{S_1}\cdot\overline{S_1}>=\frac{1}{2}[S(S+1)-S_1(S_1+1)-S_2(S_2+1)]$$
$$=\frac{1}{2}[S(S+1)-\frac{3}{4}]$$
$$=\begin{cases}-\frac{3}{4},S=0\\ \frac{1}{4},S=1\end{cases} \tag{8}$$

By the Eq.(3)-(8), we can get

$$<V>_{cc}=-\frac{8}{3}\alpha_s\frac{\beta_{cc}}{\sqrt{\pi}}+\frac{2b}{\beta_{cc}\sqrt{\pi}}-\frac{8\alpha_s}{3m_c^2\sqrt{\pi}}\left[\frac{\beta_{cc}^2\sigma^2}{\beta_{cc}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{9}$$

$$<V>_{cq}=-\frac{8}{3}\alpha_s\frac{\beta_{cq}}{\sqrt{\pi}}+\frac{2b}{\beta_{cq}\sqrt{\pi}}-\frac{8\alpha_s}{3m_cm_q\sqrt{\pi}}\left[\frac{\beta_{cq}^2\sigma^2}{\beta_{cq}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{10}$$

$$<V>_{qq}=-\frac{8}{3}\alpha_s\frac{\beta_{qq}}{\sqrt{\pi}}+\frac{2b}{\beta_{qq}\sqrt{\pi}}+\frac{8\alpha_s}{9m_q^2\sqrt{\pi}}\left[\frac{\beta_{qq}^2\sigma^2}{\beta_{qq}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{11}$$

*This project was supported by
financially supported by NSFC(F60850007) ,BJSFC(A1092012) and BIGC (Ea-09-14).

$$
\begin{aligned}
< V > &= < V >_{cc} + 4 < V >_{cq} + < V >_{qq} \\
&= -\frac{8}{3}\alpha_s \frac{1}{\sqrt{\pi}}[\beta_{cc} + 4\beta_{cq} + \beta_{qq}] \\
&+ \frac{2b}{\sqrt{\pi}}[\frac{1}{\beta_{cc}} + \frac{4}{\beta_{cq}} + \frac{1}{\beta_{qq}}] \\
&- \frac{8\alpha_s}{3\sqrt{\pi}}\{\frac{1}{m^2{}_c}\left[\frac{\beta_{cc}{}^2\sigma^2}{\beta_{cc}{}^2+\sigma^2}\right]^{\frac{3}{2}} \\
&+ \frac{4}{m_c m_q}\left[\frac{\beta_{cq}{}^2\sigma^2}{\beta_{cq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \\
&- \frac{1}{3m_q{}^2}\left[\frac{\beta_{qq}{}^2\sigma^2}{\beta_{qq}{}^2+\sigma^2}\right]^{\frac{3}{2}}\}
\end{aligned} \tag{12}
$$

2.2 The spin is one between c and q

Secondly, we discuss that the spin is one between c and q. By the same way, we can get the equations:

$$
< V >_{cc} = -\frac{8}{3}\alpha_s \frac{\beta_{cc}}{\sqrt{\pi}} + \frac{2b}{\beta_{cc}\sqrt{\pi}} - \frac{8\alpha_s}{3m_c{}^2\sqrt{\pi}}\left[\frac{\beta_{cc}{}^2\sigma^2}{\beta_{cc}{}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{13}
$$

$$
< V >_{cq} = -\frac{8}{3}\alpha_s \frac{\beta_{cq}}{\sqrt{\pi}} + \frac{2b}{\beta_{cq}\sqrt{\pi}} - \frac{8\alpha_s}{3m_c m_q\sqrt{\pi}}\left[\frac{\beta_{cq}{}^2\sigma^2}{\beta_{cq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{14}
$$

$$
< V >'_{cq} = -\frac{8}{3}\alpha_s \frac{\beta'_{cq}}{\sqrt{\pi}} + \frac{2b}{\beta'_{cq}\sqrt{\pi}} - \frac{8\alpha_s}{3m_c m_q\sqrt{\pi}}\left[\frac{\beta'_{cq}{}^2\sigma^2}{\beta'_{cq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{15}
$$

$$
< V >_{qq} = -\frac{8}{3}\alpha_s \frac{\beta_{qq}}{\sqrt{\pi}} + \frac{2b}{\beta_{qq}\sqrt{\pi}} - \frac{8\alpha_s}{3m_q{}^2\sqrt{\pi}}\left[\frac{\beta_{qq}{}^2\sigma^2}{\beta_{qq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{16}
$$

$$
\begin{aligned}
< V > &= < V >_{cc} + 3 < V >_{cq} + < V >'_{cq} + < V >_{qq} \\
&= -\frac{8}{3}\alpha_s \frac{1}{\sqrt{\pi}}[\beta_{cc} + 3\beta_{cq} + \beta'_{cq} + \beta_{qq}] \\
&+ \frac{2b}{\sqrt{\pi}}[\frac{1}{\beta_{cc}} + \frac{3}{\beta_{cq}} + \frac{1}{\beta'_{cq}} + \frac{1}{\beta_{qq}}] \\
&- \frac{8\alpha_s}{3\sqrt{\pi}}\{\frac{1}{m^2{}_c}\left[\frac{\beta_{cc}{}^2\sigma^2}{\beta_{cc}{}^2+\sigma^2}\right]^{\frac{3}{2}} + \frac{3}{m_c m_q}\left[\frac{\beta_{cq}{}^2\sigma^2}{\beta_{cq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \\
&- \frac{1}{3m_c m_q}\left[\frac{\beta'_{cq}{}^2\sigma^2}{\beta'_{cq}{}^2+\sigma^2}\right]^{\frac{3}{2}} + \frac{1}{m_q{}^2}\left[\frac{\beta_{qq}{}^2\sigma^2}{\beta_{qq}{}^2+\sigma^2}\right]^{\frac{3}{2}}\}
\end{aligned} \tag{17}
$$

2.3 The spin is one between c and $\bar{c}$

Thirdly, we discuss that the spin is one between c and $\bar{c}$. By the same way, we can get the equations:

$$
< V >_{cc} = -\frac{8}{3}\alpha_s \frac{\beta_{cc}}{\sqrt{\pi}} + \frac{2b}{\beta_{cc}\sqrt{\pi}} + \frac{8\alpha_s}{9m_c{}^2\sqrt{\pi}}\left[\frac{\beta_{cc}{}^2\sigma^2}{\beta_{cc}{}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{18}
$$

$$
< V >_{cq} = -\frac{8}{3}\alpha_s \frac{\beta_{cq}}{\sqrt{\pi}} + \frac{2b}{\beta_{cq}\sqrt{\pi}} - \frac{8\alpha_s}{3m_c m_q\sqrt{\pi}}\left[\frac{\beta_{cq}{}^2\sigma^2}{\beta_{cq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{19}
$$

$$
< V >_{qq} = -\frac{8}{3}\alpha_s \frac{\beta_{qq}}{\sqrt{\pi}} + \frac{2b}{\beta_{qq}\sqrt{\pi}} - \frac{8\alpha_s}{3m_q{}^2\sqrt{\pi}}\left[\frac{\beta_{qq}{}^2\sigma^2}{\beta_{qq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \tag{20}
$$

$$
\begin{aligned}
< V > &= < V >_{cc} + 4 < V >_{cq} + < V >_{qq} \\
&= -\frac{8}{3}\alpha_s \frac{1}{\sqrt{\pi}}[\beta_{cc} + 4\beta_{cq} + \beta_{qq}] \\
&+ \frac{2b}{\sqrt{\pi}}[\frac{1}{\beta_{cc}} + \frac{4}{\beta_{cq}} + \frac{1}{\beta_{qq}}] \\
&- \frac{8\alpha_s}{3\sqrt{\pi}}\{-\frac{1}{3m^2{}_c}\left[\frac{\beta_{cc}{}^2\sigma^2}{\beta_{cc}{}^2+\sigma^2}\right]^{\frac{3}{2}} \\
&+ \frac{4}{m_c m_q}\left[\frac{\beta_{cq}{}^2\sigma^2}{\beta_{cq}{}^2+\sigma^2}\right]^{\frac{3}{2}} \\
&+ \frac{1}{m_q{}^2}\left[\frac{\beta_{qq}{}^2\sigma^2}{\beta_{qq}{}^2+\sigma^2}\right]^{\frac{3}{2}}\}
\end{aligned} \tag{21}
$$

We get the three kinds of interaction energy of the tetraquarks($cc\bar{q}\bar{q}$).

3. DISCUSSIONS

In short summary, we have calculated the tetraquarks($cc\bar{q}\bar{q}$) interaction energy. Despite the several weaknesses of our model of the $cc\bar{q}\bar{q}$ system, we think it may provide a qualitatively valid picture of this simplest mulitiquark. If future experimental searches success to find the tetraquarks($cc\bar{q}\bar{q}$), it is sure that the Lie group in the quark model is quite successful.

REFERENCES

[1].A. Zhang, Y.R. Liu, P.Z. Huang, W.Z. Deng, X.L. Chen (Peking U.) , Shi-Lin Zhu,High Energy Phys.Nucl.Phys.29: 250 , 2005

[2].Ying Cui, Xiao-Lin Chen, Wei-Zhen Deng, Shi-LinZhu, Phys.Rev.D73: 014018 , 2006

[3].Ying Cui, Xiao-Lin Chen, Wei-Zhen Deng, Shi-Lin Zhu, High Energy Phys. Nucl .Phys. 31:7-13,2007.

[4].John Weinstein, Nathan Isgur, Phys. Rev.D27:588, 1983

[5].R.Jaffe and F.Wilczek, Phys. Rev. Lett. 91, 232003(2003).

CONFORMAL FLAT MANIFOLDS AND A PINCHING PROBLEM ON THE SCHOUTEN TENSOR *

Nan Ji, Xinghua Ma, Yunwei Xia and Yamian Peng
College of Science, Hebei Polytechnic University
Tangshan, Hebei Province, 063000,China
Email: pumpkinji@163.com

ABSTRACT

In this paper, we study the Schouten tensor, which is expressed by the Ricci curvature and scalar curvature. We find that it is a Codazzi tensor on a Riemannian manifold M (dim M>3) with harmonic Weyl conformal curvature tensor. Thus we regard it as a natural generalization of the second fundamental form of a hypersurface in real space form. Furthermore, we get some theorems and results. By using this tensor, we induce an operator, which is self-adjoint relative to the L^2 inner product. Then we characterize Einstein manifold and constant sectional curvature by inequalities between certain function on a compact local conformally flat space. This observation leads to a pinching theorem on the Schouten tensor.

Keywords: Conformally flat manifold, Schouten tensor, Pinching problem.

1. INTRODUCTION

Let M be a n-dimentional Riemannian manifold, $e_1,e_2,\cdots,e_n$ a local orthonormal frame field on M, and $\omega_1,\omega_2,\cdots,\omega_n$ its dual frame field. Then the structure equation of M are given by

$$d\omega_i=\sum_j\omega_j\wedge\omega_{ji},d\omega_{ij}=\sum_l\omega_{il}\wedge\omega_{lj}-\frac{1}{2}\sum_{k,l}R_{ijkl}\omega_k\wedge\omega_l \tag{1.1}$$

where ω_{ij} is the Levi-civita connection and R_{ijkl} is the Riemannian curvature tensor of M. Ricci tensor R_{ij} and scalar curvature are defined respectively by

$$R_{ij}:=\sum_k R_{kikj},r:=\sum_k R_{kk} \tag{1.2}$$

Now, we define Schouten tensor as the following (cf[2])

$$S_{ij}:=R_{ij}-\frac{1}{2(n-1)}r\delta_{ij} \tag{1.3}$$

Then $S_{ij}=S_{ji}$, and through the straight calculation we can get

$$trS=\frac{n-2}{2(n-1)}r,\|S\|^2=\sum_{i,j}R_{ij}^2-\frac{3n-4}{4(n-1)^2}r^2 \tag{1.4}$$

The we induce an operator □, which is similar to the operator introduced by Cheng-Yau in [4].

$$\square f=\sum_{i,j}(\delta_{ij}\sum_k S_{kk}-S_{ij})f_{ij} \quad \text{for any } f\in C^2(M)$$

When M is compact, we get following in [1]

$$\int_M\|\nabla S\|^2-\|\nabla trS\|^2+\frac{1}{2}\sum_{i,k}R_{ikik}(\lambda_i-\lambda_k)^2=0 \tag{1.5}$$

2. SOME RESULTS ON CONFORMALLY FLAT MANIFOLD

2.1 Lemma and Theorem

Following algebraic lemma is well-known (cf[3]) Lemma. Let $\alpha_i,i=1,2,\cdots,n$ be real number such that $\sum_i\alpha_i=0$, and $\sum_i\alpha_i^2=const.\ge0$, then

$$\left|\sum_i\alpha_i^3\right|\le\frac{n-2}{\sqrt{n(n-1)}}(\sum_i\alpha_i^2)^{\frac{3}{2}} \tag{2.1}$$

and equality holds in (2.1) if and only if n-1 of α_i are equal, or all $\alpha_i=0$.

Proof: We can obtain Lemma by using the method of Lagrange's multipliers to find the critical points of $\sum_i\alpha_i^3$ subject to the conditions $\sum_i\alpha_i=0$ and $\sum_i\alpha_i^2=const.\ge0$. We omit it here.

Let λ_i be the eigenvalues of S, $\alpha_i=\frac{1}{n}trS-\lambda_i$, then

$$\sum_i\alpha_i^2=\|S\|^2-\frac{1}{n}trS^2 \tag{2.2}$$

$$\sum_i\alpha_i^3=-\frac{2}{n}(trS)^3+\frac{3}{n}(trS)\|S\|^2-\frac{1}{n}trS^3 \tag{2.3}$$

we find that

1) $\sum_i\alpha_i=0$

$$2)\ \alpha_i=0\Leftrightarrow\sum_i\alpha_i^2=0\Leftrightarrow\|S\|^2-\frac{1}{n}(trS)^2=0 \tag{2.4}$$

From (2.4) we have

$$ntr(S^3)-(trS)\|S\|^2=-\frac{2}{n}(trS)^3+2(trS)\|S\|^2-n\sum_i\alpha_i^3 \tag{2.5}$$

From Lemma and (2.2), we get

$$\sum_i\alpha_i^3\le\frac{n-2}{\sqrt{n(n-1)}}(\sum_i\alpha_i^2)^{\frac{3}{2}}=\frac{n-2}{\sqrt{n(n-1)}}(\|S\|^2-\frac{1}{n}(trS)^2)^{\frac{3}{2}} \tag{2.6}$$

Taking (2.6) into (2.5), we obtain

$$ntr(S^3)-(trS)\|S\|^2\ge(\|S\|^2-\frac{1}{n}(trS)^2)\times\\(2trS-\frac{n(n-2)}{\sqrt{n(n-1)}}\sqrt{\|S\|^2-\frac{1}{n}(trS)^2}) \tag{2.7}$$

From Lemma and (2.4), we know that in (2.7) the equality holds if and only if $\|S\|^2-\frac{1}{n}(trS)^2=0$ or n-1 of λ_i are equal.

Since

$$\frac{1}{2}\sum_{i,k}R_{ikik}(\lambda_i-\lambda_k)^2=\frac{1}{n-2}(ntr(S^3)-(trS)\|S\|^2) \tag{2.8}$$

Taking (2.7) and (2.8) into (1.5), we obtain

$$\int_M\|\nabla S\|^2-\|\nabla trS\|^2+\frac{1}{n-2}(\|S\|^2-\frac{1}{n}(trS)^2)\times\\(2trS-\frac{n(n-2)}{\sqrt{n(n-1)}})\sqrt{\|S\|^2-\frac{1}{n}(trS)^2})\le0 \tag{2.9}$$

From (2.9) and Lemma, we get the following Theorem.

* This paper is supported by Scientific Fund Project of Hebei Polytechnic University (z0821), and Youth Foundation of Southwest university(SWUQ2006030)

Theorem: Let M (dimM>3) be a compact conformally flat manifold. If S satisfies

1) $\|\nabla S\|^2 \geq \|\nabla trS\|^2$

2) $trS \geq \frac{n(n-2)}{2\sqrt{n(n-1)}}\sqrt{\|S\|^2 - \frac{1}{n}(trS)^2}$

Then, either

$$\|S\|^2 = \frac{1}{n}(trS)^2$$

and M is constant sectional curvature;
or

$$trS = \frac{n(n-2)}{2\sqrt{n(n-1)}}\sqrt{\|S\|^2 - \frac{1}{n}(trS)^2} \quad (2.10)$$

and S has two different eigenvalues λ_1, λ_n , where

$$\lambda_1 = \cdots = \lambda_k = \frac{1}{n}(1 + \frac{2n}{n-2}\sqrt{\frac{(n-1)(n-k)}{k}}trS) \quad (2.11)$$

$$\lambda_{k+1} = \cdots = \lambda_n = \frac{1}{n}(1 - \frac{2n}{n-2}\sqrt{\frac{(n-1)k}{n-k}}trS) \quad (2.12)$$

2.2 Proof of the Theorem

Proof: From (2.2) we know $\|S\|^2 = \frac{1}{n}(trS)^2 \geq 0$.

Then, from condition (1), (2) we can conclude two cases,

$$\|S\|^2 = \frac{1}{n}(trS)^2, \|\nabla S\|^2 = \|\nabla trS\|^2$$

From (2.4) we know that each eigenvalue of S satisfies $\lambda_i = \frac{trS}{n} = const.$ thus, M is Eintein. Since conformally flat Eintein space is constant sectional curvature space, M is constant sectional curvature.

In the second case,

$$\|\nabla S\|^2 = \|\nabla trS\|^2, \quad trS = \frac{n(n-2)}{2\sqrt{n(n-1)}}\sqrt{\|S\|^2 - \frac{1}{n}(trS)^2}$$

Then S has two different eigenvalues λ_1, λ_n , we give the calculations of λ_1, λ_n as following

Let $\lambda_1 = \cdots = \lambda_k = \lambda$, $\lambda_{k+1} = \cdots = \lambda_n = \mu$

We have known

$$trS = k\lambda + (n-k)\mu$$

$$\|S\|^2 = \sum_{i,j} S_{ij}^2 = \sum_{i,j} \lambda_i \delta_{ij}^2 = k\lambda^2 + (n-k)\mu^2$$

Then, directly calculation, we get

$$\|S\|^2 - \frac{1}{n}(trS)^2 = (\frac{n-k}{n})k(\lambda + \mu)^2$$

Taking it into (2.10), then we can obtain (2.11) and (2.12).

Corollary A: Let M (dimM>3) be a compact conformally flat manifold with constant scalar curvature r. If

$$trS \geq \frac{n(n-2)}{2\sqrt{n(n-1)}}\sqrt{\|S\|^2 - \frac{1}{n}(trS)^2} \quad (2.13)$$

then M is constant sectional curvature space.

Corollary B: Let M (dimM>3) be a compact conformally flat manifold. If

1) $\sum_{i,j,k,l} R_{ijkl}^2 - \frac{r^2}{n-1} = c(= const.)$

2) $\|Q\| < \frac{r}{\sqrt{n-1}}$

then M is constant sectional curvature space.

REFERENCES

[1] Nan Ji, Shaohong Yan and Yamian Peng , A Codazzi Tensor on Conformally Symmetric Space and its Applications, Journal of Hebei Polytechnic University,march ,2007，pp.110-113

[2] Hertrich-Jeromin U. , Models in Meobius Differential Geometr y，2001, pp.16-22

[3] Okumura,M.,Hypersurfaces and a pinching problem on the second fundamental tensor,Amer.J.Math, 1974.pp.207-213.

[4] Cheng S. Y.,Yau S. T. Hypersurfaces with constant scalar curvature , Math. Ann. ,1977, pp.195-204.

[5] Chern S.S.,Do Carmo M.,and Kobayashi S., Minimal submanifold of a sphere with second fundamentalform of constant length ,in Fuctional Analysis and Related Fields(F.E.Browder,ed.)Springer-Verlag,New York,1970 ,pp.59-75

[6] Li,H., Hypersurface with constant scalar curvature in space forms, Math.Ann., 1996, pp.665-672.

Ji Nan: lecturer, now acting as the Math teacher in Hebei Polytechnic University. She graduated from Hebei Normal University, majoring in mathematics. Now she is the member of Parallel Computing laboratory and scientific computing laboratory. She has already attended presided over 10 research projects. Her over 10 papers have been published. Her research directions include the design and analysis of parallel computation, differential geometry, and etc.

AN IMPLEMENTATION OF UNLIMITED NUMERIC CLASS *

Kong Weiguang[1], Wang Yuan[2]
[1]**College of Computer Science, Wuhan University of Science and Engineering, Wuhan, 430074, China**
[2]**Library, Zhongnan University of Economics and Law, Wuhan, 430074, China**
Email: wgkong@public.wh.hb.cn

ABSTRACT

By analyzing the expression of digital symbol and calculation of number system, this paper develops a number system based on 2^N system, whose digital symbol can be easily expressed with built-in data type variable of high-level programming language. It researches the characteristic of scalable queue templates in visual C++ template library and gives the implementation scheme of unlimited number class founded on automatic scalable queue, In the scheme, one unlimited number object is made up by two queues: one for integer part and one for decimal part of unlimited number, each element of the queue is one numeral of the unlimited number object. This paper also researches the calculation of arithmetic calculation and tests its performance in 1024+ bits RSA cryptogram, and the result indicates that this number class is practicable, and it is significant to be a member of base library of high level programming language.

Keywords: Number System, Unlimited Numeric Class, Digital Symbol, Queue, Cryptogram.

1. INTRODUCTION

Big numeric data is useful in high precision applications such as compute algebra, geometric reasoning, cryptogram and microcosmic simulation, which calculate between big numbers from hundreds digits to thousands digits. Large primes and integers are often used in digital data encryption, RSA algorithm is an example that can be used to encrypt data and authenticate identification, its security is founded on the factorization difficulty of large integers, and it has resisted kinds of attacks successfully and kept valuable for years. Larger numbers are demanded as the developing of computing technology.

All high-level programming languages define their built-in data types, such as int, long, _int64, char and string in C++. Simple operations can be effectively performed by variables of built-in data type. It is a pity that all built-in data type variables are limited, for instance, in C++, the maximum integer type is _int64, numbers bigger than it must be turned to imprecise float. It is significant to develop a special data type to accurately express numbers larger than the maximum integer or with much more decimal digits.

One practicable solution is treating large number as a string made up of decimal digit symbols, and all arithmetic calculations such as add, sub, mul, and div are turned to operations of chars in the strings, simulating manual work pattern. For example, a 1024-bit integer may be turned into a decimal string with hundreds of chars. The strongpoint of this solution is the easiness to be understood, it avoids the puzzle of value overflow, and it has two shortcomings: the first is that simulant calculation between chars has poorer efficiency than built-in data types; the second is that it will take long time to loop on long strings of decimal digit symbols.

The other solution is treating big number as a stream of binary bits, all arithmetic calculations are made by shift and logic operations of binary bits. This solution has higher space efficiency, but operations on binary bits will take longer time, and the program code will be complex and unreadable. In real projects, a specified array is generally preferred to express big number and achieve calculations of big number.

It is troublesome to operate between classes with arrays of different length in projects based on limited array, and it is necessary to compile another similar class based on a longer array when digits of big number need to be expanded.

Reference [1] has researched the relation between operation speed and number system basement, and concluded that the operate speed is faster when the number system basement is bigger. Reference [2] has designed a big number of 2^{32} based number system, turned a 1024 bit number to an array with 32 elements of 2^{32} number system, and calculate imitating handwork pattern. Reference [3] has advanced a scheme of big number founded on chars, constructed a calculation system of and analyzed its performance. References [4-12] have researched kinds of calculations of big integers separately.

The solution of big number in this paper is based on scalable queue, the number is of 2^n based number system, and it is unlimited and can be calculated efficiently.

2. NUMBER SYSTEM AND DIGITAL SYMBOL

Common used numbers without special statement are decimal in the real world. In decimal number system, there are ten symbols (0,1,…9), the radix is 10. The numerical size of a number is the sum of every digital multiplies its right of the position, and different positions have different rights. for example, the value of number 987.65 can be computed through the Eq. (1).

$$987.65=9\cdot10^2+8\cdot10^1+7\cdot10^0+6\cdot10^{-1}+5\cdot10^{-2} \qquad (1)$$

Other number systems such as hexadecimal, binary and octal are familiar. The value of one number in different number

* This work is partially supported by Hubei Education Foundation Grant #200717005 to Mr. Hu Ming.

systems is unaltered. It is well known that data in the computer is binary finally, and the radix of binary number system is 2.
By the same token, M based number system can be defined as following:

1) There are M symbols from 0 to M-1 in the system;
2) The base radix is M.

For example, one number in M based number system can be computed as Eq. (2).

$$B_2B_1B_0.B_{-1}=B_2\cdot M^2+B_1\cdot M^1+B_0\cdot M^0+B_{-1}\cdot M^{-1} \qquad (2)$$

Inside the computer, it is necessary to figure each digital symbol of M based number system in order to implement its operation. If M equals 2^n, it is suitable to figure M digital symbols with different combination of n binary bits. For instance, in octal number system, symbols 0~7 can be represent with combination of three binary bits: 000, 001, 010, 011, 100, 101, 110, and 111. In other words, three binary bits can be treated as a digital symbol of octal number system. If n is 8, it is appropriate to use a byte to figure the digital symbols of 2^8 based number system, and if n is 16, it is seemly to use a word to figure the digital symbols of 2^{16} based number system, and if n is 32, it is suitable to use four bytes to figure the digital symbols of 2^{32} based number system, and so on.

The arithmetic and logic operations of numbers in all number systems are actually operated on all of its digital symbols. For the convenience to save and to calculate the digital symbols of M based number system, it is the best to figure each digital symbol with a built-in data type variable of high level programming language (such as byte, int, long, _int64 etc.) because the basic operations of these types have been programmed efficiently. Considering that the length of operation result of two digital symbols may be 1 or 2 digitals, the proper principle to select n is that the longest built in type of current high level language can hold two digital of 2^n based number system, namely, n is limited to half of the length of the maximum built-in integer type.

Numeric class of 2^n based number system can be organized with a queue of 2^n based digital symbols. At least two queues are required to combine a class of 2^n based number system, one for the integer part and the other for the decimal part. The queues must be scalable for unlimited number class. The two queues and a Boolean variable represent the sign of number form the basic data structure of unlimited numeric class.

3. SCALABLE DEQEUE IN C++

C++ template class library is a useful part of stand C++ language; it has defined perfect data containers and many efficient algorithms. The templates can construct classes with many more functions because c++ standard template library is expandable and the containers are scalable,

There are three kinds of scalable queues in c++ template class library, they are deque, vector and list, all of them support random access of all elements, and each of them has its own specialty. Vector is good at efficient insertion and deletion operations at the end of queue and the cost time at the head or in the middle of queue is linear. Deque's time of operating on random position of queue is efficient and invariable.

By comparison，deque has the advantage of array and list, it can be operated as convenient as array and can flex to meet special need. Deque is suitable to be the container of digital queue of unlimited numeric class. It will be convenient to insert or delete elements at the head or end of digital queues, and it will be easy to access any element randomly. More importantly, the count of elements in deque is unlimited. In visual C++, commonly used operation functions of deque are listed as following:

begin()	Return iterator point to the head of queue;
end()	Return iterator point to the end of queue;
push_front()	Insert an element to the head of queue;
push_back()	Insert an element to the end of queue;
pop_front()	Delete the element at the head of queue;
pop_back()	Delete the element at the end of queue;
insert()	Insert elements to a random position;
erase()	Delete element at specified position;
at()	Return a reference of an element;
rbegin()	Return iterator point to the converse head of queue;
rend()	Return iterator point to the converse end of queue;

In addition, it is easy to access elements with iterator variables. Generally, operation of deque is convenient.

4. WORKING PRINCIPLE OF UNLIMITED NUMBER CLASS

In visual C++， the longest built-in integer type is unsigned _int64， so it is right to define n as 32. In 2^{32} based number system, the maximum digital symbol is 0xffffffff in hexadecimal which can be seated in an unsigned long variable. The sum or product of two digital should be settled in an unsigned _int64 variable because it may overstep the limit of unsigned long. In 2^n based number system, a number with K binary bits is [K/n] digitals long (figure 1.), and the calculation of unlimited number has been turned to operations of deque elements and calculations of unsigned long variables.

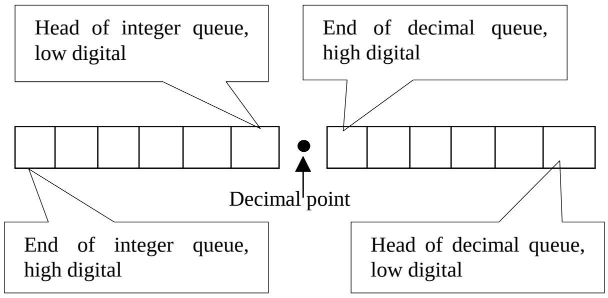


Figure 1. Digital Queues of Unlimited Numeric Class

In this unlimited numeric class, each element of integer queue and decimal queue is a digital symbol of 2^n based number system; the following is the definition of data structure:

```
deque <unsigned long> UN_Integer_Deque;
    //integer deque, the head is low position digital
deque<unsigned long> UN_Decimal_Deque;
    //decimal queue, low decimal position on head
bool        NegativeFlag;
    //number sign, true represents negative number
```

It is necessary to implement the arithmetic algorithms of this class. Addition of two numbers can be switched to addition or subtraction of two numbers with the same NegativeFlag, multiplication and division can be transformed into operations of NegativeFlag and operations of deques. Underside, the addition, subtraction, multiplication and division of unlimited number class are discussed.

Like manual work, after aligning decimal point position, the addition of two numbers is performed by a loop from the head of decimal queue to the end of decimal queue and the head of integer queue to the end of integer queue. In case of adding unlimited number X and Y, assuming that unlimited number Z is used to save the result, integer i represents the relative position of a digital, x_i, y_i and z_i are their corresponding digital in position i, unsigned long variable C(0 or 1) is the carry from digital position i-1, unsigned _int64 S is the depositary of the sum of $a_i + b_i$+C, then the lower half part of S is z_i and the high half part of S is the next carry C. The carry from the end of decimal queue will attend the addition of digitals on the head of integer queue.

The subtraction of two unlimited numbers is finally founded on the operation that the bigger number in absolute subtracts the smaller one. The loop is performed like addition. For instance, X-Y is needed to be calculated and the result will be left in Z, C (0 or 1) is the borrow bit from the lower digital and unsigned _int64 S is used temporarily, the following is the pseudo code of subtraction of one digital z_i:

S= y_i+C
If S<= x_i
Then $z_i = x_i$ -S, C=0
Else $z_i = x_i + 2^{32}$-S, C=1

Multiplication can be performed by loops of digital multiplication, shift and addition of middle results. It is necessary to save the middle result of one digital of multiplicator multiplies the multiplicand unlimited number, and the middle result will be shifted properly and added to the final result. The shift left operation is actually moving one element from end of decimal queue to the head of integer queue, if the decimal queue is empty, an element 0 will be inserted to the head of integer queue. The shift right operation is moving the head of integer queue to the end of decimal queue; the integer queue can not be empty because it should remain an element 0 even if the number does not have integer part.

Division is relatively complex because it must try the quotient each time. The quotient test procedure must recur to multiplication and subtraction of unlimited numbers. For example, there are unlimited numbers A, B, C and D, A is the dividend, B (B!=0 and A>B) is the divisor, C will be the quotient and D will be the remainder, the division arithmetic can be described as following steps:

1) Shift A and B in the same way in order to turn B into an integer unlimited number;
2) Copy the highest part of A with the length of B to D, D does not have decimal part, and then insert an element 0 to the end of integer queue of D;
3) If D<B, then insert an element 0 to the lower position of C, it means this quotient digital is 0;
   If D=B, then insert an element 1 to the lower position of C, it means this quotient digital is 1, and do D=D-B;
   If D>B, then try to find the quotient along following steps:
   a) Copying the highest two digital of D to d2 (unsigned _int64), copy the highest one digital of B to b(unsigned _int64), assuming q(unsigned _int64) is the try quotient;
   b) Minimum of q is d2/(b+1), maximum of q is min(d2/b, 2^{32}-1) , by dichotomy, it is possible to find one q settle for D<=B*(q+1) and D>=B*q. At this time, the right quotient must be q or q+1, if D<B*(q+1) then the quotient is q, else the quotient is q+1. The quotient would be inserted to the lower position of C;
   c) Subtract the product of B and the quotient from D. Temporality the highest digital of D must be 0, because D<B, and the length of D is 1 digital more than B;
4) Insert the next digital from A to the lower position of D, and delete the highest element of D (it is 0), and repeat step 3 until all elements of A has been treated. If integer part of A finished, the next digital should be the tail of decimal deque, and the quotient q will join the decimal part of quotient unlimited number C;
5) Shift D in the contrary way of step 1, then D is the remainder.

All arithmetic calculations of unlimited numbers can be founded on the algorithms upon.

5. PERFORMANCE TEST

Because C++ is an object oriented programming language, it is easy to define unlimited numeric class as a new class and implement the conversion functions with other built-in data types such as int, long, _int64. It is necessary to realize in-out operations of kinds of formats. It will be used as easy as built-in data types.

The unlimited numeric class is obviously better than other built-in types in big number expression capability, because the length of the integer part and decimal part are all unlimited and automatically scalable.

Compared with the big integer solution based on limited array in predefined length, the unlimited number class has an obvious superiority that its length can change automatically to meet the actual need, and it is efficient in memory using. On the other hand, the operation speed of unlimited number class is slower than limited array, because the element operation speed of deque is lower than array. Table I shows the result of one experiment, it lists some cost of time of unlimited numeric class to perform 1024 bit integer calculation on the same computer(P4/2.8GHZ/1G).

Table 1 Calculation Time of Unlimited Numeric Class

	1024bits	4096bits
Additions(2000 times)	300ms	1s
X^Y mod Z	15~25s	150~900s
Generate RSA keys	Average 8m	Average 50m

(ms:millisecond; s:second; m:minute)

The cost of calculation time on 1024 bit numbers of unlimited number class is acceptable. This class is valuable to be a member of base library of high level programming languages, and it can be one basement part of cryptogram class library. The speed of unlimited number class will increase availably while deque's operation speed improved.

REFERENCES

[1] JF. Liu, "Radix's selection of big integral's calculating accurately", Journal of East China Jiao tong University, vol.24 no.2, April 2007.

[2] YD. Liu, "Implementation for operation of big integer based on array", Computer and Modernization, March 2008.

[3] J. Yang, KM. Tang, XQ. Yang, "Construction and implement of characterizing large integer numerical calculation system", Journal of Yangzhou University (Natural Science Edition), vol.3 no.4, November 2000.

[4] P. Zhao, DD. Zhang, ZF. Tian, "Large integer multiplication in SIMD-LA", Microcomputer information, vol.22, December 2006.

[5] F. Gao, YZ. Wang, LQ. Sang, R. Shi, "The implementation of multiplication and division of big integer on PC machines", Journal of Logistical Engineering University, January 2007.

[6] CS. Ying, XL. Zhou, "Data structure and algorithm selection of big integer multiplication", Journal of Changchun University of Technology (Natural Science Edition), vol.29 no.2, April 2008.

[7] Y. Wang, L. Zhou, YK. Wang, B. Zhou, "The software realization for the modular arithmetic of large integer", China Information Security, February 2005.

[8] WP. Xie, DZ. Chen, "The machine testing and software realizing of primality of large number", Journal of Nanhua University (Science and Technology), vol.20 no.2, June 2006.

[9] Y. Shi, S. Yao, "Realization of arithmetic operation of the large integer", Journal of Anqing Teachers College (Natural Science), vol.10 no.2, May 2004.

[10] GL. Zhu, FM. Li, "Exact solution of multiplication of large integers", Journal of North China Institute of Water Conservancy and Hydroelectric Power, vol.28 no.3, June 2007.

[11] ZH. Wang, "A novel calculation method for the factorization of large integral numbers and the decoding of the RSA Cryptosystem", Acta Scientiarum Naturalium Universitatis Sunyatseni, vol.42 no.5, May 2003.

[12] KB. Dai, JH. Chen, "Optimization of the quadratic sieve in large integer factorization", Journal of Math (PRC), vol.25 no.6, June 2005.

[13] C. Ling, L. Mai, "Operation and performance comparisons of big integers based on two different memories", Journal of Anqing Teachers College (Natural Science), vol.9 no.1, February 2003.

[14] Bruce Schneier, Applied cryptography, John Wiley & Sons, 1996.

[15] Wang YB, Xue T, Applied cryptography, Beijing:China Machine Press, 2003.

[16] Lv FX, C++ programming, Beijing:China Publishing House of Electronics Industry, February 2002.

[17] Richard Johnsonbaugh, Object oriented programming in C++, Beijing:China Machine Press, March 2005.

THE COMPUTATION OF THE PARAMETERS β_{ss} AND β'_{ss} *

Tian Yi-Min , Zhang Ao
Beijing Institute of Graphic Communication
Beijing ,China
Email: tym@bigc.edu.cn,ahf200805@sina.com

ABSTRACT

With the increasing of the research about the multi-quark system, we perform a schematic study of the parameters β_{ss} and β'_{ss} of the tetraquarks. By the computation, we get the teo important parameters β_{ss} and β'_{ss}, which will be used to compute the tetraquarks masses.

Keywords: Tetraquark, Parameter

1. INTRODUCTION

The multi-quark system was studied many years ago using a quark model with color-magnetic interaction. Recently , Richard and Stancu used the model to estimate masses of the tetraquarks. [1].Surprisingly, we find that it is possible to accommodate the tetraquarks relatively comfortably within the restrictions imposed by experimental meson spectroscopy. We do not claim to resolve the problem by elevating multi-quark states to be very high masses [2], [3]. Several experiment are now under way and it appears useful to make precise predictions about the masses, the decay modes and the widths of the meson-baryon states expected in the color isomer model [4]. In this paper we compute the two important parameters: β_{ss} and β'_{ss}, which will help to research the tetraquark system.

2. THE COMPUTATION OF THE PARAMETER β_{ss}

Isoscalar states with the same J^{pc} will mix but mixing between the two light quark isoscalar mesons and the much heavier charmonium or bottomonium states are generally assumed to be negligible. The physical isoscalar are mixing of the SU(3) wave functon ψ_8 and ψ_1:

$$\begin{cases}\eta = \psi_8 \cos\theta_p - \psi_1 \sin\theta_p \\ \eta' = \psi_8 \sin\theta_p + \psi_1 \cos\theta_p\end{cases} \quad (1)$$

Where θ_p is the nonet mixing angle.

$$\begin{cases}\psi_8 = \frac{1}{\sqrt{6}}(u\bar{u} + d\bar{d} - 2s\bar{s}) \\ \psi_1 = \frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s})\end{cases} \quad (2)$$

$$\eta = \frac{1}{\sqrt{6}}(u\bar{u} + d\bar{d} - 2s\bar{s})\cos\theta_p - \frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s})\sin\theta_p \quad (3)$$

$$\eta' = \frac{1}{\sqrt{6}}(u\bar{u} + d\bar{d} - 2s\bar{s})\sin\theta_p + \frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s})\cos\theta_p \quad (4)$$

$$\eta = (u\bar{u} + d\bar{d})(\frac{1}{\sqrt{6}}\cos\theta_p - \frac{1}{\sqrt{3}}\sin\theta_p) - s\bar{s}(\frac{2}{\sqrt{6}}\cos\theta_p + \frac{1}{\sqrt{3}}\sin\theta_p) \quad (5)$$

$$\eta' = (u\bar{u} + d\bar{d})(\frac{1}{\sqrt{6}}\sin\theta_p + \frac{1}{\sqrt{3}}\cos\theta_p) + s\bar{s}(\frac{1}{\sqrt{3}}\cos\theta_p - \frac{2}{\sqrt{6}}\sin\theta_p) \quad (6)$$

$$\begin{aligned}&(\frac{1}{\sqrt{6}}\cos\theta_p - \frac{1}{\sqrt{3}}\sin\theta_p)\eta' - (\frac{1}{\sqrt{6}}\sin\theta_p + \frac{1}{\sqrt{3}}\cos\theta_p)\eta \\ &= [(\frac{1}{\sqrt{3}}\cos\theta_p - \frac{2}{\sqrt{6}}\sin\theta_p) \times (\frac{1}{\sqrt{6}}\cos\theta_p - \frac{1}{\sqrt{3}}\sin\theta_p) \\ &\quad + (\frac{2}{\sqrt{6}}\cos\theta_p + \frac{1}{\sqrt{3}}\sin\theta_p) \times (\frac{1}{\sqrt{6}}\sin\theta_p + \frac{1}{\sqrt{3}}\cos\theta_p)]s\bar{s}\end{aligned} \quad (7)$$

$$\begin{aligned}&\frac{1}{\sqrt{6}}(\cos\theta_p - \sqrt{2}\sin\theta_p)\eta' - \frac{1}{\sqrt{6}}(\sin\theta_p + \sqrt{2}\cos\theta_p)\eta \\ &= [(\frac{\cos^2\theta_p}{3\sqrt{2}} - \frac{2}{3}\cos\theta_p\sin\theta_p + \frac{2\sin^2\theta_p}{3\sqrt{2}})\end{aligned} \quad (8)$$

$$s\bar{s} = \frac{1}{\sqrt{3}}(\cos\theta_p - \sqrt{2}\sin\theta_p)\eta' - \frac{1}{\sqrt{3}}(\sin\theta_p + \sqrt{2}\cos\theta_p)\eta \quad (9)$$

The pseudoscalar mixing angle $\theta_p = -17.3^0$ can be measured by comparing the partial widths for radiative J/ψ decay into a vector and a pseudoscalar. Thus we can get:

$$s\bar{s} = 0.794\eta' - 0.608\eta \quad (10)$$

The masses of the meson η and η' are 547.8 Mev and 957.8 Mev, respectively. The mass of the $s\bar{s}$ is

$$\begin{aligned}m_{s\bar{s}} &= 0.794^2 \times 957.8 + 0.608^2 \times 547.8 \\ &= 806.3Mev\end{aligned} \quad (11)$$

*This project was supported by financially supported by NSFC (F60850007) ,BJSFC(A1092012) and BIGC (Ea-09-14).

$$m_{s\bar{s}} = 2m_s - \frac{8\alpha_s\beta_{ss}}{3\sqrt{\pi}} + \frac{2b}{\sqrt{\pi}\beta_{ss}} - \frac{8\alpha_s}{3m_s^2\sqrt{\pi}}\left(\frac{\sigma^2\beta_{ss}^2}{\sigma^2+\beta_{ss}^2}\right)^{\frac{3}{2}} \quad (12)$$

With the parameter $b = 0.162Gev^2$, $\alpha_s = 0.594, m_s = 0.55Gev, \sigma = 0.897Gev$ we can get the parameter value:

$$\beta_{ss} = 0.4878Gev \quad (13)$$

3. THE COMPUTATION OF THE PARAMETER β'_{ss}

The physical vector mesons are mixtures of the SU(3)wave function ψ_8 and ψ_1:

$$\begin{cases}\phi(1020) = \psi_8\cos\theta_V - \psi_1\sin\theta_V \\ \omega(782) = \psi_8\sin\theta_V + \psi_1\cos\theta_V\end{cases} \quad (14)$$

By the same way, we can get:

$$\phi(1020) = \frac{1}{\sqrt{6}}(u\bar{u} + d\bar{d} - 2s\bar{s})\cos\theta_V - \frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s})\sin\theta_V \quad (15)$$

$$\omega(782) = \frac{1}{\sqrt{6}}(u\bar{u} + d\bar{d} - 2s\bar{s})\sin\theta_V + \frac{1}{\sqrt{3}}(u\bar{u} + d\bar{d} + s\bar{s})\cos\theta_V \quad (16)$$

$$\phi(1020) = (u\bar{u} + d\bar{d})(\frac{1}{\sqrt{6}}\cos\theta_V - \frac{1}{\sqrt{3}}\sin\theta_V) - s\bar{s}(\frac{2}{\sqrt{6}}\cos\theta_V + \frac{1}{\sqrt{3}}\sin\theta_V) \quad (17)$$

$$\omega(782) = (u\bar{u} + d\bar{d})(\frac{1}{\sqrt{6}}\sin\theta_V + \frac{1}{\sqrt{3}}\cos\theta_V) + s\bar{s}(\frac{1}{\sqrt{3}}\cos\theta_V - \frac{2}{\sqrt{6}}\sin\theta_V) \quad (18)$$

$$\begin{aligned}&(\frac{1}{\sqrt{6}}\cos\theta_V - \frac{1}{\sqrt{3}}\sin\theta_V)\omega(782) - (\frac{1}{\sqrt{6}}\sin\theta_V + \frac{1}{\sqrt{3}}\cos\theta_V)\phi(1020) \\ &= [(\frac{1}{\sqrt{3}}\cos\theta_V - \frac{2}{\sqrt{6}}\sin\theta_V) \times (\frac{1}{\sqrt{6}}\cos\theta_V - \frac{1}{\sqrt{3}}\sin\theta_V) \\ &\quad + (\frac{2}{\sqrt{6}}\cos\theta_V + \frac{1}{\sqrt{3}}\sin\theta_V) \times (\frac{1}{\sqrt{6}}\sin\theta_V + \frac{1}{\sqrt{3}}\cos\theta_V)]s\bar{s} \quad (19)\end{aligned}$$

$$\begin{aligned}&\frac{1}{\sqrt{6}}(\cos\theta_V - \sqrt{2}\sin\theta_V)\omega(782) - \frac{1}{\sqrt{6}}(\sin\theta_V + \sqrt{2}\cos\theta_V)\phi(1020) \\ &= [(\frac{\cos^2\theta_V}{3\sqrt{2}} - \frac{2}{3}\cos\theta_V\sin\theta_V + \frac{2\sin^2\theta_V}{3\sqrt{2}}) \\ &\quad + (\frac{2\cos^2\theta_V}{3\sqrt{2}} + \frac{2}{3}\cos\theta_V\sin\theta_V + \frac{\sin^2\theta_V}{3\sqrt{2}})]s\bar{s} \quad (20)\end{aligned}$$

$$s\bar{s} = \frac{1}{\sqrt{3}}(\cos\theta_V - \sqrt{2}\sin\theta_V)\omega(782) - \frac{1}{\sqrt{3}}(\sin\theta_V + \sqrt{2}\cos\theta_V)\phi(1020) \quad (21)$$

With the vector mixing angle $\theta_V = 35^0$, very close to ideal mixing. Thus $\phi(1020)$ is nearly pure $s\bar{s}$. Thus we can get:

$$s\bar{s} = 0.0046\omega(782) - 0.9999\phi(1020) \quad (22)$$

The masses of the vector meson $\omega(782)$ and $\phi(1020)$ are 782.6 Mev and 1019.5 Mev, respectively. The mass of the $s\bar{s}$ is

$$m'_{s\bar{s}} = 0.0046^2 \times 782.6 + 0.9999^2 \times 1019.5 = 1019.3Mev \quad (23)$$

At the same time, we have the formula:

$$m'_{s\bar{s}} = 2m_s - \frac{8\alpha_s\beta'_{ss}}{3\sqrt{\pi}} + \frac{2b}{\sqrt{\pi}\beta'_{ss}} - \frac{8\alpha_s}{3m'^2_s\sqrt{\pi}}\left(\frac{\sigma^2\beta'^2_{ss}}{\sigma^2+\beta'^2_{ss}}\right)^{\frac{3}{2}} \quad (24)$$

With the parameter $b = 0.162Gev^2$, $\alpha_s = 0.594, m_s = 0.55Gev, \sigma = 0.897Gev$ we can get the parameter value:

$$\beta'_{ss} = 0.5716Gev$$

4. DISCUSSIONS

In short summary, we have calculated the two important parameters β_{ss} and β'_{ss}. Then we can used the two parameters to compute the tetraquark masses. If future experimental searches success to find the tetraquarks, it is sure that the two parameters will become more important to research the tetraquarks.

REFERENCES

[1] R.Jaffe and F.Wilczek, Phys . Rev . Lett . 91 , 232003 (2003) .
[2] Ying Cui, Xiao-Lin Chen, Wei-Zhen Deng, Shi-LinZhu, Phys.Rev.D73: 014018 , 2006
[3] Ying Cui, Xiao-Lin Chen, Wei-Zhen Deng, Shi-Lin Zhu, High Energy Phys. Nucl .Phys. 31:7-13,2007.
[4] John Weinstein, Nathan Isgur, Phys . Rev.D27:588, 1983

A PROTOTYPE FOR DETECTING DEFECTIVE PILLS DURING MANUFACTURING *

Craig C. Douglas [1],Gundolf Haase [2],Thad Hannel [3] , David Link [3], and Robert A. Lodder [3], Deng Li [4]
[1] Mathematics Department, University of Wyoming,Laramie, WY 82071-3036, USA and
Computer Science Department, Yale University,New Haven, CT 06520-8285, USA
[2] Mathematics and Computational Sciences Department, Karl-Franzens University of Graz
A-8010 Graz, Austria
[3] Chemistry Department, University of Kentucky,Lexington, KY 40506, USA
[4] JSOL Corporation, Chuo-ku, Tokyo 104-0053, Japan
Email: [1]craig.c.douglas@gmail.com,[2] gundolf.haase@uni-graz.at,
[3] tshannel@gmail.com, [3] dlink4884@gmail.com, and [3] lodder@uky.edu,[4] deng17jp@yahoo.co.jp

ABSTRACT

Contaminants and impurity formation in pharmaceutical products during manufacturing cause inconvenience and costly recalls. Determining when to adjust a process to control impurity formation, or halt production and destroy end products should be an automatic task using experts' knowledge, data (audio and chemical spectral footprints), adaptive models and algorithms, and different scales (data and models). Integrated sensing and processing (ISP) optimizes sensing systems that integrate the traditionally independent units of sensing, signal processing, communication and targeting. By employing ISP, computational complexity within traditional sensing system is substantially reduced through determining efficient low-dimensional representations of those sensing problems that were originally posed in high-dimensional settings by traditional sensing architecture. An ISP-based imaging spectrometer produces detector signals directly correlated to desired sample information, obviating the need for post-collection chemometrics and converting data directly to knowledge. We apply ISP techniques to novel acoustic spectral sensors designed and created as part of this project. The ISP is reprogrammable on demand. It uses data libraries that are created offline. The creation is expensive computationally, but it makes the use of the sensor easy. We have to use high performance computing facilities to quickly generate new or modified libraries.

Keywords: Defect detection, manufacturing, integrated sensing and processing, dynamic data-driven application systems, DDDAS, cyber physical systems, CPS, parallel processing, and algorithms.

1. INTRODUCTION

Dynamic data-driven application systems (DDDAS) creates a rich set of new challenges for applications, algorithms, systems software, and measurement methods. DDDAS research [1-10] typically requires strong, systematic collaborations between applications domain researchers and mathematics, statistics, and computer sciences researchers, as well as researchers involved in the design and implementation of measurement methods and instruments. Consequently, most DDDAS projects involve multidisciplinary teams of researchers.

DDDAS enabled applications run in a different manner than many traditional applications. They place different strains on high performance systems and centers due to dynamic and unpredictable changes in resources that are required during long term runs. An on demand environment is required. In this paper, we will also categorize many of these differences.

Cyber physical systems (CPS) are typically small, embedded systems using data assimilation, like DDDAS, and have a distinct personal interaction or commitment. The devices we are designing and putting in the field are a cross between traditional DDDAS and CPS devices and applications.

Sensors are collecting more and more data with greater bandwidth. The ratio of pixels-to-pupils is heading toward infinity, meaning that no one person can make sense of all of the data. Sensors need to produce more high level information (like chemical concentrations) and less fields of physical information (e.g., reflectance values at hundreds or thousands of wavelengths) to implement real-time control of continuous processes [11]. Precise control and lean manufacturing will help to reduce the high cost of drugs.

Our devices are essential to reducing the problems of information overload in many pharmaceutical manufacturing environments. As an added bonus, similar our devices will also be useful to guaranteeing the correct delivery of the manufactured product to patients in hospitals and hospices.

2. MOTIVATION

Diabetes is a problem in the US and worldwide. About 17 million persons in America have Diabetes mellitus, but five million of them do not know it. Nearly 1 million new cases are diagnosed each year. The disease affects men and women of all ages and ethnic groups. African Americans, Latinos, American Indians, Alaskan Natives, Asian Americans, and Pacific Islanders are more greatly affected than other groups [12]. With Type 2 diabetes, the more common type, your body does not make or use insulin well. Without enough insulin, the glucose stays in your blood. Over time, having too much glucose in your blood can cause serious problems. It can damage your eyes, kidneys, and nerves. Diabetes can also cause heart disease, stroke and even the need to remove a limb. Pregnant women can also get diabetes, called gestational diabetes [13].

- The total annual economic cost of diabetes in 2007 was estimated to be $174 billion. Medical expenditures totaled $116 billion and were comprised of $27 billion for diabetes care, $58 billion for chronic diabetes-related complications, and $31 billion for excess general medical costs.

* This research was supported in part by NSF grants OISE-0405349, ACI-0305466, CNS-0719626, and ACI-0324876, DOE grant DE-FC26-08NT4, and Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

Indirect costs resulting from increased absenteeism, reduced productivity, disease-related unemployment disability, and loss of productive capacity due to early mortality totaled $58 billion. This is an increase of $42 billion since 2002. This 32% increase means the dollar amount has risen over $8 billion more each year.

- The 2007 per capita annual costs of health care for people with diabetes is $11,744 a year, of which $6,649 (57%) is attributed to diabetes.
- One out of every five health care dollars is spent caring for someone with diagnosed diabetes, while one in ten health care dollars is attributed to diabetes [14].

The U.S.S Food and Drug Administration (FDA) has noted that U.S.S drug products are of generally high quality, but there is an increasing trend toward manufacturing-related problems that lead to recalls, disruption of manufacturing operations and loss of availability of essential drugs. Low manufacturing process efficiency (<30%) has also led to increased cost of drugs. Emphasis on cGMP (current good manufacturing practice) as the means of controlling drug quality has led to reluctance among companies to innovate in the manufacturing sector. Such problems have led the FDA to conclude that a new scientific understanding of the drug production process achieved through the use of new sensing technologies can provide science-based, risk-based approaches to the regulation of drug quality, thereby alleviating these problems. Process analytical technologies (PAT) like near infrared (NIR) spectroscopy have been selected as the model for the U.S. to use in shifting successfully from empirical standards like cGMP to science-based standards for achieving manufacturing process quality [15]

Providing a reprogrammable, networked embedded system that can automatically determine quality and control end products in a manufacturing line will impact not just the pharmaceutical industry, but any similar chemically-based production facility. A likely offshoot is a handheld acoustic device that medical providers can use to correctly identify all medications that will be given to patients. Taking the wrong medications is the fifth leading cause of death in the United States.

The identification of pills by using acoustic waves is a challenging problem in technical pharmacy. Manufacturing lines are extremely noisy and we typically only have 8-10 milliseconds to identify each defective or incorrect pill. Our goal is to catch 100% of the possible problem pills and get them out of the manufacturing line.

In the U.S., a pilot line is created, tested, and eventually approved by the FDA. Once approved, it becomes a production line and cannot be modified in any significant way. Any significant modification to a production line makes it a pilot line again with all of the testing and approvals required again. Further, the FDA has to approve to conversion back to a pilot line, which it may not do.

Spherix is a small publicly traded pharmaceutical company that is now manufacturing Tagatose [16-18]. The manufacturing line is in Italy and is owned by a larger company that manufactures on contract for others, including Spherix. We are in a position to create a complete CPS and test it on a real pharmaceutical manufacturing line in Italy, where the rules for testing new pharmaceutical technology is different than in the U.S.

3. MANUFACTURING PROTOTYPE

The system we will create corresponds to a CPS because it involves combining a physical process (pill manufacturing and environmental factors) with integrated sensing and processing (ISP) with a number of human factors that come into play based on who is working on the manufacturing line during a shift and the peculiarities of the individual workers. Incoming raw materials vary, too, and the system needs fine tuning on a regular basis, which requires some significant computing to create downloadable libraries for the actual CPS devices.

We are creating the CPS in a series of steps. First, we have a semi-empirical Simulink [19] program TagSim that models the production process and predicts output yield and impurities for the tagatose production process. Pilot-scale TagSim runs have been executed safely over the Internet. They provide training for graduate and undergraduate students in process control. Remote access to the system uses a client-server paradigm. Servers connected to the processes through standard data acquisition hardware and Matlab Communication Toolbox routines address issues regarding safety, security, data validation, and session management. Implementing TagSim in Simulink and Matlab allows the team to design, test, and verify control strategies in real time over the Internet. Interlocks and cutoffs are easily included to ensure safe remote operation of the process units. Studies to date suggest that computation and network speeds can sometimes become a concern when process dynamics occur on the timescale of the chemometrics and network I/O. Performing computation on the detectors through SCISP should eliminate the problem, leading to tighter process control.

The CPS device uses Integrated sensing and processing acoustic resonance spectroscopy (ISP-ARS), which is a novel approach to acoustic spectroscopy that can be implemented using instruments as simple as an MP3 player or far more complex depending on the environment that ISP-ARS will be used in. Pharmaceutical manufacturing lines for pills are extremely noisy environments. Simple acoustic devices are worthless unless a sound box is created to house the acoustic device while letting pills pass by.

The CPS device must also be very fast since pills pass by quickly and we want to analyze all of the pills and be able to mark defective ones immediately. To date we have a prototype CPS device, but it is neither fast enough now capable of marking defective pills.

We can deliver an infinite number of acoustic spectra, but that defeats the creation of a small, embedded CPS device that is useful in itself. Instead we choose a small number of spectra, which changes slightly over time based on environmental and personnel factors.

Once the spectrum of a sample has been collected, it will be classified to determine the substance present. The Bootstrap Error-adjusted Single-sample Technique (BEST) [20] is the analytical basis of our ARS-ISP CPS device, and the foundation for the pill chemical identification library. The BEST metric is a clustering technique for exploring distributions of spectra in hyperspace.

A sample spectrum will be compared to each substance in a biogeochemical and industrial library based on its direction and distance, measured in standard deviation units, from the known substances. BEST handles asymmetric standard deviations surrounding each substance nonparametrically, allowing more precise discrimination than other metrics, e.g., a Mahalanobis distance [Casis et al 1993]. A sample within 3 standard deviation units of a substance will be considered to be composed of the matching substance while others will be classified as unknown substances.

For a given library entry, the BEST algorithm will be suitably approximated using multiple linear regression to substantially reduce computational requirements. In this implementation, BEST standard deviation units will be precalculated before the ARS-ISP CPS device is deployed in a large number of directions from the population means, and multiple linear regression will be used to fit the standard deviation contours as a function of direction.

The BEST classification algorithm will be performed in situ, allowing a sensor to classify many samples, only notifying the simulation when an interesting substance is found. An initial library will be computed based on substances likely to be found in the target environment. When a substance unknown to the BEST library is found or is out of range (indicating a defect or foreign substance entirely), the CPS will mark the oddities for removal from the manufacturing line.

A secondary use for our CPS is after the correctly manufactured pills are produced and packaged, it can identify that the contents of the packaging is what is on the labels. We can do this easily by combining a bar code scanner with the ARS-ISP CPS device to once again verify the contents. Millions of pills each year have to be recalled due to mislabeling.

4. CONCLUSIONS

We have a system that works in a laboratory setting. It can be used to

- Reduce errors in pill dispensation and usage.
- Identify defective pills in the manufacturing process in real-time.
- Identify pills in the wrong packaging at end of manufacturing.

We have a willing test partner for the manufacturing application. Testing the device on a real manufacturing line is essential to verifying the research ideas and goals of the project. We have a partner in Italy (who does not need FDA approval) who makes pills willing to put our system on a production line. The company cannot reliably make a new drug that will be useful for diabetes patients and does not understand where the process is failing. Our devices will isolate where they are failing and hopefully lead to a change in the manufacturing process of the drug.

REFERENCES

[1] http://www.dddas.org. Includes project descriptions, many DDDAS workshop virtual proceedings, and links to DDDAS software.

[2] C. C. Douglas, A. Deshmukh, M. Ball, R. E. Ewing, C. R. Johnson, C. Kesselman, C. Lee, W. Powell, R. Sharpley, Dynamical data driven application systems: Creating a dynamic and symbiotic coupling of application/simulations with measurements/ experiments, National Science Foundation, Arlington, VA, 2000. http://www.dddas.org/nsf-workshop-2000/worksop_report.pdf.

[3] K. Baldridge, G. Biros, A. Chaturvedi, C. C. Douglas, M. Parashar, J. How, J. Saltz, E. Seidel, A. Sussman, January 2006 DDDAS Workshop Report, National Science Foundation, 2006, http://www.dddas.org/nsf-workshop-2006/wkshp_report.pdf.

[4] 2003 Dynamic Data-Driven Application Workshop, F. Darema, ed., in Computational Science - ICCS 2003: 3rd International Conference, Melbourne, Australia and St. Petersburg, Russia, June 2-4, 2003, Proceedings, Part IV, P.M.A. Sloot, D. Abramson, A.V. Bogdanov, J.J. Dongarra, A.Y. Zomaya, Y.E. Gorbachev (Eds.), Lecture Notes in Computer Science, Vol. 2660, Springer-Verlag Heidelberg, 2003, pp. 279-384.

[5] 2004 Dynamic Data-Driven Application Workshop, F. Darema, ed., in Computational Science - ICCS 2004: 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part III, Marian Bubak, Geert Dick van Albada, Peter M. A. Sloot, and J.J. Dongarra (eds.), Lecture Notes in Computer Science series, vol. 3038, Springer-Verlag Heidelberg, 2004, pp. 662-834.

[6] 2005 Dynamic Data-Driven Application Workshop, F. Darema, ed., in Computational Science - ICCS 2005: 5th International Conference, Atlanta, Georgia, USA, May 22-25, 2005, Proceedings, Part II, Vaidy S. Sunderam, Geert Dick van Albada, Peter M.A. Sloot, Jack J. Dongarra (eds.), Lecture Notes in Computer Science series, vol. 3515, Springer-Verlag Heidelberg, 2005, pp. 610-745.

[7] 2006 Dynamic Data-Driven Application Workshop, F. Darema, ed., in Computational Science – ICCS 2006: 6th International Conference, Reading, UK, May 28-31, 2006, Proceedings, Part III, edited by V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, and J.J. Dongarra, Lecture Notes in Computer Science 3993.

[8] 2007 Dynamic Data-Driven Application Workshop, F. Darema, ed., in Computational Science – ICCS 2007: 7th International Conference, Beijing, P.R. China, May 28-31, 2007, Proccedings, Part I, Y. Shi, G. D. van Albada, P. M. A. Sloot, and J. J. Dongarra (eds.), Springer-Verlag, Heidelberg, Lecture Notes in Computer Science series, vol. 4487, 2007, pp. 955-1245.

[9] 2008 Dynamic data-driven application systems - DDDAS 2008, C.C. Douglas, ed., in Computational Science – ICCS 2008: 8th International Conference, Krakow, Poland, June 23-25, 2008, Proceedings, M.Bubak, G. D. van Albada, P. M. A. Sloot, and J. J. Dongarra (eds.), Springer-Verlag, Berlin Heidelberg, Lecture Notes in Computer Science series, vol. 5103, 2008, pp. 3-63.

[10] 2009 Dynamic data-driven application systems - DDDAS 2008, C.C. Douglas, ed., in Computational Science – ICCS 2008: 8th International Conference, May 25-27, 2009, Proceedings, Part II, Gabrielle Allen, Jaroslaw Nabrzyski, Edward Seidel, Geert Dick van Albada, Jack J. Dongarra, and Peter M. A. Sloot (eds.), Lecture Notes in Computer Science series, vol. 5545, Springer-Verlag, Heidelberg, 2009, pp. 445-509.

[11] B. Dai A. Urbas C C. Douglas R. A. Lodder, Molecular factor computing for predictive spectroscopy. Pharm Res. 8 (2007), pp.1441-1449.

[12] http://www.fda.gov/diabetes/qna.html. Checked 2/25/2009.

[13] http://www.nlm.nih.gov/medlineplus/diabetes.html. Checked 2/25/2009.

[14] http://www.diabetes.org/diabetes-statistics/cost-of-diabe

tes-in-us.jsp. Checked 2/25/2009.
[15] http://www.fda.gov/ohrms/dockets/ac/02/briefing/3869 B1_08_woodcock/sld001.htm. Checked 2/25/2009.
[16] Y. Lu, G. V. Levin, and T. W. Donner, Tagatose, a new antidiabetic and obesity control drug, Diabetes Obes Metab. 10 (2008), pp. 109-134.
[17] J. R. Beadle, J. P. Saunders, and T. J. Wajda, Jr, Process for Manufacturing Tagatose, Biospherics Incorporated, 1991 (US patent 5,002,612). Available from URL: http:// www.uspto.gov/patft/index.html. Checked 7/5/2009.
[18] J. R. Beadle, J. P. Saunders, and T. J. Wajda, Jr, Process for Manufacturing Tagatose, Biospherics Incorporated, 1992 (US patent 5,078,796). Available from URL: http: //www.uspto.gov/patft/index.html. Checked 7/5/2009.
[19] http://www.mathworks.com, Simulink and the Communications Toolbox. Checked 2/26/2009.
[20] W. Dieter, R. A. Lodder, J. E. Lumpp, Jr., Scanning for extinct astrobiological residues and current habitats (SEARCH) using integrated computational imaging, IEEE Aerospace and Electronic Systems, 2005.
[21] L. A. Cassis and R. A. Lodder, Near-IR imaging of atheromas in living arterial tissue, Analytical Chemistry, 65 (1993), pp. 1247-56.

Craig Douglas is a School of Energy Resources Distinguished Professor of Mathematics at the University of Wyoming, where winter never ends. He is also a research affiliate in the computer science department of Yale University. He graduated from the University of Chicago in 1977 and from Yale University in 1982 with a specialty in scientific computing and numerical analysis. He has been a professor of mathematics, mechanical engineering, and computer science at the University of Kentucky, a research staff member at the IBM Thomas J. Watson Research Center, and an assistant professor at Duke University. He has also been a foreign guest professor at Wuhan University, a Zi-Qiang Professor at Shanghai University, and visiting professorships at the University of Pavia, Yale Univesity, and Texas A&M University. His research interests are in parallel processing, Grid computing, hardware acceleration algorithms, data-driven applications, and computational sciences.

Robert Lodder is a Professor Chemistry, Pharmacy, and Electrical and Computer Engineering at the University of Kentucky. He is a graduate of Xavier University and Indiana University. He is also President of Spherix, Inc. (member of NASDAQ), has founded two other pharmaceutical product related companies, and holds the distinguished classification of a Dr. Wizard. He was an assistant and associate professor at the University of Kentucky. His research interests are in pharmacy, chemistry, and integrated sensing and processing devices.

Gundolf Haase is a Professor of Mathematics and Computational Sciences at Karl-Franzens University of Graz. He is a graduate of Karl-Marx-Stadt University (now the University of Chemnitz). He has been an assistant professor at the Johannes Kepler University of Linz and a visiting professor at the University of Kentucky. His research interests are in parallel computing, algebraic multigrid, and GP-GPU hardware acceleration of algorithms.

Li Deng is a research scientist at the JSOL Company (Tokyo) and a visiting scientist at Rensselear Polytechnic Institute. She graduated from Sun Yat-Sen University (Guangzhou) in 1984 and the University of Electro-Communications (Tokyo) in 1998. She has held positions at the Japan Research Institute, Tokyo Institute of Technology, and the University of Kentucky. Her research interests include parallel finite element methods for eigenvalue problems related to reducing noise in moving vehicles (and has a patent in Japan for one of her algorithms), data-driven applications, and numerical analysis.

Thad Hannel and **David Link** are research assistants in Chemistry at the University of Kentucky. Each expects to graduate with a Ph.D. during the second half of 2009.

A PARALLEL STATISTICAL LEARNING APPROACH TO THE PREDICTION OF BUILDING ENERGY CONSUMPTION BASED ON LARGE DATASETS

Hai Xiang ZHAO, Frédéric MAGOULÈS
Applied Mathematics and Systems Laboratory, Ecole Centrale Paris
Grande Voie des Vignes, 92295 Châtenay-Malabry Cedex, France
Email: zhaohaix@ecp.fr, frederic.magoules@hotmail.com

ABSTRACT

The prediction of future energy consumption of buildings based on historical performances is an important approach to achieve energy efficiency. A simulation method is here introduced to obtain sufficient clean historical consumption data to improve the accuracy of the prediction. The widely used statistical learning method, Support Vector Machines (SVMs), is then applied to train and to evaluate the prediction model. Due to the time-consuming problem of the training process, a parallel approach is applied to improve the speed of the training of large amounts of data when considering multiple buildings. The experimental results show very good performance of this model and of the parallel approach, allowing the application of Support Vector Machines on more complex problems of energy efficiency involving large datasets.

Keywords: Support Vector Machines (SVMs), Prediction, Model, Energy Efficiency, Parallel Computing.

1. INTRODUCTION

Building energy consumption plays an important role in the total energy consumption of end use. Energy efficiency in building management and retrofit is important for energy conservation which significantly benefits both individuals and society. For a specific building, the energy transformation is treated as a complex system with many factors involved, such as building construction and materials, ambient weather conditions, occupants' behaviors, inner facilities, etc. In fact, most of these factors vary irregularly from one building to another building and from time to time, making analysis of the energy performance extremely difficult in practice. Many works have been done in the past few decades in order to find out the empirical rules of energy performance. Some work aimed to simulate and evaluate the energy efficiency of buildings or facilities from the engineering point of view [1, 2]. A large number of tools were developed such as DOE-2, Apache, AkWarm, etc. Other works took the historical energy consumption behaviors into consideration, and tried to find out the rules from those recorded data and then applied these rules to predict the performance in unknown periods or conditions. Some models and methods were proposed for this kind of data mining approach, e.g. regression model [3, 4], time-series analysis [5], neural networks [6–10] and statistical learning method [11–13]. This last method is widely used in industry to analyze complex real-life problems due to the great generalization ability of Support Vector Machines (SVMs) in solving non-linear problems [14, 15].

Dong et al. [11] first applied SVMs to predict the monthly electricity consumption of four buildings in the tropical region. Three years' data was trained and the derived model was applied to predict the landlord utility in one year. Lai et al. [12] applied robust regression model on one year's data of one building with electricity consumption and climate involved, then they applied the model to three month's data to test the predication ability of their model. They also trained models on each daily basis datasets and then compared the obtained models to verify the stability of this method. In addition, they added some perturbations to a certain part of the data and tried to detect these perturbations by examining the change of contributing weights. Li et al. [13] compared SVMs with the conventional back-propagation neural network in predicting the hourly cooling load of an office building. SVMs method was proved to be better than the traditional solution. The above three work has shown good performances of SVMs method in predicting hourly and monthly building energy consumption. However, some problems remain in these analyses. Firstly, the predictions are based on the historical performance of the same building, which means that we do not know the ability of SVMs in predicting the energy performance in a completely new building. Secondly, only the dependence of energy requirements on a limited number of features, such as weather conditions, is considered. Obviously this limitation leads to inaccuracy in the previous models.

To improve the accuracy of the predicting model, this paper presents a simulation method to obtain enough energy consumption data for multiple buildings. Also, it applies SVMs to predict the performance of a new building with construction parameters involved. A parallel implementation is then applied to very large dataset in order to make SVMs more practical in solving complex problems. The plan of the paper is the following. In section 2, we briefly introduce the theory of SVMs method and its parallel approach. In section 3, we present the simulation method for obtaining energy consumption data of multiple buildings. The data analysis and numerical experiments are then presented in section 4. Finally, conclusions are given in section 5.

2. SUPPORT VECTOR REGRESSION METHOD

Support vector machines aim at finding a decision function (model) to represent the relationship between the features and the target. Let vector x_i indicates the *ith* sample of the features and y_i represents the corresponding target value. Therefore, all the samples can be represented as:

$$(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)$$

where $x_i \in K^n$, $y_i \in K$ and l is the number of samples. If y is a continuous variable with real values, then this method is called a support vector regression (SVR). The decision function we are looking for can be expressed as:

$$f(x) = \omega x + b \qquad (1)$$

where $f(x_i)$ is the estimation of the corresponding y_i and where ω and b are coefficients. If such a function is found, we can use it to estimate the unknown y_i with the new feature x_i (this is called the prediction ability of the extracted model). The process to find this function is usually called model training with the main task to find out the proper values of ω and b . We approximate them by minimizing the empirical risk with respect to the following loss function:

$$L(y - f(x)) = \begin{cases} 0 & if \left| y - f(x) \right| \le \varepsilon \\ \left| y - f(x) \right| - \varepsilon & otherwise \end{cases}$$

This ε -insensitive function supposes an assumption that there is no deviation of the estimated value to the measured one if their distance is less than ε . That means that we only consider the points outside the ε -tube around the measured values while we are optimizing the decision function. To find out the vector ω , the problem is equivalent to the following quadratic optimization problem with two new variables ξ_i and ξ_i^* , *i = 1,2,..,.l*, where we minimize

$$\frac{1}{2}\|\omega\|^2 + C\left(\sum_{i=1}^{l}\xi_i^* + \sum_{i=1}^{l}\xi_i\right) \tag{2}$$

under the constraints

$$y_i - f(x_i) \le \varepsilon + \xi_i^*$$
$$f(x_i) - y_i \le \varepsilon + \xi_i$$
$$\xi_i^*, \xi_i \ge 0, \qquad i = 1,2,...,l$$

where C is a regularizing constant, which determines the trade off between the capacity of $f(x)$ and the number of points outside the ε -tube. To solve the problem of Eq. (2), we can use a Lagrange function by introducing four Lagrange multipliers, α^* , α , γ^* , γ .

$$L(\omega,b,\xi^*,\xi,\alpha^*,\alpha,\gamma^*,\gamma) = \frac{1}{2}\|\omega\|^2 + C\left(\sum_{i=1}^{l}\xi_i^* + \sum_{i=1}^{l}\xi_i\right)$$
$$-\sum_{i=1}^{l}(\gamma_i^*\xi_i^* + \gamma_i\,\xi_i\,) - \sum_{i=1}^{l}\alpha_i\left[y_i - (\omega\cdot x_i) - b + \varepsilon + \xi_i\right]$$
$$-\sum_{i=1}^{l}\alpha_i^*\left[(\omega\cdot x_i) + b - y_i + \varepsilon + \xi_i^*\right] \tag{3}$$

The four Lagrange multipliers satisfy the constraints $\alpha^* \ge 0$, $\alpha \ge 0$, $\gamma^* \ge 0$ and $\gamma \ge 0$, *i = 1, 2, ..., l*. If the relations

$$\frac{\partial L}{\partial \omega} = \frac{\partial L}{\partial b} = \frac{\partial L}{\partial \xi^*} = \frac{\partial L}{\partial \xi} = 0$$

occur, we can get the following conditions,

$$\omega = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)x_i \tag{4}$$

$$\sum_{i=1}^{l}\alpha_i^* = \sum_{i=1}^{l}\alpha_i \tag{5}$$

$$0 \le \alpha_i^*, \alpha_i \le C \tag{6}$$

$$C = \alpha_i^* + \gamma_i^* = \alpha_i + \gamma_i\,, \qquad i = 1,2,...,l \tag{7}$$

Putting them back into the Lagrange function, we can obtain the solution of the optimization problem through maximizing Eq. (3) with respect to the Lagrange multipliers. In the resulting function, there is a dot product of two vectors x_i and x_j . In practice, it is difficult to find out a linear function $f(x)$ for problems involving large dataset. Therefore, we need to map the 1-dimensional problem into a higher dimensional feature space where it is easier to find a linear function similar to $f(x)$ in a lower dimensional space. Fortunately, it is not necessary to express explicitly the mapping during the computation. Actually, the dot product can be replaced by a kernel function $K(x_i, x_j)$. Putting Eq. (4) into Eq. (1), and taking into consideration the kernel function, the decision function can be alternated as:

$$f(x) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(x_i\cdot x) + b$$

and the optimization problem becomes: Maximize the quadratic function:

$$W(\alpha_i^*,\alpha_i) = \sum_{i=1}^{l}y_i(\alpha_i^* - \alpha_i) - \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i)$$
$$-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i\cdot x_j) \tag{8}$$

under the constraints (5) and (6). Normally, only a certain parts of the samples satisfy the property of $\alpha_i^* - \alpha_i \ne 0$, these parts are called support vectors (SVs).

There are four frequently used kernel functions, the linear function, the polynomial function, the radial basis function (RBF) and the sigmod function. They represent different decision shapes in the feature space. In our work, we chose RBF in the training process because it has been tested to be proper for a couple of industrial applications [11]. RBF is also called the Gaussian kernel, it has the form $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ where $\gamma > 0$ is the kernel parameter.

Sequential minimal optimization (SMO) method was proposed in [16] to solve Eq. (8) instead of standard quadratic problem (QP) techniques. The main idea of SMO is to divide the entire optimization problem into sub-problems and solve each sub-problem independently. In each step, two Lagrange multipliers are optimized analytically until the Karush-Kuhn-Tucher (KKT) conditions are satisfied. Which of the two Lagrange multipliers are selected in the current step is decided by a heuristics algorithm. The parameter b is re-computed in each step. The SMO method has been proved to be a better alternative to handle large datasets because it successfully solved memory load and numerical problems.

However, the problem solving is a time-consuming process when applied to large datasets. In [15], Brugger proposed a parallel approach to solve the quadratic problem by using SMO method. By profiling the performance of the training process on different datasets, the author found that the most time-consuming part is located in the kernel evaluation. Therefore, the author parallelized the kernel evaluations and gradient updates, then combined with inner sequential QP solver and distributed storage of kernel rows, and achieved linear speedup for regression. The tool Pisvm was developed accordingly by the author. In our work, we applied this tool to train our model in order to accelerate the training process in high burden conditions.

In our work, we applied two performance evaluation methods. The first one is the mean squared error (MSE) which gives the average deviation of the predicted values to the measured one. Lower the MSE, better is the performance of the prediction. The second one is the squared correlation coefficient (SCC) which lies between [0,1] and gives the ratio of successfully predicted number of target values on total number of target values, i.e. how accurate the predicted values are compared to the measured one. Higher the SCC, stronger is the evaluating ability.

3. OBTAINING HISTORICAL DATA

The main energy form in a building is the electric and gas consumption for every day use and the space heating demand in cold weather. The energy requirement for a building depends on many factors, so that it would be quite different for each building and quite different upon the time for one given building. The most important influence factors are the ambient weather conditions, building constructions and materials, inner occupants' behaviors. Therefore, as indicated in [17], it is difficult to obtain accurate consumption data from the real world, normally from measurement. Here we adopt a simulation method by using EnergyPlus [2]. We choose this software because it is a succession of the widely used building energy simulation software including BLAST and DOE-2. It is well tested and has comprehensive ability in calculating complex systems.

In this paper, an office located in Paris-Orly is simulated. The most important parameters for this building are given in Table 1. It is a simple building model with only one room and one thermal zone. The occupants' behaviors, e.g. the people using of facilities, air ventilation, are scheduled as normal office hours. We suppose that the space is heated by district heating, which means that the inlet temperature of the heating system keeps constant. We decided to do this simulation with really recorded weather data from November 1st to March 31th because this duration is a typical heating season in this location and we will analyze how much district heating energy would be consumed by this building. The weather conditions contain solar radiation, dry bulb temperature, relative humidity and so on. To indicate the weather data, the dry bulb air temperatures of the first 20 days in January and July are plotted in Table 1.

Table 1. Information of A Single Building (in metric units)

Parameters	Values
Location	Paris-Orly, City
Duration	From Nov 1 to Mar 31
Building Shape	Rectangle
Structure	Length:11 Width:10 Ceiling Height:4 North axis: 10^o
Walls	1IN Stucco, 0.0253m 8IN Concrete HW, 0.2033m Insulation, 0.0679m
Fenestration surface	$14m^2$ for each wall
Thermal Zones	1
Number of people	14
Air infiltration	0.0348 m^3/s
Heating type	District heating
Cooling type	Heating, ventilating, and air conditioning (windowAirCondioner)
Other facilities	Light, Water heater

The output of the simulation consists of hourly recorded data related to the energy transformation of the building. These data are preprocessed in order to extract useful information and reorganized according to the format required by the analyzing tool. In our model training step, we take district heating demand as the target variable and 25 other variables as the features including day type which indicates if the current day is holiday or not, hourly weather conditions, zone mean air temperatures, infiltration volume, heat gain through each window, heat gain through lights and people, zone internal total heat gain.

The energy consumption data for multiple buildings is generated based on the parameters of the above single one. An interface was created to automatically simulate multiple buildings under EnergyPlus. The input file is divided into two parts: an alterable part containing parameters which would be different for each building and a stable part containing parameters which would be kept the same for every building. For example, the building constructions are in the alterable part while weather data is in the stable part. For the alterable parameters, the values are obtained by a stochastic method in a reasonable range. The two parts are then combined together to form the final input file. After successfully simulating one building, we update the alterable part and repeat the simulation process for a new one. In order to analyze multiple buildings, it is necessary to put the output results of all buildings into one single output file in the proper format. The whole controlling flow is presented in Figure 1.

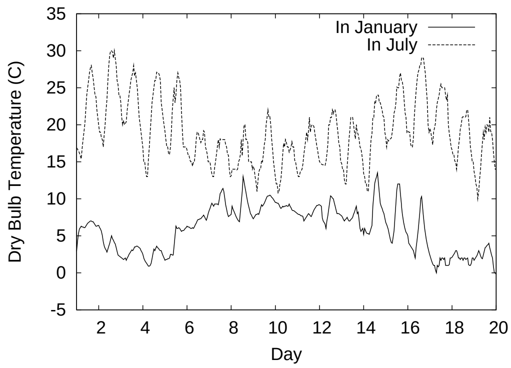


Figure 1. Dry Bulb Temperature in The First 20 days of January and July.

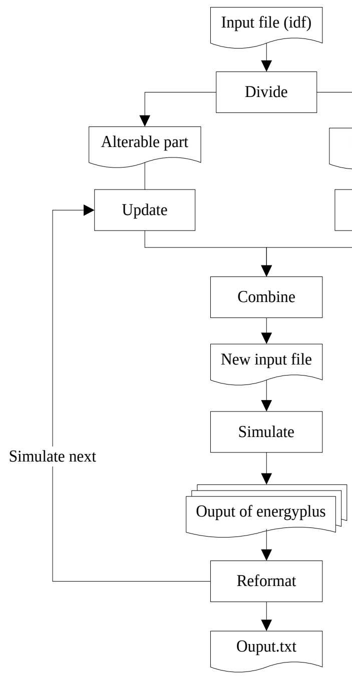


Figure 2. Flow chart of generating energy consumption for multiple buildings.

4. EXPERIMENTS AND NUMERICAL RESULTS

The dataset is divided into two parts, one is used for training models which indicates the dependence of the target on the features and the other one is used for testing the prediction performance of the obtained models. In the testing step, the predicted target regarding to the new values of features is compared with the measured one. Before training the data by SVR, one has to scale the values linearly into a small range in order to avoid numerical problems in the calculations. We scale

the training subset into the range [0,1] and then apply the scaling function to scale the testing subset. The experiments are performed on a cluster composed by two homogeneous workstations with one Gigabit Ethernet network. Each computer has 8 processors (2.5GHz), 1333MHz FSB and 4G memory. The operating system used is Linux, the kernel version is 2.6.
Optimizing the parameters of SVR is important to improve the generalization ability of the models. The best choice will give the model well performance without over-fitting problem. In our work, RBF kernel is selected to train the model, so that the parameters need to be optimized are C, γ, ε. The estimation of γ is solved by $\gamma = l^2 / (\sum_{i,j=1}^{l} \|x_i - x_j\|^2)$ as indicated in [15] and [18]. The parameters C and ε are optimized by 5-fold cross validation on randomly selected 3000 samples from the training set.

One hundred buildings located in the same place in heating season are here simulated. The energy consumption data of the first 99 buildings is taken as training set, and the data for the last building is taken as testing set. Therefore, the number of samples for training is equal to 358776 and the number of samples for testing is 3264. The demand of district heating is taken as the target and the other 28 parameters such as weather conditions, the building structures, are taken as features in the analyzing process. The parameters for SVR and RBF kernel are set as $C = 4$, $\gamma = 0.3179$ and $\varepsilon = 0.01$. As the result shows, the number of support vectors are 27501, MSE is 5.01e-5 and SCC is 0.997639. The predicted target compared with the measured one for the first 100 samples is plotted in Figure 3. The model shows very good generalization performance in the result. It proves that SVR with optimized parameters has a comprehensive ability in predicting energy consumption of a new building if we provide enough historical information to the model training.

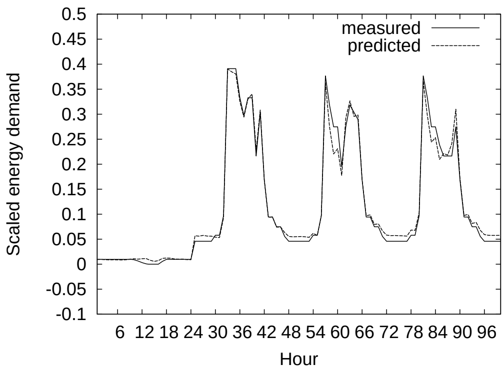


Figure 3. Measured and Predicted District Heating Demand for The Last Building in Heating Season.

In order to improve the precision of the prediction, one has to collect more historical samples for training. However, the calculating time on processors would increase very fast while the data size is growing. It is a trade-off between the modeling time and the precision. In the above experiment, the model training process needs 31.4 hours in sequential implementation of SVMs — Libsvm [19]. It is really long for analyzing 100 buildings. To make the analyzing method more practical, it is necessary to reduce the learning time on large datasets. In next experiment, we turn to a parallel approach of SVMs to train the model and test how it performs on this problem.

We analyzed the same dataset as in the above experiment on 1, 2, 4, 6, 8 processors in Pisvm. The cache size is set as 256MB for each process of MPI. Every training process is repeated three times. The average running time and speedup are calculated and shown in Figure 4. From the curves we can see that the speedups of the parallel implementation on 2 and 4 processors are quite close to linear speedup. But the performance enhancement seems not obvious when the number of processors is increased to 6 and 8. That is because of the limitation of our testing environment. On one side, SMP structure inside one node can not make distributed cache approach achieving its full effect when the number of processes is big. On the other side, gigabit network connection is very slow compared to the high performance processors. However, from the performance under 2 and 4 processors, a better speedup can be expected on servers connected by higher speed devices such as infiniBand.

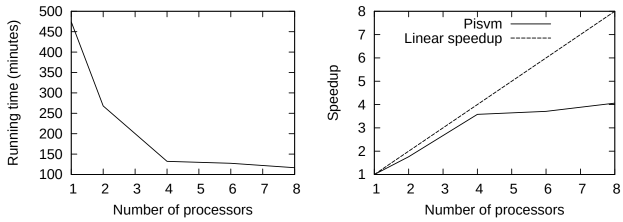


Figure 4. (a) Running Time of The Training Process Using A Parallel Implementation of Svms. (b) Comparison of The Speedup with A Theoretical Optimal Linear Speedup.

The results obtained in those experiments, the number of SVs, MSE and SCC, keep stable under different number of processors. The average results of parallel implementation are compared with sequential implementation in Table 2. This table indicates that, the prediction performance of the parallel solution on large datasets is quite close to the sequential one performed on smaller datasets. The parallel solution can be applied to predict the building energy performances in more complex situations which will allow longer time duration and more features and sample buildings.

Table 2. Comparison of parallel and sequential implementations: The number of Support Vectors (SVs), Mean Squared Error (MSE), Squared Correlation Coefficient (SCC).

Implementations	SVs	MSE	SCC
Sequential	27501	5.01097e-05	0.997639
Parallel	27382	5.08532e-05	0.997571

5. CONCLUSIONS

This paper introduces a simulation approach to collect enough historical time series data for multiple buildings' energy consumption. A statistical learning method is then applied to predict the energy behavior in a completely new building. A parallel implementation of support vector regression with RBF kernel is applied to analyze large amounts of energy consumption data. Experimental results demonstrate that SVR with optimized parameters show a very good generalization ability on such problems. The good performance of the parallel implementation enhances the potential application of SVR on more complex models involving large amounts of samples or features.

REFERENCES

[1] M.S. Al-Homoud. Computer-aided building energy

analysis techniques. Building and Environment, 36:421–433, 2001.

[2] EnergyPlus. Available on line at: http://www.EnergyPlus.gov, 2008. (Accessed July 2009).

[3] D. Ruch and D.E. Claridge. A four parameter change-point model for predicting energy consumption in commercial buildings. ASME Journal of Solar Energy Engineering, 114(2):77–83, 1992.

[4] S. Katipamula, T.A. Reddy and D.E. Claridge. Multivariate regression modeling. ASME Journal of Solar Energy Engineering, 120:177–184, 1998.

[5] A. Kimbara, S. Kurosu, R. Endo, K. Kamimura, T. Atsuba and A. Yamada. On-line prediction for load profile of an air-conditioning system. ASHRAE Transactions, 101(2):198–207, 1995.

[6] G. Mihalakakou, M. Santamouris and A. Tsangrassoulis. On the energy consumption in residential buildings. Energy and Buildings, 34:727–736, 2002.

[7] J.F. Kreider, D.E. Clardige and P. Curtiss. Building energy use prediction and system identification using recurrent neural networks. Journal of solar energy engineering, 117(3):161–166, 1995.

[8] P.A. Gonzàlez and J.M. Zamarreo. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. Energy and Buildings, 37(6):595–601, 2005.

[9] S. Karatasou, M. Santamouris and V. Geros. Modeling and predicting building's energy use with artificial neural networks: Methods and results. Energy and Buildings, 38(8):949–958, 2006.

[10] B.B. Ekici and U.T. Aksoy. Prediction of building energy consumption by using artificial neural networks. Advances in Engineering Software, 40(5):356–362, 2009.

[11] B. Dong, C. Cao and S.E. Lee. Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37(5):545–553, 2005.

[12] F. Lai, F. Magoulès and F. Lherminier. Vapnik's learning theory applied to energy consumption forecasts in residential buildings. International Journal of Computer Mathematics, 85(10):1563–1588, 2008.

[13] Q. Li, Q.L. Meng, J.J. Cai, Y. Hiroshi and M. Akashi. Applying support vector machine to predict hourly cooling load in the building. Applied Energy, 86(10):2249–2256, 2009.

[14] V.N. Vapnik. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[15] D. Brugger. Parallel support vector machines. In VLSI-SoC: Proceedings of the IFIP International Conference on Very Large Scale Integration of System on Chip. Springer, 10 2007.

[16] J.C. Platt. Fast training of support vector machines using sequential minimal optimization. MIT Press, Cambridge, MA, USA, 1999.

[17] L. Pérez-Lombard, J. Ortiz and C. Pout. A review on buildings energy consumption information. Energy and Buildings, 40(3):394–398, 2008.

[18] I.W. Tsang, J.T. Kwok and K.T. Lai. Core vector regression for very large regression problems. In ICML'05: Proceedings of the 22nd international conference on Machine learning, pages 912–919, New York, NY, USA, 2005. ACM.

[19] C.C. Chang and C.J. Lin. LIBSVM: A library for support vector machines, 2001. Available online at: http://www.csie.ntu.edu.tw/˜cjlin/libsvm. (Accessed July 2009).

RESEARCH AND DESIGN OF MOBILE MESSAGE SYSTEM BASED ON DISTRIBUTED AGENT

Ou Yangquan, Xu Hongyun
School of Mathematics & Computer Science, Jianghan University
WuHan,430056,China
Email:oyq8888@163.com

ABSTRACT

To transmit information among different networks inside a distributed system is a problem encountered by distributed applications. Currently, the solution is to use a message system to transmit information between applications residing in different networks. Message system supports data communication for distributed application, which could be applied in any computer operation system platform, and associates with distributed application system based on data communication. The purpose of the current research is to design a set of service solutions upon the integrated and reliable JMS message system for distributed application in the mobile communication field. The demonstrations of this study dissertate the advantage of using the ORB to construct JMS message system, synchronously the structure model of mobile JMS message system, function definition and designing intention, which could be found in this study, finally the efficient mobile JMS message system designing solution is constructed.

Keywords: ORB, JMS, Distributed agent

1. INTRODUCTION

With sustaining technical development of mobile communication devices, more and more people could obtain the access to receive all kinds of messages at any time, any place through comprehensive mobile network, which strongly supports the mobile distributed application. In the office automatic system based on Internet and communications network, remote office, remote file management, information transfer and other related tasks are the new requirements of mobile application. These requirements could be implemented by distributed computing model. The finding of this study is on how to construct the transfer of distributed computing messages between mobile network application and regular network application, therefore the distributed application of enterprises could be extended into the mobile application platform.

2. SYSTEMATIC ANALYSIS

In current distributed application, model is directly adopt the most bottom network API to communicate between mobile client and gateway. Byte stream as the way of information exchange, but the issue which applies these APIs of byte stream is that the communication protocol of message client and gateway have to be defined, moreover, creating the corresponding protocol stack and transferring all kinds of information via the event object of protocol stack are strict to be demanded. Therefore the developers of message system have to face to all kinds of problems, such as network connection, data buffer, data packaging/unpacking; request transacted and executed, some other abundant and complicated basic operation. Furthermore, due to the close bound between message handling and protocol stack, adding and modifying the functions of message system will be limited, which is adverse to update and expand. So, this kind of function call based on bottom protocol stack broadens the difficulty and complication of development of message gateways and the libraries of JMS mobile client support.

This study uses ORB (Object request Broker) technology to construct mobile JMS message system. ORB is an object of conforming JAVA CORBA norms, combining the advantages of distributed computing and object –oriented, makes the programs which are in the environment of distributed computing be able to realize cross-platform, cross-network communication capability of the component object[1,2].

The advantages of using ORB are as follows:

1) The systematic model based on the real world: With the use of object model, as long as the problem domains of the transaction are adopted, the defined objects, including those properties and behavior, inter-object of interrelation and so on, constructing a software system which is based on the real world will become possible.

2) The partition of logical objects of system function: Each object is so clear to the designers and developers because it only fulfils some special operation. As the various parts of the system are defined as objects and operation, the errors will decrease in the software system, which makes the communication become smoother and easier to be understood. In addition, the software is also much more modular.

3) The extension of the modules as requirements changed: It’s not difficult to alter a well-designed OO system changing the requirements. If all the required objects have been existed in the question fields, to the new requirements it possibly only needs to do a simple action, which adds some new behavior and efficient methods to combine with the existent objects. Adding a new requirement isn’t to add a new relationship; it usually only needs to extend existent objects. Contrarily, adding a new requirement to the model of function-oriented system may conduce the re-designing of the main portion of the system.

Therefore, ORB as core communication framework, defines a set of interfaces similar with standard JMS API, to realize the encapsulation to JMS Messaging service package, finally this method facilitates the development of the libraries of mobile JMS API and brings some convenience into the later maintenance and administration, as well as the change of JMS API in the background only affects the codes encapsulated by addition, without modification and deploying those mobile applications.

What’s more, this framework makes it possible for J2ME application qualified for communication in the remote component objects, which implements the abstraction of application communication based on the Generic Connection Framework of J2ME.It encapsulates low-level communication mechanism such as the technical complexity, hides the details of communication technology library to the application developers. With the service layer object communication, the technical details may not be What's more, such as operating system, network protocols low-level information and physical network addresses, accessing to the remote application, improving mobile applications, interoperability, flexibility, transplantation and maintainability. Hereby, the function and scope of J2ME application could be improved greatly.

3. THE STRUCTURE AND FUNCTION DEFINITION OF MOBILE JMS MESSAGE SYSTEM

With the construction of Enterprise distributed mobile application system, the differences are eliminated between the application of mobile network and the application of solid network, which makes the communication level of wireless network more transparent to the mobile application, the developers don't care which communication is used, and they only need to know how to call those standard APIs. Mobile JMS message system composes of three components: mobile JMS client, mobile message gateway and JMS provider, which can be better shown in Fig. 1.

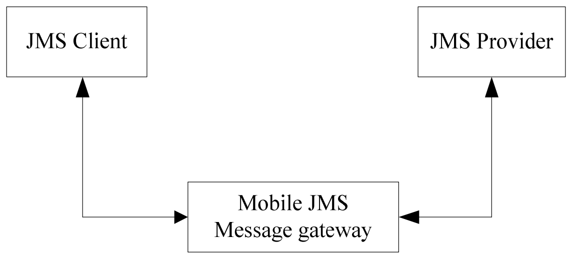


Figure 1. Structure of mobile JMS message system

The mobile JSM client is the user of mobile JMS message system, is the jumping-off point and the end point of the messages, reside in the device. In the pure Java solution, mobile devices must be able to support J2ME, and then a set of lightweight mobile JMS libraries to support the application is quite significant.

The mobile JMS message gateway provides the message service for the wireless device, which is Java–oriented, there are two important functions, saving the messages and transferring the messages. Saving the messages is that the gateways isn't only responsible for saving the messages from the background, but also the messages from the foreground, All the messages are saved into the local messages list and prepare to transfer; transferring the messages is that the gateways choose the corrective path to send out the messages by the destination address and flow direction.

The mobile message gateway is as a standard JMS client communicated with JMS service, on the other hand it is also the service interface of the mobile client to provide the message service eligible of JMS norms, accordingly makes the JMS service be transparent to the mobile client.

With providing the transfer service between mobile application and the message system of the background, the mobile message gateway is just likes a bridge to serve for wire network and wireless network. On this basis, the mobile application client could be constructed quickly and efficiently on implementation of mobile users and enterprise application systems to the seamless integration.

In addition, the message gateway still could support the message logs, user administration, and control of the amount of information transfer, such features as well as good call interface and extensive commonality.

JMS provider is a server and runs in the background. It can be either a stand-alone JMS server, or a JMS Services application server, the final message exchange will be executed in the JMS Provider.

3.1 The Design Objective of Mobile JMS Message System

1) The integrity of the service: Mobile JMS message system could provide an intact message service system. The content of integrity includes two kinds of message transfer module (point to point, publish / contract model), the receiving of asynchronous message and the persistent message, and the basic message types and properties.

2) The compatibility of service interface: the API interface of the message service is similar with the standard JMS API, which could help the developers master the basic developing skill, reduce the access to be as JMS professional engineer and improve development efficiency threshold.

3) The reliability of service: messages will be safely transferred in the system and, not be lost in any condition, especially in the wireless mobile network environment.

3.2 The Functions of Mobile JMS Message System

1) Store and forward / offline operation: message queue should exist in the clients and services. When the connection of the mobile client and the mobile server is valid or terminal, the mobile client saves the messages into the client message queues; those messages will be distributed into the mobile server until the connection is reloaded. In the mobile server the message queues may be temporary or permanent, the messages in the permanent message queues are saved in the database and could be recovered when the system crashes and restarts.

2) Wireless data transmission and optimization for data Compression: the optimization for data compression is a technology to reduce the content of the data compressed, accelerates the transmission of wireless network and decreases the load of the Broad Band.

3) Push technology: it means that the client may set the message channels which they need in their initial using to the mobile message system, later which will be automatically downloaded from the mobile network, without asking actions. A key component of mobile application programs could push data to client.

4) The fast cache technology: bandwidth throttling is the bottlenecks of mobile network, it's reasonable to diminish the access probability of the mobile terminal to the mobile server by adopting caching techniques. This kind of technique is to put the data frequently accessed into the local mobile terminal.

3.3 The Message Stream Of The Mobile JMS System

1) From the JMS provider to the mobile client: In this kind of message stream, the mobile client firstly creates some JMS topic subscriber to inform the mobile server gateway what should be received. When JMS provider publishes a message of the topic, the message will be accepted by the mobile gateway. Meanwhile, the mobile gateway will create the copy of this

message, and save it into the permanent message queues in order to avoid losing all the data when the mobile devices loses the connections. Then, the gateway uses some wireless transmission modes, such as GPRS, transferring the messages to the clients. After the clients confirm the messages received the mobile gateway will delete those messages from the queues.

2) From the mobile client to JMS provider: in this message stream, the mobile client firstly creates some message queue sender, then creates the messages and sends the messages to the mobile gateway through the sender. Once the gateway receives this message, it will be transmitted to the background JMS provider.

3.4 The Cooperation Method of the Mobile Message System

When sending the messages, the mobile message client firstly registers a dialogue related with the mobile gateway, which could attain the necessary message service objects, and then the messages will be created and sent out. For ensuring the security of messages transmission, all the messages will be buffered by the mobile gateway, finally all the buffered messages will be sent to the JMS provider.

Receiving the message composes of synchronous receiver and asynchronous receiver. Asynchronous receiver includes online and offline. To synchronous receiver and asynchronous online receiver the application of the client has to be active while synchronous offline receiver doesn't.

As the synchronous receiver, after the client opens a dialog the obstructive operation of synchronous message receiving maybe be implemented; Through registering an event listener asynchronous online receiver could be constructed, whereas asynchronous offline receiving, the message client firstly needs to register in the MIDP push registry, the mobile gateway will take the messages into the cache after it receives the messages from the JMS provider, at the same time it will send a MIDP push to the mobile custom to alert the new arriving messages.

4. THE DESIGEN OF MOBILE JMS MESSAGE SYSTEM

For realizing the design objective, the mobile message system composes of three parts which can be shown in Fig. 2, public ORB running library in the object communication layer, which is the lightweight ORB communication layer, JMS message gateway of the message service layer and JMS message agent of the mobile client. JMS message gateway locates in the ORB running library and provides a series of remote JMS service objects, for instance, session, topic, queue and so on, furthermore, it still has the functions of message cache, user administration, gateway configuration and so on. The mobile client message agent also locates in the ORB running library, and encapsulates a series of defined mobile JMS API for the mobile JMS client, at the same time it also provides local messages caching, asynchronous message listening, service quality and fault tolerant and so on.

JMS message agent is JMS API encapsulation of mobile client, which unifies the service interface. The services of JMS message agent are to send and receive messages, cache the local messages and push message be listened. JMS API encapsulation depends on a series of gateway service objects which implement the mobile version of the JMS service interfaces. All the mobile JMS operation will submit to the gateway service object execution by stub. To be emphasized, the standard JMS API is an interface framework in J2EE platform, and all the standard JMS APIS are applied in the mobile clients won't be accepted and necessary. Therefore, part of them should be modified or omitted for accessing the JMS basically.

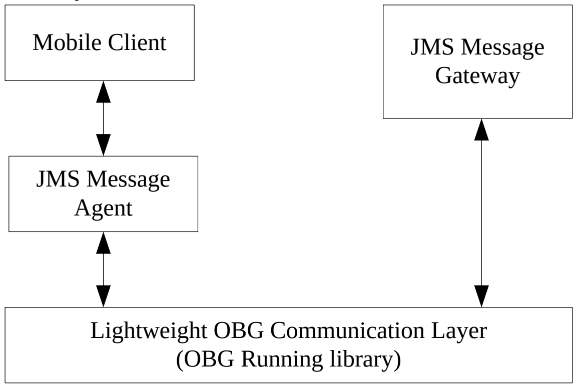


Figure 2. Distribution mobile JMS message system based on ORB

JMS message gateway is the core service part of distributed message system. It provides a series of remote object interface, client with the administration of client dialogue, user message cache, JMS message sending with listening and some other background services. The remote object interface is the operation interface of mobile gateway and client, mobile JMS message agent will accomplish all the jobs depending on these message agents, whereas the service object serves the remote object interface in the mobile gateway.

For the transfer of all the messages accurately, the mobile message system adopts persistent data storage technology to save the data. The advantage of persistent data storage is to ensure that the messages will not lose when the system is breakdown and the messages will be transferred continually as the gateway is recovered. The services of persistent data storage come from external RDBMS system, which doesn't need the corresponding RDBMS system.

The communication layer in the mobile message system is designed by ORB technology, which is the CORBA norm of Java system, and can establish a relationship between the client / server object. Through ORB the client could call the methods of server objects in the same computer or the same network[3,4]. ORB will support this kind of calling and find an object to construct the request, and then send the parameters to those objects, calling the method, finally return the result. ORB could accomplish inter-operate of various computers' application and multi-objects seamless connection in the distributed environment[5,6]. In this study ORB will serve as the communication layer for the transfer tasks of mobile JMS. Lightweight ORB system technology is the application of ORB in the J2ME platform. Normally ORB is applied in the desktop host and server system which has good computing ability and strong support from the other software system[7]. Whereas the J2ME platform which is restricted by hardware and software system ORB couldn't be adopted in the basic ORB of the desktop system, modification and cutting are essential to the system function.

4.1 Construction of Message Agent

In Fig.3 the construction of message agent is described. The entire message APIs are provided by the class WJMSsession. The class WJMSSession uses the method createSession to create the dialogues of the same message gateway; there are three parameters on the method createSession of WJMSSession, user id, password and address of gateway. Thereinto, the format of gateway address is decided by the type of network protocol.

After creating the dialogue, the message Producer and message consumer, which is related with the dialogue, will be built. With the difference in standard more than one message

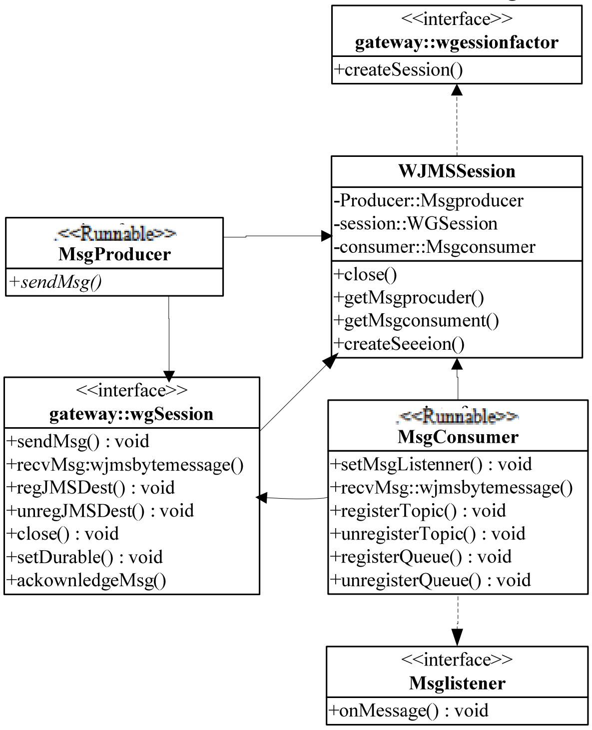


Figure 3. Construction of mobile JMS message agent

producers and message consumers could be built in the dialogue of standard JMS, nevertheless, in this study a dialogue only connects one message producer and one message consumer for simplifying the systematic designing. But some message producer or message consumer may send or receive the messages toward more than one gateway address.

For the transfer of messages reliably, there is a local message queue in the message producer, and the runnable interfaces are executed in the independent thread. The method sendMsg of Message Producer only places the messages into the queue, the thread of message producer will get the item of the queue and send the messages. It’s just for avoiding the sudden break as the transfer process, for instance, network disconnection or other possible failures, which will lead to the lost of data. If some messages are failed, message producer will go on sending it till it is sent successfully without removing from the queue. Message consumer fulfils the receiving job to the message client, which receiving the messages by synchronous / asynchronous and off-line asynchronously. The method registerTopic and unregisterTopic of message consumer is to register and remove the message topic to the message client, the method registerQueue and unregisterQueue register and remove the message queue to the message custom.

For accomplishing receiving the online asynchronous message, the runnable interface has to be built in message consumer, the method setMsgListener registers message event listening. The Message consumer’s thread will send the new messages to the gateway constantly, once the messages are arriving the message event listening will be called.

In order to receive the off-line asynchronous message, MsgConsumer provided the method setDurable. It is registered in the local push registry, the gateway will be as a drable client to make the gateway be able to listen the message to the client.

4.2 Construction of Message Gateway

Figure 4 describes the construction of mobile JMS message gateway.

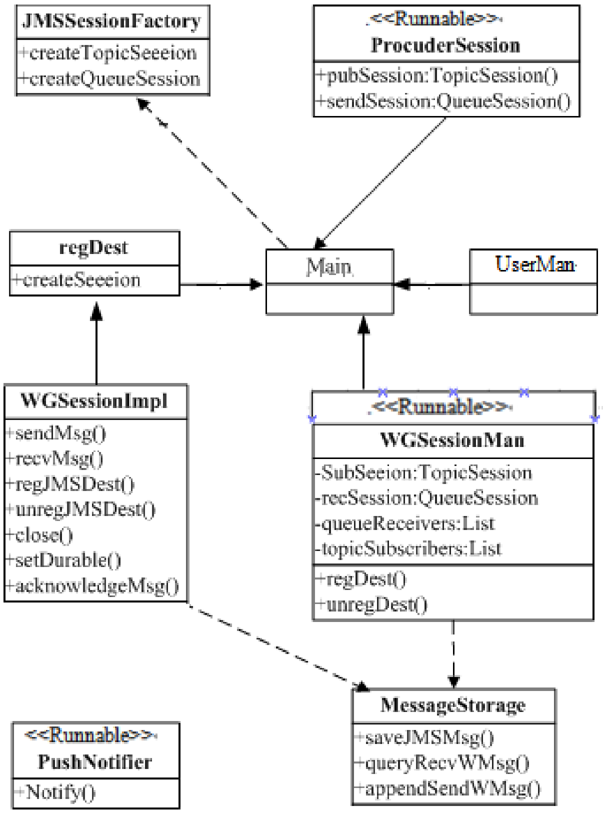


Figure 4. Construction of mobile JMS message gateway

The class WGSessionMan is to administrate all the dialogues, including creating and revoking the dialogues for the custom and monitoring the overtime and invalid dialogues. The object interface WGSessionFactory of WGSessionMan could be used in the remote interface by the client. A series of session objects which named WGSessionImpl are managed by WGSessionMan, and the object interface inherits from WGSessionImpl, a series of JMS message API methods are encapsulated, involving in sending and receiving the messages, initiating and clearing the message address, signing in a label to the persistence customer.

The functions that manage the mobile clients are built by the class UserMan, add, deletes and modifies the users, and some other usual operation.

The class MessageStrorage provides the access ability to saving persistence messages. It includes saving the JMS messages by classification, adding outgoing message to the cache appendSendWMsg, obtaining the client messages by client id and the address of the message, which is the method queryRecvWMsg.

For the sending and receiving of JMS messages, the gateway mainly depends on three objects, which are JMSSessionFactoy. ConsumerSession and ProducerSession.
JMSSessionFactoy will provide two kinds of connecting type communicated with the background, involving in topic

connection and queue connection, which extend into the two message dialogues, Topicsession and QueueSession. The two dialogues could be called by the object ConsumerSession of gateway and ProducerSession.

ConsumerSession is to receive the JMS messages. ConsumerSession has an interface runnable, which runs in the independent thread. It will create more than one QueueReceiver and TopicSubscriber by the subscribed messaging address in the system and save them in the queue. The thread ConsumerSession will proceed those QueueReceivers and TopicSubscriber, save them into the received queue of the custom after receiving JMS messages.

ProducerSession is to send the messages from the custom. ProducerSession has an interface Runnable, which could run in the independent thread. It proceeds to read and send the messages from the message queues into JMS Provider.

Any operation in JMSSessionFactoy, ConsumerSession and ProducerSession must follow as the class MessageStrorage.

Finally, to the permanent push message notice will be sent from PushNotifier. PushNotifier is also the interface Runnable, therefore it could run in the independent thread. The method saveJMSMsg in the MessageStrorage triggers the operation notify in PushNotifier.

5. CONCLUSIONS

Through adopting a kind of high-level and abstractive call method In this study, which adds an OO communication service layer between mobile JMS API and mobile network API, without directly relation to the network API. There are two advantages on abstracting the public and shared functions to design and modify independently. One is to facilitate the design of the mobile JMS API libraries and avoid the complexity of the mobile JMS protocol stack between clients and service; another is to improve and expend the functions and scope of J2ME application.

REFERENCES

[1] Ulbert Attila. Object request brokers in mobile computing. Wireless Personal Communications, 2003,24(4):507～526

[2] Gupta Anil, Kar Subrat. The common object request broker architecture (CORBA) and its notification service,2002,19(1):31～45

[3] Vinoski S. CORBA : Integrating Diverse Applications withinDist ributed Heterogeneous Environment s [J] . Communica2tions Magazine , IEEE ,1997 ,35 (2) :46255.

[4] Czajkowski G,Von Eicken T. J Res :A Resource Accountinginterface for J ava [J] . Proceedings of the 13th ACM SIGP2LAN Conference on Object2oriented Programming ,Systems ,Languages ,and Applications. 1998 :21235.

[5] TANG Han, SHI Haoshan. Middleware Analysis and Design of Wireless Middleware. Computer Engineering ,2004,30(20):121～123

[6] Deng Lihua Huang Hua. The Discussion on Characteristics of CORBA Notification Service, Computer Applications and Software,2003,20(7):77～79

[7] MU Ying, LI Zhongliang, WANG Peng, HUANG Meifeng. WU Rongquan. Design and Implementation of Agent-based Message Oriented Middleware. Computer Engineering,2004,30(22):91～92,151

Ou Yang Quan 1994.06 Graduated from HuaZhong University of Science and Technology. 2007.06 Graduated from Wuhan University of Science and Technology for master degree of computer science. Since 1994, He has addressed in research and education of computer network and database at school of mathematics and computer science in JiangHan university mainly . He has published a books, over 10 Journal papers.

RESEARCH ON DATA EXTRACTION OF LANDSLIDE STABILITY EVALUATION SYSTEM UNDER DISTRIBUTED MULTI-TIER DATA ENVIRONMENT*

Yuanyou Xia, Lingwei Liu
School of Civil Engineering and Architecture, Wuhan University of Technology
Wuhan, Hubei 430070, P.R.China
Email:liulingwei_2005@126.com

ABSTRACT

Stability evaluation of landslide is an important research content in landslide hazards prevention. As the development of database technology, distributed and multi-tier databases have been becoming the main data source for the landslide stability evaluation system. Data Extracting is the key issue for landslide stability evaluation system under distributed multi-tier data environment. In this paper a data extraction engine is designed and implemented for extracting valid data from the distributed database of bottom layer, in order to meet the demands of data extracting of the landslide stability evaluation system under the distributed multi-tier data environment. The engine achieves the purpose of data extracting by data searching, clustering analysis and data uniform handling. It is designed in a hierarchical way with a simple structure, so it is easy to realize. The landslide stability evaluation system on the base of this engine for Three Gorges Reservoir Area is developed successfully and the execution results prove that the data extracted by the data extraction engine discussed in the paper can meet the information requirements of the landslide stability evaluation.

Keywords: Landslide hazard, Distributed database, Landslide stability evaluation, Data extraction, Clustering analysis

1. INTRODUCTION

As a kind of geological disasters, landslide hazards bring a huge threat to human life and properties. So the prevention of landslide hazards is an important research content in the field of geological disasters prevention. Landslide stability evaluation , which can provide Theoretical basis for the prevention of landslide hazards and help experts or decision-makers to make a right decision, is a process of numerical analysis based on the lithology and working condition of the landslide[1,2]. Software for stability evaluation of landslides is another application of Information Technology in the field of prevention of geological disasters[3].In the current landslide stability software, the information required by the calculation is either input by the users directly[4], or from a single data source[5]. The way of inputting data directly by users can ensure the integrity of the data, but the users must be of high professional quality in order to operation the software and this is not conducive to the promotion of stability evaluation software[6]. The way of obtaining data from single data source is difficult to ensure the integrity of the data and always needs auxiliary input by the users. In recent years, as the development of distributed database technology[7,8], the information used in landslide stability evaluation is developing in a distributed and multi-tier trend, and the information is becoming more and more comprehensive. So there is a growing demand of landslide stability evaluation under the distributed multi-tier data environment. Under this data environment, the data types are various and the structure of data level is complex, for example, there are not only inherent attribute data, but also monitoring data and geographic data. Because the collection source of the data is uncertain, these data are always repetitive, noisy, fuzzy and stochastic, and how to extract valid data which can be used in stability evaluation system is a key problem for the application under the distributed multi-tier data environment. It will be of great practical significance to design a data extraction engine for landslide stability evaluation system under this data environment.

Taking the stability evaluation of landslides in Three Gorges Reservoir Area as the application background, the data extraction theory and method under the distributed multi-tier data environment are studied in the paper. Then a data extraction engine is proposed in this paper. This engine can analyze data which are massive, noisy, fuzzy and stochastic and extract useful data for the landslide stability evaluation system. The engine can provide better data source for the application system, improve the operating convenience for the users and reduce the workload of the operators, and this is conducive to the promotion of the application software. The most importance of all, because the landslide stability system can get comprehensive and accurate data, the accuracy of the system is improved a lot.

Based on the data extraction engine proposed in the paper, a landslide stability evaluation system for Three Gorges Reservoir Area is developed. The experimental results show that the engine is reliable and effective, designed simply and easy to implement, and it provides a feasible solution for the data extraction for the landslide stability evaluation system under distributed multi-tier data environment.

The remaining of this paper is structured as follows: section 2 analyzes the data environment and requirement for landslide stability evaluation system of Three Gorges Reservoir Area, section3 presents the model, the workflow and the key algorithm of the data extraction engine proposed in the paper, and section 4 discusses the construction of the landslide stability evaluation system based on the engine and testify the effectiveness of the engine. Conclusions, observations and future improvements are then given at the end.

*Special fund for Three Gorges Reservoir Area of the Ministry of Land and Resources P.R.C: the early warning and command of the geological disasters in Three Gorges Reservoir Area (SXJC-3ZH1B2).

2. DATA ANALYSIS OF THE SYSTEM

2.1 Data Environment of The System

Comprehensive databases have been established for the disasters in Three Gorges Reservoir Area, on the basis of geological survey and monitoring network. These databases can provide data supporting for various application systems, and the landslide stability evaluation system is a subsystem developed on the basis of this data environment for the early warning and command of the geological disasters in Three Gorges Reservoir Area. The databases of geological disasters in Three Gorges Reservoir Area are designed in distributed and multi-tier way.

The databases are divided into 2 tiers: one is source data tier which is on the bottom, the other is handling tier which is above the source data tier. The databases in source data tier mainly store basic and raw data, including basic geographic data, basic geological data, geological disaster information, humanities economic information, weather, earthquakes and other spatial data, high slope information, investigation information of Reservoir banks, survey data and monitoring data. The databases in this tier aren’t open to the application systems. The application systems can only get data from the databases in handling tier. The data in source data tier databases must be converted and processed before they are stored into the databases in handling tier. There are three types of database in this tier: spatial database, profession attributes database and management database. Data operation of application systems can only be in this tier. The database hierarchy is shown in Figure 1.

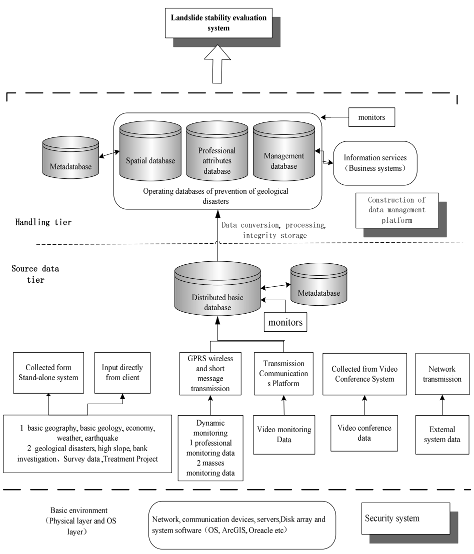


Figure 1. Database Hierarchy

In source data tier, because of the variety of the data collecting approaches, the databases are distributed. The databases are composed of many distributed and collaborative database server nodes. These nodes provide information services for the client on the network collaboratively. Cooperative database server nodes are the core of the database system. They are not only independent body of computer and data storage and management body, but also execution body of calculating and analyzing tasks. Data, including landslide spatial information, attributes information, monitoring information and multimedia information etc, are stored in these database server nodes.

2.2 Data Need Analysis of The System

The landslide stability evaluation system is to use the limit equilibrium method to analyze the stability of landslides on the basis of the geological databases of Three Gorges Reservoir Area which can provide the necessary information for the system. A coefficient returned by the system after the numerical analysis is on behalf of the stability of the landslide. The data used in the landslide stability system include landslide spatial data, such as slope line, water line, rock layer, slip plane etc, attributes data, such as cohesion of the slip plane, friction angle of the slip plane, natural unit weight of the rock, dry unit weight of the rock etc, and monitoring data, such as slip displacement, change of water line, change of reservoir water level etc. Figure 2 shows the data needs model of the system.

The data used in the landslide stability evaluation system is a subset of the data in geological databases of Three Gorges Reservoir Area. And because the collecting sources of these data are various, these data in databases of source data tier are always massive, noisy, fuzzy and stochastic and can’t be used directly. So there must be a mechanism to identify, analyze and process these data and make them be useful for the landslide stability system. Because only the databases in the handling tier can be operated by the application systems, the problem becomes that how to extract useful data from source data tier to handling tier. In this paper a data extraction engine is proposed, which can solve that problem.

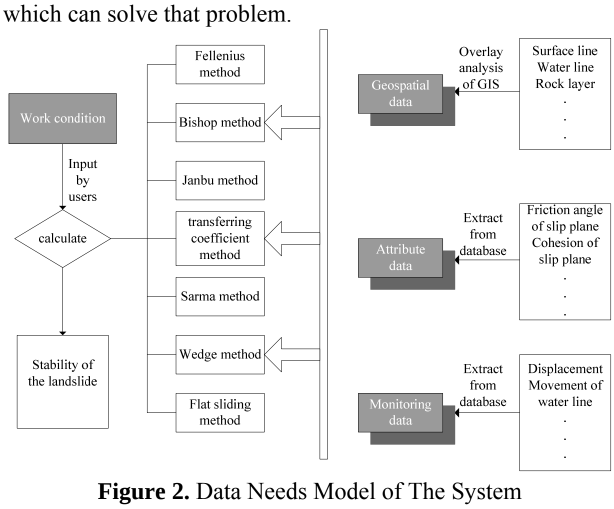


Figure 2. Data Needs Model of The System

3. DATA EXTRACTION ENGINE UNDER DISTRIBUTED MULTI-TIER DATA ENVIRONMENT

3.1 Model of The Data Extraction Engine

The data extraction engine, presented in this paper, incorporates multiple technologies used to extract data and make them available as needed. Figure 3 gives a high level view of the data extraction engine architecture.

This engine is a platform and a set of tools which can retrieve, integrate, unify and analyze the data in bottom tier databases via various processing algorithms as required by application. The engine, developed with J2EE, incorporates web service and mobile agent technologies, can provide data services for the landslide stability evaluation system.

Elements of the data extraction engine adopted include the following entities: Data Acquisition Unit (DAU), Retrieve Unit (RU), Data Flow Processor (DFP), Data Transfer Layer (DTL), Message Layer (ML), Data Storage Unit (DSU), Security Layer (SL) and Control Layer (CL).

DAU (Data Acquisition Unit): is an entity that is responsible for the data acquisition. There is a cache in this entity and the data in the bottom tier databases which are distributed can be transferred to this cache with the help of the DTL.

Retrieve Unit (RU): is used to retrieve the data in the databases of source data tier. Because there is a large amount of data in the databases and the useful data must be retrieved efficiently, the SU must adopt optimum algorithms and strategies. And because both attributes data and spatial data are used in the landslide stability system, all of them must be taken into account in the process of data retrieval. The retrieval of the attribute data is relatively simple, and a strategy for the spatial data is discussed here. To accommodate queries of data holdings, a catalog data model with a common spatial frame of reference is generated and stored on XML files following standards used in Geo-Spatial data manipulation. The geospatial data which is used most frequently is stored in these XML files, and they are maintained as part of the data in handling tier databases. The catalog is in a format that it efficiently handles geospatial queries. A parser and extractor are used for loading and extracting records. Geospatial coverage for cataloged datasets is created as polygonal areas, lines, or multipoint objects. There is a spatial data calculator responsible for the spatial analysis. When RU receives a request for spatial data retrieval, it retrieves data in the geospatial coverage created from catalog data first, if the data can’t be find there and then it retrieves data in the records of the databases. Because geospatial retrievals take a lot of time， this strategy improves the efficiency of geospatial data retrieval greatly to adopt this strategy.

DTL (Data Transfer Layer): The internal functionality to transfer and convert data from source data database to DAU is handled by the DTL.

DFP (Data Flow Processor): As mentioned previously, the data in the geological disasters databases of bottom tier is always noisy and fuzzy, so they must be processed before they are stored to the handling tier databases. DFP is responsible for the data processing. For example, the data must be in a uniform format, including uniform data unit, uniform data precision, and DFP can convert them to the required format. DFP also provides functions such as calculating and validity check based on the requirement of the landslide stability evaluation system. If the data can’t be converted into the format which is required by the system, the data will be discarded by DFP.

DSU (Data Storage Unit): is responsible for storing data processed by DFP to the handling tier databases. In the application background of landslide stability system, the data only need to be stored in the databases, but in other applications, the data may be required to store in TXT, XML or HTML files.

ML (Message Layer): is used to listen in internal and external events of the engine and distribute messages. ML also is responsible for the correspondence between different units or layers.

SL (Security Layer): the data processed and extracted by the engine must be ensured to be safe and of integrity. SL checks the security of the data via testifying the source of the data. SL is also responsible for verifying the legitimacy of the request.

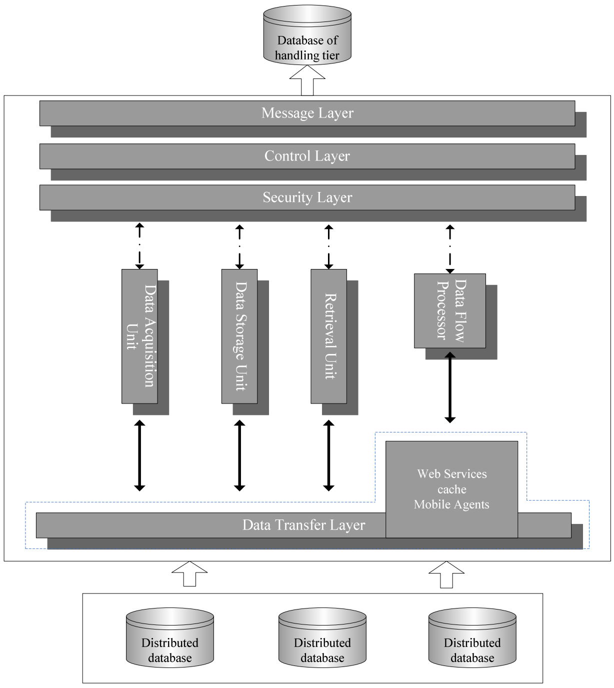


Figure 3. Architecture of The Data Extraction Engine

CL (Control Layer): is an important layer of the engine. CL is responsible for the scheduling and encapsulating of the resources, the coordination of different parts of the engine and the exception handling. Because this engine works under distributed multi-tier data environment, the resources of different nodes, including computing resources, network resources and storage resources, must be encapsulated and integrated. In order to make every part work cooperatively, CL controls them by sending commands, for example, it can send a command to an agent to let it fetch data in a node. Commands are sent by ML, instead of by CL itself, in order to reduce the burden of CL. The commands are sent to ML in the form of messages firstly, and then parsed by ML and sent to the target unit or layer. Because the data is always chaotic, any exceptions may appear in the process of data extraction, and CL has the power to handle these exceptions, for example terminate the task or restart the task.

3.2 Workflow of The Data Extraction Engine

When ML has heard the data extraction request outside, the request is sent to SL to verify the legitimacy of the request and if it is accepted by SL then it is sent to CL. Then CL begins to schedule, encapsulate and distribute resources. After that CL sends data acquisition command to DAU. And then the command is parsed and query conditions are constructed by DAU. Then DAU sends data transfer request to CL and CL commands the DTL to transfer data from databases of sourced data tier to DAU's cache with the help of SU, web services and mobile agents in DTL are the executors of data transfer. ML will send messages to CL When data in DAU's cache reaches a certain amount, and CL will send data handling command to DFP. Then DTL transfers the data in DAU's cache to DFP. And then the data is transferred back to DAU's cache after unified handling, denoising, clustering analysis and amendment process. Then DAU sends the data to SL and the data is stored in databases of handling tier by DSU after SL checks the safety of the data. That is a complete process of the data extraction.

3.3 Clustering Analysis Algorithm of The Data Extraction Engine

One function of the data extraction engine is to discover useful and valid data from mass data, and clustering analysis algorithm is used in DFP. Clustering algorithms have been extensively used in many application fields. In recent years ,a lot of efficient clustering analysis algorithms have been proposed by many experts, such as DENCLUE[9] based on kernel density estimation, CN2[10] and CL2[11], Nonlinear algorithms based on neural networks[12]. However, due to the limitations of their own theories, these algorithms, or difficult to determine the parameters, or difficult to implementation, or the analysis results are not satisfied, are not appropriate to use in this paper.

Instead, a clustering analysis algorithm based on ELECTRE[13-17] is proposed in this paper. And it is described in detail below.

Step 1: Input a set of decision rules X_D and objects Y_T to be classified.

Step 2: Fetch an object y_i from Y_T and Estimate the resemblance between the object and every decision rule. The purpose of Step1, is to identify the resemblance between object $y_i \in Y_T$ and each of the decision rules $x_j \in X_D$ for every attribute a_k that evaluates both y_i and x_j. Calculation formula is as follows:

$$D(x_{j,a_k}, y_{i,a_k}) = \frac{m_{x_{j,a_k}} + m_{y_{i,a_k}}}{m_{x_{j,a_k}} . m_{y_{i,a_k}}} . \delta(x_{j,a_k}, y_{i,a_k}) \tag{1}$$

where x_{j,a_k} is the value of decision rule x_j on a_k, y_{i,a_k} is the value of object y_i on a_k, $m_{x_{j,a_k}}$ is the number of times x_{j,a_k} appears in the set of the decision rules on a_k, $m_{y_{i,a_k}}$ is the number of times y_{i,a_k} appears in the set of the objects on a_k. $D(x_{j,a_k}, y_{i,a_k})$ is based on the dissimilarity measure of Huang [18].

$$\delta(x_{j,a_k}, y_{i,a_k}) = \begin{cases} 0, & if \quad x_{j,a_k} = y_{i,a_k} \\ 1, & if \quad x_{j,a_k} \neq y_{i,a_k} \end{cases} \tag{2}$$

Step 3: Construct concordance coalition $Con(y_i, x_j)$, discordance coalition $Dis(y_i, x_j)$ and zero coalition $ConZero(y_i, x_j)$ based on the resemblance calculated in step 2.First, parameter q_{i,a_k} should be calculated,

$$q_{i,a_k} = Min\{D(x_{j,a_k}, y_{i,a_k}) : D(x_{j,a_k}, y_{i,a_k}) \neq 0, j = 1,2...n, i = 1,2,...m\} \tag{3}$$

The attributes that satisfy relation $D(x_{j,a_k}, y_{i,a_k}) \le q_{i,a_k}$, belong to a concordance coalition $Con(y_i, x_j)$. And attributes that satisfy relation $D(x_{j,a_k}, y_{i,a_k}) > q_{i,a_k}$, belong to a discordance coalition $Dis(y_i, x_j)$ .And attributes that satisfy relation $D(x_{j,a_k}, y_{i,a_k}) \le q_{i,a_k}$ and $x_{j,a_k} = y_{i,a_k}$, belong to a zero coalition $ConZero(y_i, x_j)$.

Step 4: Calculate the concordance indices CI and concordance thresholds between the object and every decision rule, and then select rules which may take part in the classification process. The concordance indices for the comparisons of object y_i with each of the decision rules x_j are calculated as follows:

$$CI(x_j, y_i) = \left[\left(\frac{\sum_{c=1}^{u_j} W_c}{\sum_{h=1}^{e_j} W_h} \right) + bonus(j). \left(\frac{\sum_{z=1}^{p_j} W_z}{\sum_{c=1}^{u_j} W_c} \right) \right] \tag{4}$$

Where W_h, h=1,..., e_j is the weight of attributes $a_h \in$F that belong to the concordance and the discordance coalition, that is, the attributes of x_j that take part in the classification process, and ej is number of these attributes. W_c, c= 1,..., u_j is the weight of attributes $a_c \in$F that belong to the concordance coalition, that is they belong to $Con(y_i, x_j)$ and u_j is the number of these attributes. W_z, z= 1,..., p_j is the weight of attributes $a_z \in$F that belong to the zero coalition, and p_j is the number of these attributes. Parameter *bonus*(j) constitutes an essential factor in order to properly sort in descending order,

$$bonus(j) = \begin{cases} p_j.b, & ConZero(y_i, x_j) = Con(y_i, x_j) \\ p_j, & ConZero(y_i, x_j) \subset Con(y_i, x_j) \\ 1, & ConZero(y_i, x_j) = \Phi \end{cases} \tag{5}$$

The concordance thresholds $CT(x_j, y_i)$, corresponding to the concordance indices $CI(x_j, y_i)$, are calculated as follows:

$$CT(x_j, y_i) = m(i) + bonus(j). \left(\frac{\sum_{z=1}^{p_j} W_z}{\sum_{c=1}^{u_j} W_c} \right) \tag{6}$$

$m(i)$, represents a lower limit of the sum of the weights of the attributes that belong to the concordance coalition, is calculated as follows:

$$m(i) = \begin{cases} 0.7, & N_z(i) = 0 \\ 0.85, & N_z(i) = 1 \\ 1, & N_z(i) > 1 \end{cases} \tag{7}$$

If CI ≥CT, Decision rule x_j , will take part in the classification process, is add to set *PR*.

Step 5: The classification process will continue for those decision rules not excluded from Step 4.

On the premise of $Cov(x_t) \neq 0$ and $Acc(x_t) \neq 0$, if $P_t \le L_1 \cdot b$ and $CovAcc(x_t) \ge L_3$ or $Acc(x_t) \ge L_2$, x_t will be excluded from *RP*.

Where P_t is the number of attributes that belong to $ConZero(y_i, x_j)$ and b is the total number of attributes of the decision rules set taking part in the classification process as a whole. $Cov(x_t)$ is the coverage of decision rule x_t , t= 1, ..., d_1, where d_1 is the number of the decision rules not excluded from

Step4, and $Acc(x_t)$ is the corresponding accuracy. The coverage and accuracy of a decision rule are important evaluation measures [19]. $CovAcc(x_t)$ is an increasing function of both the accuracy and coverage of decision rule x_t. it is calculated as follows:

$$CovAcc(t) = Acc(x_t).(1 + Cov(x_t)) \quad (8)$$

L_1, L_2, L_3 are calculated below,

$$L_1 = \begin{cases} 1/2 & A_V > 1 \\ 1/3, & 1/3 \le A_V \le 1 \\ 0, & 0 \le A_V \le 1/3 \end{cases} \quad (9)$$

$$A_v = \frac{\sum_{t=1}^{d1} P_t}{d1} \quad (10)$$

L_2 is a constant 0.95 and L_3 is calculated as follows:

$$L_3 = \begin{cases} CovAcc(t)[N_{dis}/2], A_v > 1 \\ CovAcc(t)[2*N_{dis}/3], 1/3 \le A_v \le 1 \\ CovAcc(t)[N_{dis}/2], 1 \le A_v < 1/3 \end{cases} \quad (11)$$

Step 6: For the rules in RP, if $D(x^*_{t,a_k}, y_{i,a_k}) > U_{i,a_k}$, object y_i is classified to the class of x^*_{t,a_k}.

Step 7: If the object isn't classified to any class, it will be classified to a default class. And if all the objects in object sets are classified to a class, the process will be terminated, if not, the next object which is not classified in object sets is fetched out and jumps to step 4.

4. LNADSLIDE STABILITY EVALUATION SYSTEM BASED ON THE ENGINE

4.1 Implementation of the System

The data required for the landslide stability evaluation system can be extracted by the engine from source data tier databases to handling tier databases, and these data can be used by the landslide stability system directly. In this paper, an application system used for the stability evaluation of landslides in Three Gorges Reservoir Area is constructed by JSP technology. The system adopts popular three-tier B / S model, web performance layer, application logic layer and data services layer. Browsers are used as the user interfaces to submit application request for the users in web performance layer. Application logic layer is the core of the whole system, which is a series of component modules and algorithms. Data services layer is responsible for data exchange between Application logic layer and database and provides data services for the application. Oracle 10g is used for the handling tier database software and database connectivity is based on the JDBC.

4.2 Running Instance

Calculation instance is randomly selected from the Three Gorges reservoir area. It is a rock slide, lithology structure of which is complex , the required data for calculation is relatively large. All the data, including spatial data of profile, lithology information and monitoring data ect, are extracted from the source tier databases by the data extraction engine proposed in the paper. The stability calculated by the system is consistent with the real state of this landslide. This proves the data extraction engine works successfully. Figure 4 shows the running result of the system.

A large number of similar experiments have been done in the paper, results of which are all satisfying. These experiments show that the data extraction engine proposed in the paper is feasible and effective. Operating convenience is improved a lot with the help of this engine, because data need to be input by users is much less.

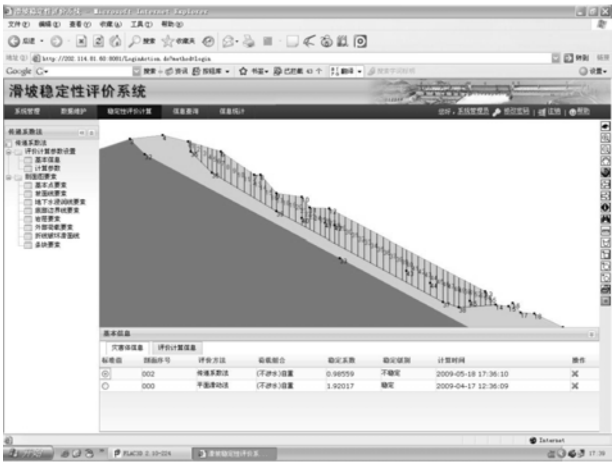


Figure 4. Running Instance

5. CONCLUSIONS

Taking the data transfer, message delivery, data processing, data retrieval, security and global control into account, based on the stability evaluation of landslides in Three Gorges Reservoir Area, a data extraction engine is proposed to meet the requirements for the application under distributed multi-tier data environment in the paper. A number of experiments show that the engine is efficient. As distributed databases are used more and more extensively in recent years, research fruits of this paper also provide a solution for the similar applications.

In the field of data extraction, some exploring work has done in this paper, and this work will be continued. The next work will focus on: 1) Design of more optimized data retrieval algorithm; 2) The application of distributed computing technology based on mobile agent in data extraction under distributed multi-tier data environment.

REFERENCES

[1] Jianzheng Pan. Buildings stability against sliding and landslide analysis. Beijing:China WaterPower Press，1980(in chinese)

[2] Y.M.Cheng,T.Lansivaara,W.B.Wei.Two-dimensional slope stability analysis by limit equilibrium and strength reduction methods[J].Computers and Geotechnics,2007,34:137～150

[3] Donald, I. B. And Giam, P. The ACADS slope stability programs review. Proc. 6Th International Symposium on Landslides. 1992. Vol. 3. 1665～1670

[4] Luyu Zhang. Development of software ZSlope for slope stability analysis[J].Chinese journal of rock mechanics and engineering, 2004,23(16)：2830～2835

[5] Yusheng Wu,Shenghong Chen. Study on the software of limit equilibrium method for slope stability analysis[J]. Journal of Hydroelectric Engineering,2005,24(6):35～39

[6] Luyu Zhang,Xiaoxiu Ouyang,Yingren Zheng.Problems and thoughts of development of slope stability analysis software in china[J]. Chinese journal of rock mechanics and engineering, 2003,22(1)：166～169

[7] Nicola Pio Belfiore.Distributed Databases for the development of Mechanisms Topology[J].Mechanism and Machine Theory,2000,35:1727～1744

[8] R.Warren,A.E.Solomonides,C.del Frate etc. Mammo Grid— a prototype distributed mammographic database for Europe[J]. Clinical Radiology, 2007, 62:1044～1051

[9] Xiao gaoyu,Yin jian.A new clustering algorithm based on knn and denclue. In:Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, 2005.p:2033-2038

[10] Clark P, Niblett T. The CN2 induction algorithm. Machine Learning 1989,3(4): 261–83.

[11] Boutsinas B, Antzoulatos G, Alevizos P. A novel classification algorithm based on clustering. In: First international conference “From scientific computing to computational engineering”, Athens, Greece, 2004.

[12] Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: explorations in the microstructure of cognition. Cambridge: MIT Press; 1986. p. 318–63.

[13] Bouyssou D. Outranking methods. In: Floudas CA, Pardalos PM, editors. Encyclopedia of optimization, vol. 4. Dordrecht: Kluwer; 2001. p. 249–55.

[14] Figueira J, Mousseau V, Roy B. Electre methods. In: Figueira J, Greco S, Ehrogott M, editors. Multiple criteria decision analysis: state of the art surveys. New York: Springer; 2005. p. 133–53.

[15] Pirlot M. A common framework for describing some outranking methods. Journal of Multi-Criteria Decision Analysis 1997,6:86–92.

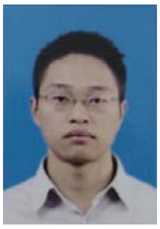
[16] Roy B. Multicriteria methodology for decision aiding. Dordrecht: Kluwer Academic Publishers; 1996.

[17] Roy B. The outranking approach and the foundations of electre methods. Theory and Decision 1991,31:49–73.

[18] HuangZ. Extensions to the k-means algorithm for clustering large datasets with categorical values.Data mining and Knowledge Discovery,1998,2:59–77.

[19] Tsumoto S. Characteristics of accuracy and coverage in rule induction. In: Lecture notes in computer science. 2003. p. 237–44.

Yuanyou Xia is a Full Professor of School of Civil Engineering and Architecture, Wuhan University of Technology. He received his phD in Mining Engineering from the Beijing University of Science and Technology. His research interests include the stability of geotechnical structures, landslide hazard assessment and prevention decision-making, as well as landslide stability Intelligent Analysis.



Lingwei Liu is a phD candidate in Geotechnical Engineering at Wuhan University of Technology. He received his master degree in computer application technology from Wuhan University of Technology. His research interests lie in software engineering, GIS, application of Information Technology in Geotechnical Engineering.

LARGE-SCALE PARALLEL SIMULATION OF HIGH-DIMENSIONAL AMERICAN OPTION PRICING*

CHANG Hong-xu[1, 2], LU Zhong-hua[1], and CHI Xue-bin[1]
[1] Supercomputing Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China
[2] Graduate University of Chinese Academy of Sciences
Beijing, 100190, China
Email: {hongxu, zhlu, chi}@sccas.cn

ABSTRACT

High-dimensional American option pricing is computational challenging in both theory and practice. We use stochastic mesh method combined with performance enhancement policy by bias reduction to solve this practical problem in classic Black-Scholes framework. We effectively parallelize this algorithm, use MPI for implementation and execute large-scale numerical experiments on heterogeneous supercomputer DeepComp7000. Numerical results of parallel simulation demonstrate that the parallel simulation has good scalability in different parallel environments, and can obtain higher speedups for large-scale numerical simulation. The convergence performance is also demonstrated. The estimated option value converges with the increase of mesh size; when using smaller mesh size, the stochastic mesh method with bias reduction can underestimate the true American option value.

Keywords: High-dimensional, American Option, Option Pricing, Large-scale, Parallel Simulation.

1. INTRODUCTION

In the inspiring field of computational finance, high-dimensional American option pricing poses great challenges. The general method for this intractable problem requires large computational effort, and can't satisfy the low-latency need of the practitioners [5][6][9]. Popular Monte Carlo simulation-based method can easily dispose of high-dimensional problems. The convergence rate doesn't depend on the dimensionality of the specific problem on hands. This method is quite flexible to different option payoff types and can be easily implemented. Large-scale parallel Monte Carlo simulation can efficiently overcome the computational challenges in high-dimensional American option pricing.

Several methods have been proposed to deal with this problem. Longstaff and Schwartz [8] suggested the regression-based method to estimate the continuation values during the simulated paths and then to price the American option. This method estimates the current continuation values via the linear combinations of selected basis functions of current asset price and uses regression techniques to obtain the optimal weight coefficients. It is broadly applicable and its accuracy largely depends on the carefully chosen basis functions. Moreover, this approach can only compute the lower bound of American option value. Broadie and Glasserman [3] constructed random tree approach for American option pricing. The computational effort of this method exponentially increases with the number of exercise opportunities, which makes this approach computational prohibitive for high-dimensional problems.

* This work was funded by National High Technology Research and Development Program of China (863 project) grant 2006AA01A116 and National Natural Science Foundation of China grant 60873113.

Broadie and Glasserman [4][5] later proposed stochastic mesh method for high-dimensional American option pricing. This algorithm firstly constructs the mesh of assets states and then recursively estimates current continuation value by linear combinations of option values at the next exercise date. The weights are computed via transition density function of assets price, unlike the regression approach in [8]. Then the optimal exercise policy can be obtained. The method's computational effort is linear in the number of exercise opportunities and quadratic in mesh size. This approach can estimate both lower and upper bounds of option value and then the confidence interval and point estimates of the option value can be obtained. As for the convergence of this novel method, Avramidis and Matzinger [2] recently derived the asymptotic upper bound on the probability of the estimated error under mild assumptions. Both the estimated error and the probability bound vanish gradually with the increasing of mesh size. The convergence of this algorithm is only related to the mesh size, which suggests that stochastic mesh method should be of great interest of the practitioners. Moreover, some performance enhancement policies for this method are proposed. Broadie and Glasserman [5] used control variate technique, while Avramidis and Hyden [1] suggested the performance enhancement policies using bias reduction or importance sampling. These general policies can obtain more accurate results and less bias than the naïve stochastic mesh method. In this paper, we will focus on parallelizing the stochastic mesh method combined with estimated bias reduction policy, and pricing high-dimensional American option with large-scale parallel simulation.

The rest of the paper is organized as follows: section 2 describes the background of this problem in classic Black-Scholes framework and presents the serial stochastic mesh method combined with bias reduction policy; in section 3, the parallelization strategy is discussed in details; numerical results of large-scale parallel simulation on DeepComp7000 is presented and analyzed in section 4, followed by the conclusion.

2. SERIAL STOCHASTIC MESH METHOD

2.1 Background of High-dimensional American Option Pricing

Consider the American option allowing the holders to exercise at finite times (also called Bermudan option). Suppose that the American option underlies n assets, where n is the dimensionality of the problem; the strike price is K and the maturity T ; price vector of n assets at time t $S_t = (S_t^1, S_t^2, \cdots, S_t^n)$ is a Markov process on R^n with fixed initial value S_0 ; $0 = t_0 < t_1 < \cdots < t_M = T$ are exercise opportunities (also called exercise dates) of the American option and assume that $t_{i+1} - t_i = \Delta t$, $i = 0,1,\cdots,M-1$, we use $t = 0,1,2,\cdots,M$ for short of t_i ; the interest rate r is constant; the ith asset has dividend rate δ_i and volatility σ_i , $i = 1,2,\cdots,n$. Consider that the risk-neutral dynamics of n assets follow multi-dimensional correlated geometric Brown motion, i.e.

$$dS_t^i = (r-\delta_i)S_t^i dt + \sigma_i S_t^i dW_t^i,\ i=1,2,\cdots,n \quad (1)$$

where W_t^i are standard Brownian motion, the covariance between W_t^i and W_t^j $\mathrm{cov}(W_t^i, W_t^j) = \rho_{ij} t$ ($i,j=1,2,\cdots,n$, $i \neq j$), the element of the covariance matrix Σ is $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, $\Sigma = (\Sigma_{ij})_{n\times n} = CC^T$ is the Cholesky decomposition of the covariance matrix Σ.Then the price of the ith asset at time t is $S_t^i = S_0^i \exp\{(r-\delta_i - \frac{1}{2}\sigma_i^2)t + Z_t^i\}$ (2)

where $Z_t = C \cdot RNum$, $RNum$ follows n -dimensional standard normal distribution $N(0,I)$, $i=1,2,\cdots,n$. Moreover, the transition density function of S_{t+1} given S_t is

$$f(S_t, S_{t+1}) = (2\pi\Delta t)^{-\frac{n}{2}} \,|\, C^{-1} \,|\, (\prod_{i=1}^{n} S_{t+1}^i)^{-1} \exp\{-\frac{\| \Theta(S_t, S_{t+1}) \|^2}{2\Delta t}\} \quad (3)$$

where $\Theta(S_t, S_{t+1}) = \left[\log\frac{S_{t+1}^i}{S_t^i} + (\frac{1}{2}\sum_{1\le l\le n} C_{il}^2 - r + \delta_i)\Delta t \right]_{1\le i\le n}$

Let $h(t,S_t)$ be the payoff function of high-dimensional American option at exercise date t. For example, the payoff function of American max call option underlying n assets is $h(t,S_t) = \max\{\max\{S_t^1, S_t^2, \cdots, S_t^n\} - K, 0\}$ and the payoff function of geometric average call option on n assets is $h(t,S_t) = \max\{(S_t^1 S_t^2 \cdots S_t^n)^{\frac{1}{n}} - K, 0\}$.

High-dimensional American option pricing problem can be described in the following different forms:

(1) Optimal Stopping Problem

$$P = \max_{\tau} E[e^{-r\tau} h(\tau, S_\tau)] \quad (4)$$

where $\tau \in \{t_0, t_1, \cdots, t_M\}$ are stopping times.

(2) Stochastic Dynamic Programming Problem

$$\begin{cases} P(M,S_M) = h(M,S_M) \\ P(t,S_t) = \max\{h(t,S_t), C(t,S_t)\} & t = M-1,\cdots,1,0 \end{cases} \quad (5)$$

where $C(t,u) = e^{-r\Delta t} E(P(t+1,S_{t+1}) \mid S_t = u)$ is the continuation value of the American option at exercise date t.

We view the high-dimensional American option pricing problem in the later form and effectively estimate the option value $P(0,S_0)$ using stochastic mesh method.

2.2 Serial Stochastic Mesh Method with Bias Reduction

Based on the risk-neutral dynamics of n assets, we generate the mesh consisting of b independent paths of assets motion, with the ith path $S_t(i)$, $t=1,2,\cdots,M$, $i=1,2,\cdots,b$. We follow the mesh generation method in [5], "forgetting" the path to which each point in the mesh belongs. We call b the mesh size. The high-biased estimator of the option value at exercise date t on the ith path $\hat{P}_H(t,S_t(i))$ can be derived using the dynamic programming procedure

$$\begin{cases} \hat{P}_H(M,S_M(i)) = h(M,S_M(i)) \\ \hat{P}_H(t,S_t(i)) = \max\{h(t,S_t(i)), \hat{C}(t,S_t(i))\} \end{cases} \quad (6)$$

where $\hat{C}(t,S_t(i)) = e^{-r\Delta t}\frac{1}{b}\sum_{j=1}^{b}\hat{P}_H(t+1,S_{t+1}(j))w(t,S_t(i),S_{t+1}(j))$, $w(t,S_t(i),S_{t+1}(j))$ is the weight attached to the arc jointing $S_t(i)$ and $S_{t+1}(j)$, $t=M-1,\cdots,1,0$, $i,j=1,2,\cdots,b$. A better choice of weights $w(t,S_t(i),S_{t+1}(j))$ can be expressed in the for

$$w(t,S_t(i),S_{t+1}(j)) = \frac{f(S_t(i),S_{t+1}(j))}{\frac{1}{b}\sum_{k=1}^{b} f(S_t(k),S_{t+1}(j))} \quad (7)$$

where $f(S_t(i),S_{t+1}(j))$ is the same as in Eq. (3).

To obtain the low-biased estimator $\hat{P}_L(0,S_0)$, we forward simulate $nbMC$ paths of assets dynamics respectively, and derive the option value by using the suboptimal exercise policy derived in the process of estimating $\hat{P}_H(t,S_t(i))$, which determines whether to exercise or continue to hold the option at exercise time t. Replicate the mesh for N times, average the high-biased and low-biased estimators of option value respectively, and then we can construct a $1-\alpha$ (α =5%, 10%, ⋯) level confidence interval of the option value, at the same time the point estimate.

The above is the standard stochastic mesh method. Related detail proofs of the low and high estimators can be referred to [5]. Owning to the existence of the estimated bias, here we use the bias reduction policy suggested in [1] to improve the performance of this naïve algorithm. Let $\hat{P}(t,S_t(i))$ be the estimated value of the true option value, $\hat{P}_{HA}(t,S_t(i))$ be the high-biased average estimator and $\hat{P}_{LA}(t,S_t(i))$ the low-biased average estimator, $i=1,2,\cdots,b$, $t=0,1,\cdots,M$.

The high-biased average estimator of the true option value is defined as

$$\begin{cases} \hat{P}_{HA}(M,S_T(i)) = h(M,S_M(i)) \\ \hat{P}_{HA}(t,S_t(i)) = \max\{h(t,S_t(i)), e^{-r\Delta t}\frac{1}{b}\sum_{j=1}^{b}\hat{P}(t+1,S_{t+1}(j))w(t,S_t(i),S_t(j))\} \end{cases} \quad (8)$$

where $i=1,2,\cdots,b$, $t=M-1,\cdots,1,0$.

Suppose the set $B=\{1,2,\cdots,b\}$, $B_j = B-\{j\}$. Define the estimate of the continuation value only using the mesh points in I at exercise date t, $\hat{C}(t,S_t(i),I)$, as

$$\hat{C}(t,S_t(i),I) = e^{-r\Delta t}\frac{1}{|I|}\sum_{j\in I}\hat{P}(t+1,S_{t+1}(j))w(t,S_t(i),S_{t+1}(j)) \quad (9)$$

where $w(t,S_t(i),S_{t+1}(j)$ is the same as in Eq. (7) and $|I|$ is the total number of elements in I. Then the low-biased estimator of the true option value only using the mesh points in I, $\hat{P}_L(t,S_t(i),I)$, is defined as

$$\hat{P}_L(t,S_t(i),I) = \begin{cases} h(t,S_t(i)) & if\ h(t,S_t(i)) \ge \hat{C}(t,S_t(i),I) \\ \hat{C}(t,S_t(i),I^c) & otherwise \end{cases} \quad (10)$$

where I^c is the complementary set of I relative to universal set U, i.e. $I^c = U - I$. Finally, the low-biased average estimator of the true option value is represented as

$$\begin{cases} \hat{P}_{LA}(M,S_M(i)) = h(M,S_M(i)) \\ \hat{P}_{LA}(t,S_t(i)) = \frac{1}{b}\sum_{j=1}^{b}\hat{P}_L(t,S_t(i),B_j) \end{cases} \quad (11)$$

where $i=1,2,\cdots,b$, $t=M-1,\cdots,1,0$.

Therefore, the estimator of the true option value is

$$\begin{cases} \hat{P}(M,S_M(i)) = h(M,S_M(i)) \\ \hat{P}(t,S_t(i)) = \frac{1}{2}(\hat{P}_{HA}(t,S_t(i)) + \hat{P}_{LA}(t,S_t(i))) \end{cases} \quad (12)$$

where $i=1,2,\cdots,b$, $t=M-1,\cdots,1,0$. With Eq. (12), we can obtain a more accurate point estimate of each point in the mesh and hence more precise point estimate of $P(0,S_0)$. Replicate the mesh for N times, average the estimated values, and then a better point estimate of the true option value can be given.

3. PARALLELIZATION OF STOCHASTIC MESH METHOD

As described on the above, the stochastic mesh method can be split into three stages: mesh generation; dynamic program-ming; mesh replication. We mainly study the parallelism of dynamic programming stage, which dominates the whole computation process.

3.1 Parallel Strategy for Dynamic Programming

Suppose we use np processors, where $np = 2^x$, $x = 1,2,\cdots$. We do not generate the mesh on the root process and then broadcast the mesh to other processes, but construct the whole mesh on each processor respectively, so that each one has the full information of the mesh and the communication in this step is avoid. Consider the generated mesh consisting of b paths of assets dynamics. We split the mesh by row, and then there is a "real" submesh consisting of $pb = b / np$ paths on each processor, i.e. the pth processor actually only deal with the paths $pb \cdot p+1,\cdots, pb\cdot(p+1)$, $p = 0,1,\cdots, np-1$.

We focus on the weights calculation firstly. To compute the weights $w(t, S_t(i), S_{t+1}(j))$, we need to know the transition density of state $S_{t+1}(j)$ at exercise date $t+1$ given $S_t(k)$, i.e. $f(S_t(k), S_{t+1}(j))$, where $i, j, k = 1,2,\cdots,b$, $t = M-1,\cdots,2,1$. We compute the transition density $f(S_t(k), S_{t+1}(j))$ on the pth process, where $j = 1,2,\cdots,b$, $t = 1,2,\cdots, M-1$, $p = 0,1,\cdots, np-1$, $k = pb\cdot p+1,\cdots, pb\cdot(p+1)$. We use all the np processes to compute the denominator in Eq. (7) in parallelism, i.e. the pth process compute the partial sum $\sum_{k=pb\cdot p+1}^{pb\cdot(p+1)} f(S_t(k), S_{t+1}(j))$, and then gather the results on each processor to get the total sum. Then we compute the local mesh weights $w(t, S_t(i), S_{t+1}(j))$ on the pth process, since the whole transition density information is known by finite communication among different processes, where $p = 0,1,\cdots, np-1$, $j = 1,2,\cdots,b$, $t = M-1,\cdots,2,1$, $i = pb\cdot p+1,\cdots, pb\cdot(p+1)$.

Then, we discuss the parallel computation strategy for the estimated continuation value calculation. Owning to the pth process only knows the estimated values $\hat{P}(t+1, S_{t+1}(j))$ at exercise date t, where $j = pb\cdot p+1,\cdots, pb\cdot(p+1)$, we need to gather values distributed on the other processes and then backward compute the estimated value $\hat{P}(t, S_t(j))$, $t = M-1,\cdots,2,1$. At exercise date t, we broadcast $\hat{P}(t, S_t(j))$ ($j = pb\cdot p+1,\cdots, pb\cdot(p+1)$) among different processors once these values are computed, where $t = M,\cdots,2,1$. This is the main communication of the parallel algorithm. Each processor needs to broadcast $N\cdot M\cdot pb$ estimated values during the whole dynamic programming process assuming that we replicate the mesh for N times.

With the parallel computing results of weights and continuation values at each exercise date, we recursively compute the point estimate of each point in the mesh. After replicating the mesh for N times, the better point estimator of the true value is obtained by averaging these estimated values.

3.2 Implementation

To generate the mesh, we need to use random number sequences firstly to simulate several independent paths of assets dynamics. Because the convergence performance of Monte Carlo method using low discrepancy sequences, such as Sobol, Faure and Halton sequences, is better than using pseudorandom sequences, we generate the Sobol sequences with skipping the first 256 points of the sequences to improve the quality of the sequences [6]. This is the so-called Quasi-Monte Carlo method. Then we use inverse method to transform Sobol sequences into ones following n-dimensional standard normal distribution [7].

We use MPI (Message Passing Interface) to efficiently implement the parallel algorithm described on the above, since in different computing environments the performance of using MPI is more flexible than the one obtained using OpenMP. In the computational process of dynamic programming, we construct a new MPI datatype, which consists of the estimated option values at current exercise date on each processor, to associate the global communication at each exercise date.

4. NUMERICAL STUDY

4.1 Experimental Environment

We implement our parallel algorithm and run large-scale simulation on supercomputer DeepComp7000, which ranked 31th in the Top500 issued in Jun. 2009. The supercomputer consists of two heterogeneous parts: cluster part (we use "Cluster" for short) and SGI Altix4700 nodes (we use "SGI" for short). The C code of this algorithm runs in both parallel computing environments to measure the performance. We show the system parameters of both environments in Table 1. The default code optimization option of the compiler is used.

Table 1. System parameters of "Cluster" and "SGI"

parameter	**Cluster**	**SGI**
Operating System	Red Hat Enterprise Linux Server release 5.1	SUSE Linux 10SP2
Linux Kernel	2.6.18-53.e15	2.6.16.60-0.21.default
Complier	Intel C/C++ Compiler 11.0.081	Intel C/C++ Compiler 10.1.008
MPI Library	Intel MPI 3.2.011	SGI MPT 1.2
Processor	Intel Xeon E5450, Quad-core, 3.00Ghz	Intel Itanium2 9140M, Dual-core, 1.66Ghz

4.2 Numerical Examples

We use the subset of test cases shown in [5]. Assume that the underlying assets of the American option follow multi-dimensional geometric Brownian motion. To be simple, we let the private parameters of each asset identical, i.e. $S_0^i = S$, $\delta_i = \delta_j$, $\sigma_i = \sigma_j$, $\rho_{ij} = \rho$, $i, j = 1,2,\cdots,n$, $i \neq j$. We mainly focus on the following test cases:

(1) 5-Dim Case: American max call option on five assets

$n = 5$, $S = 100.00$, $K = 100.00$, $r = 0.05$,
$\delta_i = 0.10$, $\sigma_i = 0.20$, $T = 3$, $i = 1,2,\cdots,n$

(2) 7-Dim Case: American geometric average option on seven assets $n = 7$, $S = 100.00$, $K = 100.00$, $r = 0.03$,
$\delta_i = 0.05$, $\sigma_i = 0.40$, $T = 1$, $i = 1,2,\cdots,n$

The payoff functions $h(t, S_t)$ of each case are described in Section 2. The parameters of the algorithm are as follows: b, the mesh size; N, the total number of mesh replication; np, the number of processors. We use speedup to measure the performance of parallel algorithm and estimated bias to study the performance of stochastic mesh method. The speedup is the ratio of the serial computational time to the parallel consumed time; while the estimated bias is calculated via “true” value subtracting point estimated value.

Firstly, we show the results in the standard settings: $\rho = 0.0$ in both cases; $M = 3$ in 5-Dim test case, and $M = 10$ in 7-Dim test case. In each test case, we let the mesh size b equal to 1024, 2048 and 4096 respectively and we replicate the mesh for 50 times. Then, we execute the codes in both computing environments for each level of mesh size, using 1, 2, 4, 8, 16, 32 and 64 processors respectively. The results are illustrated in Table 2. For each level of mesh size, the above row in the table is the computing time in seconds on “SGI” while the below on “Cluster”. The values in brackets below the point estimators are the corresponding estimated variances. We compute the “true” values of both options by setting $b = 32768$, $N = 50$ and running the codes on “Cluster” using 128 CPUs. The computations cost 1003.86 and 4957.28 seconds respectively. The “true” values of both options are 25.224569 and 3.325708, with variances 0.000203 and 0.000079 respectively. The corresponding European options are 23.051029 and 2.419403 respectively, based on 100,000,000 Monte Carlo simulations.

Table 2. Results in the standard 5-Dim and 7-Dim settings

n	b	Point Estimator	Estimated Bias	np: 1	2	4	8	16	32	64
5	1024	25.028061 (0.068107)	0.196508	86.00	41.65	21.62	12.49	7.67	6.12	10.70
				26.00	12.35	6.15	3.35	2.67	2.13	2.31
	2048	25.073362 (0.013985)	0.151207	371.00	200.61	89.74	47.03	25.33	32.74	27.73
				185.00	95.63	46.76	24.54	10.79	5.30	6.13
	4096	25.103284 (0.05256)	0.121285	1490.00	780.85	416.41	233.47	100.27	60.10	54.99
				686.00	351.06	210.95	146.18	72.37	26.30	14.03
7	1024	3.179396 (0.02966)	0.146312	629.00	287.72	157.23	83.10	44.26	33.50	31.40
				181.00	88.19	67.40	26.82	18.66	10.62	16.52
	2048	3.261448 (0.01806)	0.06426	2624.00	1240.38	646.91	315.88	160.91	98.14	73.86
				980.00	525.83	349.72	168.27	63.58	31.85	27.24
	4096	3.299665 (0.01100)	0.026043	10581.00	7731.02	2851.46	1363.94	636.94	348.06	337.50
				4147.00	2271.78	1140.39	732.55	417.04	151.71	85.46

As demonstrated in Table 2, the estimated bias decreases nonlinearly as the mesh size increases. The decrement in the 7-Dim case is about 82.2%, much larger than 38.3% in the 5-Dim case. However, as the mesh size increases, the variance of point estimator decreases 92.3% in the 5-Dim case while 62.9% in the 7-Dim case. The estimated bias in 7-Dim setting is much smaller than the bias in 5-Dim case using the same mesh size. When the mesh size increase k times, the computational time on “SGI” increase about k^2 times while more than k^2 times on “Cluster”. In 5-Dim setting, when we run the codes on “SGI” using 64 processors, the computational time of test cases with mesh size 2048 and 4096 is respectively just 2.59 and 5.14 times of the consumed time of test case with mesh size 1024. While on “Cluster”, the corresponding values are 2.66 and 6.07 respectively.

Based on the results in Table 2, we show the speedups of the above test cases on “SGI” in Figure 1. The legend “Linearity” describes the ideal linear speedup. The legend “5-Dim1024” represents 5-Dim test case when mesh size equals to 1024, and the others are likewise. The speedups of using much more processors don’t increase along the “Linearity” line, just below that line. As for 5-Dim test case, some speedups don’t increase when using more than 32 processors, while all the speedups in 7-Dim cases increase all the time. The main reason is that communications among too many processors can dominate the whole process. The speedups of both cases on “SGI” using fewer processors have little difference but vary greatly when using much more CPUs. We can also observe the above phenomenon based on the speedups on “Cluster”.

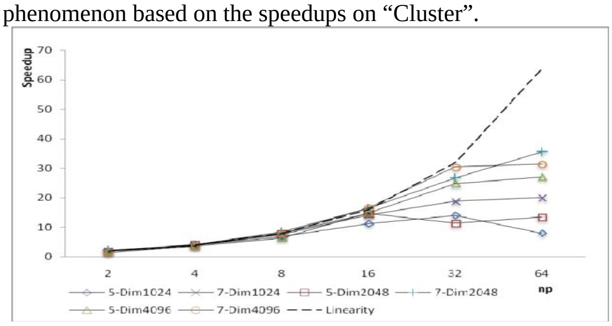


Figure 1. Speedups of test cases on “SGI”

We then compare the parallel performance between “SGI” and “Cluster”. The speedups of 5-Dim cases on “SGI” and “Cluster” are demonstrated in Figure 2. “5-Dim1024Cluster” in the figure represents the speedups of 5-Dim test case with mesh size 1024 when running on “Cluster”. The speedups begin to vary greatly when using more than 8 processors. Better parallel performance on “Cluster” is obtained when the computational effort is quite large and the procedure runs using more processors. This observation is demonstrated by the “5-Dim 2048Cluster” using 16, 32 and 64 CPUs and “5-Dim4096Cluster” using 64 CPUs. The parallel speedups in both environments don’t move toward the same direction simultaneously. The speedups of “5-Dim2048Cluster” and “5-Dim2048SGI” agree with this observation. As for the 7-Dim case, the discrepancy between the speedups on “SGI” and “Cluster” is smaller than the one in 5-Dim case.

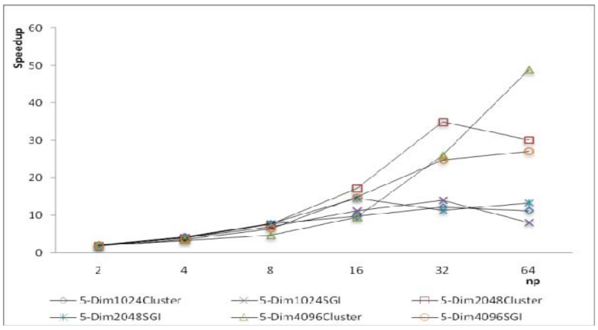


Figure 2. Speedups of 5-Dim test cases on “SGI” and “Cluster”

We also present the numerical results of the 7-Dim case with different numbers of exercise dates. We set the number of exercise opportunities equal to 10, 20, 30, 40 and 50 respectively. In each test case, $b = 4096$, $N = 50$. Table 3 shows the results. For each number of exercise dates, the first row in the table shows the computational time on “SGI” and the second row on “Cluster”. The “-” in the table represents that the consumed time exceeds 6 hours. As the number of exercise dates increases, the estimated option value decreases and the computational time increases almost linearly. On “Cluster” using 64 CPUs, the computational time of test cases with exercise dates 20, 30, 40 and 50 is respectively 2.15, 3.36, 4.35 and 5.98 times of the consumed time of test case with exercise dates 10.

Table 3. Results in the 7-Dim case with different exercise dates

Exercise Dates	Point Estimator	np						
		1	2	4	8	16	32	64
10	3.299665 (0.001100)	10581.00	7731.02	2851.46	1363.94	636.94	348.06	337.50
		4147.00	2271.78	1140.39	732.55	417.04	151.71	85.46
20	3.089466 (0.000984)	-	11368.46	6078.62	3040.96	1379.79	742.55	538.96
		8776.00	4780.91	2605.75	1781.82	956.36	374.65	183.78
30	2.833293 (0.000654)	-	16301.30	10349.91	4650.35	2521.30	1237.52	981.32
		14109.00	7753.29	3972.82	2926.89	1509.14	655.24	287.44
40	2.619965 (0.000614)	-	-	12189.84	5963.58	2966.40	1672.48	1066.51
		17980.00	9910.18	5526.19	3888.15	1619.10	655.05	371.75
50	2.446534 (0.000532)	-	-	15246.06	7959.88	3649.26	1954.20	1316.79
		-	12574.81	6527.18	4510.93	2015.47	998.87	510.66

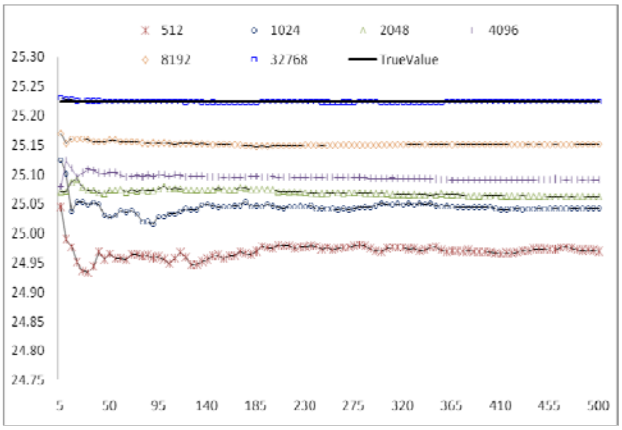


Figure 3. Estimated option values with N from 5 to 500

Furthermore, we observe the influence of the number of mesh replication N on estimated option values, i.e. the behaviors of estimated option value when replicating different numbers of meshes. We use test cases with mesh size 512, 1024, 2048, 4096, 8192 and 32768 in 5-Dim setting. We let N vary from 5 to 500 with step 5. The results are demonstrated in Figure 3. The “TrueValue” is calculated with parameters $b = 32768$ and $N = 50$. For each level of mesh size, the smaller the number of mesh replication N, the severer fluctuation the estimated option value; the lines of estimated option values move toward the “TrueValue” line, i.e. the estimated option values converge with the increase of mesh size, which is consistent with results in [2]. For reasonable level of N, the line of estimated values then becomes quite flat; when the mesh size is large enough, the line of estimated values behaves much smoother, almost parallel to the horizontal axis. However, there is a quite frustrating observation. When the mesh size isn’t large enough, the estimated value appears to increase with the increasing of mesh size, i.e., the larger the mesh size, the higher the estimated option value. The stochastic mesh method combined with performance enhancement policy using bias reduction might underestimate the true option value when using smaller mesh size. In practice, constrained with the computational power and latency time, the practitioners often simulate the option value with lower level mesh size and may obtain lower-estimated result, which is not in favor of them to trade or hedge the American options.

5. CONCLUSIONS

In this paper, we studied the parallelism of stochastic mesh method for high-dimensional American option pricing, implemented this effective algorithm using MPI, ran large-scale parallel simulation and analyzed the performance based on several standard numerical examples on DeepComp7000. The parallel performance is much better for cases with large-scale simulated parameters and has good scalability in different parallel environments. Moreover, the approach with bias reduction policy using smaller mesh size might underestimate the true option value. Our results can motivate the further performance enhancement of stochastic mesh method and then enable the practitioners to embrace this effective method.

Since modern financial derivatives are more and more complicated and effectively pricing these instruments requires challenging computational effort, parallel derivatives pricing algorithm will be increasingly popular. Future research involves development of efficient parallel algorithms to solve more complex financial practical problems, such as large-scale portfolio optimization and risk management.

REFERENCES

[1] A. N. Avramidis, and P. Hyden, "Efficiency improvements for pricing American options with a stochastic mesh", Proceedings of the 1999 Winter Simulation Conference, 1999, pp.344~350.

[2] A. N. Avramidis, and H. Matzinger, "Convergence of the stochastic mesh estimator for pricing American options", Journal of Computational Finance, 7(4), 2004.

[3] M. Broadie, and P. Glasserman, "Pricing American-style Securities using Simulation", Journal of Economic Dynamics and Control, 21(8-9), 1997, pp.1323~1352.

[4] M. Broadie, and P. Glasserman, "A stochastic mesh method for pricing high-dimensional American options", working paper, Columbia University, 1997.

[5] M. Broadie, and P. Glasserman, "A Stochastic Mesh Method for Pricing High-Dimensional American Option", Journal of Computational Finance, 7(4), 2004, pp.35~72.

[6] Glasserman, P., Monte Carlo Methods in Financial Engineering, Springer-Verlag, New York, 2004.

[7] C. Lemieux, and P. L'Ecuyer, "On the Use of Quasi-Monte Carlo Methods in Computational Finance", Computational Science -- ICCS 2001, May 2001, Lecture Notes in Computer Science, Springer-Verlag, 2001, pp.607~616.

[8] F. A. Longstaff, and E. S. Schwartz, "Valuing American Options by Simulation: a Simple Least-square Approach", Review of Financial Studies, 14, 2001, pp.113~147.

[9] B. Sentance, "Analytics, Data and the Time for High Performance Computing in Financial Markets", white paper, Microsoft, Nov. 2007.

RESEARCH ON RMI-BASED DISTRIBUTED APPLICATION ARCHITECTURE

ZHOU Xiaojian, Ren Xuehui
School of Computer and Communication, Lanzhou Univ. of Tech.
Lanzhou,Gansu 730050, China
Email:zhouxj_lz@163.com

ABSTRACT

RMI provides the necessary mechanisms for the operation and control of distributed object. This paper analyzes detailedly the core and architecture about RMI, depicts an RMI distributed object model, and studies the developing process of a RMI distributed object application through a case. These reveals that RMI possesses powerful function and predominance on distributed object.

Keywords: RMI, Distributed Object, Architecture, Stub, Design pattern

1. INTRODUCTION

The communication between different hosts and address space was realized by distributed system. Java provides a basic communication mechanism which was called sockets. It can satisfy the requirement of general communication. However, the encoding and decoding must be implemented on the client and server which locate at application layer. And the protocol is difficult to design. In addition, it is well known that RPC(Remote Procedure Call) is a kind of sockets, but it can not support the distributed object.

Java RMI(Remote Method Invocation) is an intrinsic mechanism to implement the network function in Java [1]. RMI allows to design the distributed object in Java and communicate in memory or JVM(Java Virtual Machine). The concept of RPC was absorbed by RMI, So, it was permitted to communicate between the distributed object instead of activate the method in program only. Then the network codes can be made into a whole object in order to take a lot of advantages of the object oriented programming, such as inheritance, encapsulation, polymorphism, etc.

2. RMI-BASED DISTRIBUTED OBJECT APPLICATION

RMI application is usually made up of the client program and server program [2]. The server build remote object and its reference, then wait the client to refer the method belong to the remote object. The client obtains the remote object from the server and then makes method invocation. RMI provides a mechanism for client and server to communicate and send messages to each other.

When RMI is used to distributed object, there are two ways to obtain the reference of the remote object: one of them uses the tools rmiregistry, which is named simply by RMI, to register the remote object, other one send the reference of the remote object as the part of normal operation. The detail of communication between the remote object is handled by RMI, and pellucid to programmer. It seems as a local method invocation for the programmer. RMI allows the caller to visit the remote object by object, so RMI provides a necessary policy to upload the codes of the object.

Figure 1 describes a RMI distributed application system which obtains the quote of remote object with registry. The server makes a contact with the remote object by invocating registry, and the client looks up the remote object and refers the method of the remote object with the server's registry. RMI distributed system may upload the class byte code, and the java platform support RMI to upload it with URL protocol, such as HTTP, FTP, file, etc.

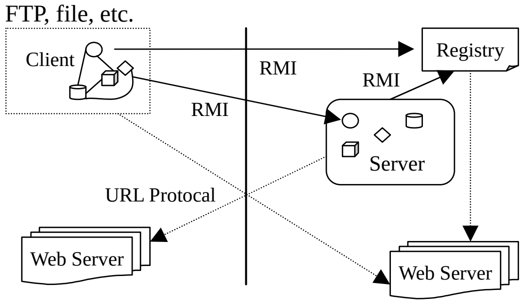


Figure 1. An RMI distributed application system

3. ARCHITECTURE OF RMI

The architecture of RMI is designed in order that a distributed object can be created by integrating java programming language and local object model [3]. RMI frame can expand the security and strength to the domain of the distributed computing successfully.

3.1 Interface: Core of RMI

The architect of RMI is based on such a great programming idea: detaching the definition and implement of method. RMI allows the definition of method and codes implemented method to be separated and run in different JVM.

Interface defines the public information of object, such as the name and the parameters of method, etc. The implementing provides the core logical program for object, including some special algorithm, logic and data. A distributed system only require clients to provide the definition of service, and the service is provided by server. As shown in Figure 2, the remote service is defined by interface and is implemented by interface and is implemented by classes in RMI.

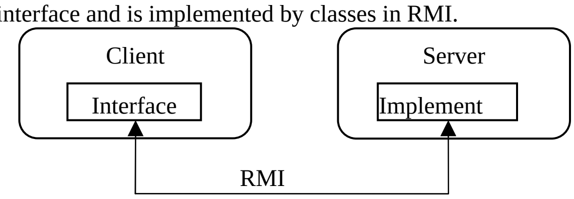


Figure 2. Interface and implementation are separated

Java interface does not contain any running codes, RMI support to implement two classes in a same interface, one is running in server, another is running in client as a remote service agent, which is shown as Figure 3.

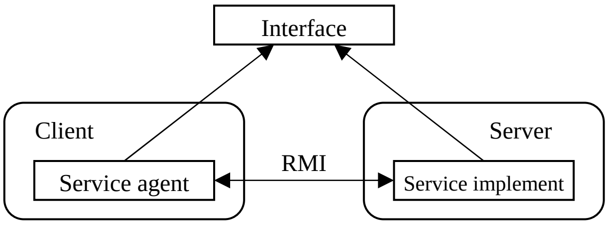


Figure 3. Service proxy and implementation

3.2 Layer Structure of RMI

Figure 4 shows that the implementing of RMI is built on the basis of 3-layer abstraction [4]. The first layer is called Stub/Skeleton layer under the developer view, which interprets the method invocations about reference variable on the client interface or redirects these invocations to remote RMI server. The second layer is called remote reference layer, which mainly interprets and manages the reference from clients to the remote service object. The third one is transport layer, which is the connection between the computers in internet based on TCP/IP protocol and provides basis connection, such as the firewall passing policy.

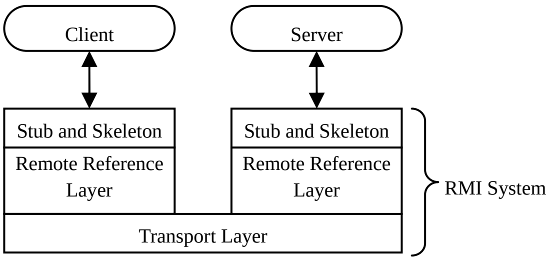


Figure 4. RMI is a layered architecture

On Stub and Skeleton layer, RMI expressed with proxy design pattern [5]. The proxy design pattern provides an agent for other object in order to control it to be visited. Figure 5 present a structure of proxy design pattern.

In RMI applications, the Stub class plays a role as agent, the remote service implement class plays a role as RealSubject class does. Skeleton layer knows how to communicate with Stub by RMI link and executes their conversation. It reads parameters for method invocation from link, then builds a vocation to remote implement object and accepts the return values, finally write the return values back to Stub. In Java2, the Skelet0n has been disabled, so RMI build the connection with remote object by mapping.

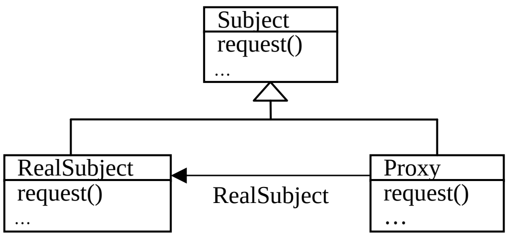


Figure 5. RMI uses the proxy design pattern

The remote reference layer defines and supports the connection invocation semantic meaning for RMI, and provides a RemoteRef object to express the connection to the remote service implementing. The Stub object sends method invocation with the invoke() method from RemoteRef, and RemoteRef object knows the invocation semantic meaning of the remote service. In Java2, the implementation of RMI appends a semantic meaning for client/server connection. Because RMI supports activated remote object, when the proxy builds a method invocation of an activated object, RMI make sure whether the remote service implementation object is in sleeping state. RMI will instantiate the remote service implementation object and restore its state from disk file if it is in sleeping state.

Figure 6 illustrates that the transport layer build a connection between JVM, and all the connection uses network connection by TCP/IP based on streaming. Although two JVM is running in the same physical computer, they may be connected by the TCP/IP protocol stack of its host.

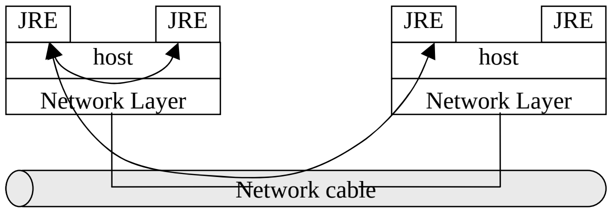


Figure 6. Transport layer’s network connection

4. DEVELOPING RMI-BASED DISTRIBUTED SYSTEM

Based on the above discussion, it can be seen that developing the client/server program based on RMI is a very directly process. The general step of developing a RMI-Based distributed system [6] is as follows:

(1) Defining remote interface: We must create a remote interface. It contains every method that the remote objects provide to remote host, and the interface must belong to java.rmi.Remote.

```
import java.rmi.Remote;
import java.rmi.RemoteException;
public interface TestInterface extends Remote {
public String getDescription() throws RemoteException;
}
```

(2) Implementing remote interface: The remote object can be expanded, so it can provide Pre-activated, sleeping, activatible and RMI/IIOP server object behavior.

```
import java.rmi.*;
import java.rmi.server.*;
public class TestImpl extends UnicastRemoteObject
implements TestInterface {
  public TestImpl(String n) throws RemoteException {
    name = n;
}
public String getDescription() throws RemoteException {
  return “I am ” + name + “!”;
  }
  private String name;
  }
```

(3) Creating RMI Stub/Skeleton: With rmic tools we can create the main frameand stub of RMI server and client, which are as follows:

TestImpl_Skel.class, TestImpl_Stub.class.

(4) Implementing RMI server registry: Firstly implementing a class, then register a RMI server to RMI registry, or register an activatible RMI/IIOP server to an activated daemon process of RMI, or register a RMI/IIOP server to RMI/IIOP naming

service of RMI/IIOP. After an initial RMI server object is looked up by a RMI client, other RMI server object will be instantiated and registered by the RMI server created initially or by the RMI server created subsequently.

```
  import java.rmi.*;
  import java.rmi.server.*;
  public class TestServer {
  public static void main( String args[] ) {
    try {
        System.out.println("Create RMI Server……");
        TestImpl t = new TestImpl("Test");
        System.out.println("Binding……");
        Naming.rebind("test", t);
      }
      catch(Exception e) {
      System.out.println("Error: " + e);
    }
  }
}
```

(5) Implement RMI client: Now a RMI client can be created by the RMI stub that is created in the preceding, RMI client must look up all reference of RMI server object using RMI registry or RMI/IIOP naming service. RMI client uses RMI stub pellucidly.

```
  import java.rmi.*;
  import java.rmi.server.*;
  public class TestClient {
  public static void main( String args[] ) {
    System.setSecurityManager(new RMISecurityManager());
    String url = "rmi://dataserver.cad.gsut.edu.cn/";
      try {
TestInterface t=(TestInterface)Naming.lookup(url+"test");
        System.out.println( t.getDescription );
        }
        catch ( Exception e ) {
        System.out.println("Error: " + e);
        }
      }
    }
```

(6) Deployment: After testing the programs and its deployment status, then we may deploy them to clients and servers. Certainly, it is too difficult to deploy RMI because sometimes it runs normally and sometimes it runs with implicit errors. The status can be improved in a great extend by taking an expansion after making deploying in local host.

5. CONCLUSIONS

The architecture of RMI and the process to implement was discussed in the paper. As an important part and base of J2EE architecture, RMI had a great advantage and function to build distributed object. Especially when it works with other Java API, it will give an enormous impetus to enterprise application. For example, JNDI provides name and directory service, JDBC provides database connection and access technology. Moreover, RMI is a necessary part of EJB. Many distributed object use RMI as a network standard API. The RMI technology is worth to be further studied and applied.

REFERENCES:

[1] http://java.sun.com/rmi/

[2] Du Feng-Lei . The Distributed Application Based on Java With RMI [J]. macroelectronics & compter，2006，(07)

[3] Meng-Chun Wueng， Fu—Fang Yang， Cheng-Zen Yang. A novel Java RMI middleware design for active networks. TENCON 2004. 2004 IEEE Region 10 Conference Volume C，21—24 Nov. 2004 Page(s):68—71 V01. 3

[4] Paul J.Perrone, et al. Building Java Enterprise Systems with J2EE [M]. USA:Sams Indianapolis.2001

[5] Erich Gamma, Richard Helm, Ralph Johnson, et al. Design Patterns:Elements of Reusable Object-Oriented software [M]. USA:Addison Wesley.2000

[6] Rickard Oberg. Mastering RMI: Developing Enterprise Applications in Java and EJB. John Wiley & Sons,Inc. 2003,7(7):67-70

DESIGN AND APPLICATION OF THE AUTOMATIC ORIENTATION DEVICE FOR NOISE TEST OF SOUND TOYS

Yin Lijuan1, Xu Xiaoguang1, Wang Jinsong2, Zhang Yanqun1, Du Zhuoxin1
1.Shenzhen Entry-Exit Inspection and Quarantine Bureau, Shenzhen Guangdong China
2.Shanghai Entry-Exit Inspection and Quarantine Bureau, Shanghai China
Email: yinlj411@163.com

ABSTRACT

Based on the requirements of noise test of safety test standard on toys, the method of noise test and defined position of measurement point were described. The automatic orientation device for noise test of sound toys from construction design, control principle and disposition of hardware were explained. Its application resolved key problem of defined position of test point on the noise test of sound toys.

Keywords: acoustics; sound toys; noise test; measurement point; automatic orientation

1. INTRODUCTION

Noise is a kind of sound which appears in the form of vocality in the human daily society communication. It makes people uncomfortable and worried, so it is a kind of unwelcome sound for people, but usually unavoidable. Sound toys positively makes children happy, but which also makes potential dangerous to hearing of children at that moment. People hope that children's toys are safe, so that American, European and National safety standard on toys institute the ultimate noise weighted requirements which sound toys permit.

The noise test method of safety test standard on toys is very complicated. Different kinds of toys have different test methods. The noise test methods of National standard of toys GB6675, European standard of toys EN71 and International standard of toys ISO8124 are almost sameness. At present, sound toys test methods are artificially, which makes toy noise test more difficult, test result inaccurate, repetition poor, error great and test efficiency low and so on, so that it is very important to actualize toy noise test automatization.

2. TEST REQUIREMENTS

Using sound pressure level show the noise strong or weak on the noise test, sound pressure level defines as Eq. (1):

$$L_p = 10\lg\frac{p^2}{p_0^2}(dB) \qquad \text{Eq. (1)}$$

Eq. (1): $P_0 = 2\times10^{-5}$ P_a is benchmark sound pressure level, P is availability sound pressure level.

IEC standard 2002 IEC61672-2002 measure with the A-weighted sound level(L_{pA}), the A-weighted single-event emission sound pressure level ($L_{pA,1s}$) and the C-weighted peak sound level ($L_{pC\ peak}$). A-weighted sound level shows the noise is similar to human feeling, C-weighted peak sound level keeps straight answer in most frequency area and attenuates less in low frequency and high frequency of hearing range. Take European standard for safety of toys EN71-1:2005 for example, when tested according to the determination of emission sound pressure levels, toys which are clearly designed to emit sound shall conform to the following requirements:

(1) The A-weighted emission sound pressure level, L_{pA}, produced by close-to-the-ear toys shall not exceed 80 dB when measured in a free field; The A-weighted emission sound pressure level, L_{pA}, produced by close-to-the-ear toys shall not exceed 90 dB when measured using an ear coupler.

(2) The A-weighted single event emission sound pressure level, $L_{pA,1s}$, produced by rattles or squeeze toys shall not exceed 85 dB.

(3) The C-weighted peak emission sound pressure level, L_{pCpeak}, produced by rattles or squeeze toys shall not exceed 110 dB.

(4) The C-weighted peak emission sound pressure level, L_{pCpeak}, produced by a toy using percussion caps shall not exceed 125 dB.

(5) The C-weighted peak emission sound pressure level, L_{pCpeak}, produced by any type of toy excluding toys using percussion caps shall not exceed 115 dB.

(6) If the C-weighted peak emission sound pressure level, L_{pCpeak}, produced by a toy exceeds 110 dB, the potential danger to hearing shall be drawn to the attention of the user.

3. TEST METHOD ANALYSE

In toy sound pressure levels testing, microphone is mounted on the measuring position points for measurement test with modular precision sound analyzer. Measuring positions and numbers are determined by variety of toy. Reference box and measurement box are two important conceptions in toy sound pressure levels testing, first one means a minimum cuboid of toy, latter one is a cuboid with certain distance from reference box, they are both virtual tanks in practical test, all of measurement positions should on measurement box.

European standard for safety of toy EN 71-1 classified sound pressure levels testing method with 6 sorts according with variety of toy:

(1)Close-to-the-ear toys: The distance required between microphone and the surface of main sound source exists of toy is 2.5 ± 0.5cm, and the toy with a test distance at least 100 cm above the floor. Locate the position of the maximum sound pressure level (L_{PA}) of a close-to-the-ear toy by moving the toy or the measuring microphone so that the measuring distance is (2.5 ± 0.5) cm from that surface of the toy where the main sound source exists. This position is the microphone position for measurements. For earphones and headphones the microphone positions are given by the coupler.

(2) Cap-firing toy: Mount the main sound emitting part of the toy at the measuring coordinate system in its normal operating orientation in which is the center of the rotate axes. Mount the main sound emitting part of the toy at the origin of the measuring coordinate system in its normal operating orientation

in such a way that the main axes of the toy coincide with the axes of the measuring coordinate system. If the length of the toy exceeds 50 cm, rotate the toy in the xy-plane 45° around the z-axis without changing the microphone positions. As shown in Figure 1, in which "1" is the measurement positions of microphone and there are six positions, the measuring distance is 50 cm (the distance from microphone to the strike point), the center as shown in figure 1 is the strike point of cap-firing toy of sound emitting origin.

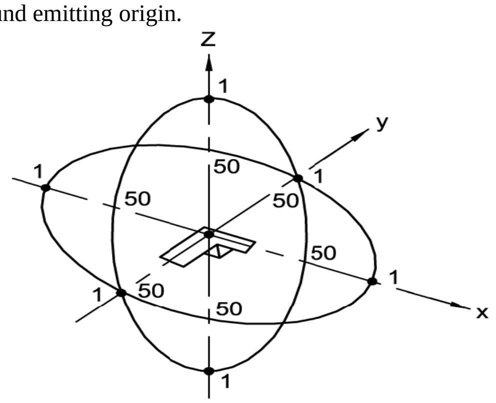


Figure 1. Cap-firing toy measurement position

(3) Rattles and squeeze toys: A distance of measuring is 50cm(distance from microphone to sound emitting origin of toy), Mount the microphone 1.2 m above the floor, Measure the A-weighted single-event emission sound pressure level, LpA,1s, and the C-weighted peak emission sound pressure level, $L_{pC\ peak}$ for ten movements (squeezes, beats). Carry out the test three times.

(4)Other hand-held toys: Select six microphone positions on a box-shaped measurement surface, as specified in figure 2, black point positions on the measurement box at a measuring of 50cm distance from the reference box of the toy is the measurement place, The positions are at the centres of the sides of the reference box. In Figure 2 "1" is measurement box, "2" is reference box.

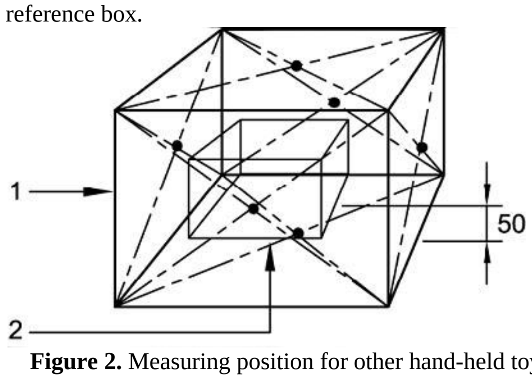


Figure 2. Measuring position for other hand-held toys

(5) Stationary and self-propelled table-top and floor toys: Place toy on measuring table, The sides of the measurement box are always 50 cm from the sides of the reference box. Select five, or if the length or width of the toy is larger than 100 cm, nine microphone positions all on measurement box surface as specified in Figure 3. The sides of the measurement box with height H are always 50 cm from the sides of the reference box. All microphone positions are on the measurement box. H= height of reference box+50cm, A is reference box and B is measurement box. H is height of toy plus 50cm, 1~9 is measuring position, therein, 1~4 on the side of measuring box, 5~8 on the tip of measuring box, 9 on the top center of measuring box, all microphone positions are on the measuring box.

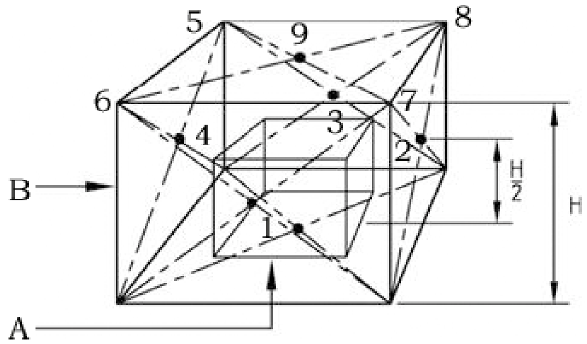


Figure 3. positions for the measurement of stationary and self-propelled table-top and floor toys

(6) Pull and push toys and hand-activated spring-propelled toys: For toys with a width(w) of 25cm or less, use two microphones at distances(d) 50cm from the x-axis of the measuring coordinate system; For toys with a width(w) of more than 25cm, use two microphones at distances(d) 40cm plus half the width of the toy from the x-axis(40 + w/2) as shown in Figure 4, "1" is microphone, "2" is measurement end-point, "3" is toy.

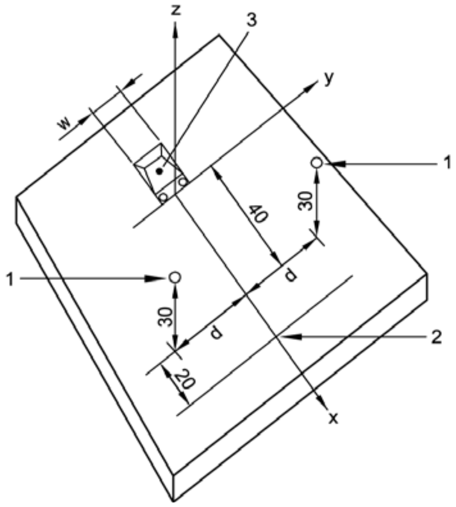


Figure 4. Positions for the measurement of pull and push toys and hand-activated spring-propelled toys

4. DESIGN OF AUTOMATIC ORIENTATION DEVICE

Base on analysis of noise test, how to locate the measurement point is the key work which is the premise to get the accurate measurement result. After the measurement point is located, noise test can be performed with the microphone at the measurement point. If there is more than one measurement point, the microphone is always moved from one position to another. Whenever is practicable, it is an alternative to rotate the test object instead. Attention must be paid to maintaining the correct measuring distance.

Structural graph of Automatic orientation device for the noise test of sound toys is in the broken line of figure 5 as follows. It includes test platform, human-machine interface device, control part, drive part and orientation device.

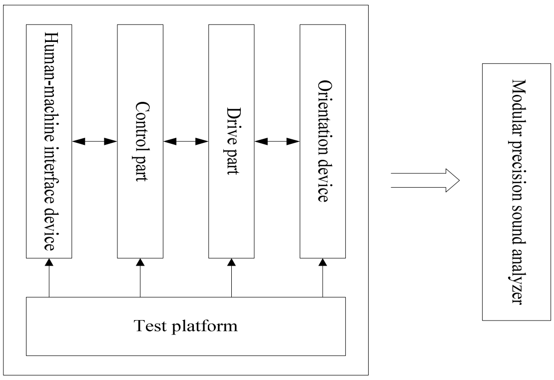


Figure 5. Structure of orientation device

The human-machine interface adopts the operating panel F920 of Mitsubishi. The control part adopts the PLC FX1S-10MT of Mitsubishi. The drive part is formed of four walking machines and four drivers of JQF-MD808. There are two kinds of fixed device: platform device and desktop device with turnplate . The former is used to fix the close-to-the-ear toys, cap-firing toys, rattles and squeeze toys. The latter is used to fix the stationary and self-propelled table top, floor, and crib toys, pull and push toys and hand-activated spring propelled toys and other hand-held toys. Consider that the reflection of noise will increase the test data of noise; device to fix the toys should not be too big, as it will influence the measurement result. We adopt the sound pressure level analyzer BK2260 of B&K Corp. Ltd in Danmark to test the sound pressure level with the microphone which is at the measurement point.

5. CONTROL ELEMENTS

The control elements of automatic orientation device as shown in figure 6. The hardcore is made up with PLC control system, drivers and steppers. There are four steppers which use to control microphone up and down, go ahead and back off, go left and right, the clockwise rotating and anticlockwise rotating of test platform, the three preceding steppers control the movement of microphone of axis X,Y,Z, another stepper control the movement of the test platform which with fixed device.

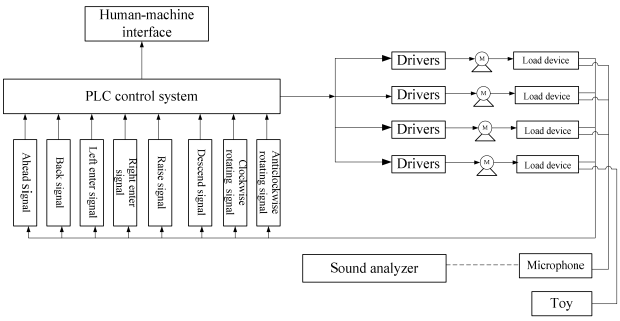


Figure 6. Control Elements

Before test, first mount the toy on the test table with the automatic orientation device, then input the test standard, the type of toy and other test information with test panel. PLC will analyse and dispose the output signal then input it to the drivers when it received these information, drivers will amplify and match the control signal and then control the movement of the stepper. The drivers observe the principle which control the movement of microphone mainly, control the movement of test panel secondly, confirm each measurement point on the correct place according the requirement of standard. The drivers will change the movement direction of stepper by setup the limit switch on the different moving directions. Lastly, display the noise parameter of the sound analyzer, the test parameter and result will be displayed on the operate panel at the same time.

6. TEST AND APPLICATION

Test steps: first, choose standard; second, choose the type of the toy on the control panel with the characteristic of the toy; third, when the test positions relate to the size of the toy, should input the length, width, height of the toy and the height of the test table or the height of the surface of the clamp; fourth, confirm the centre of the reference box and the main sound emitting part of the cap-firing toy is coaxial with the stepper for control angle when fixing the toy. The normal operational diretion make the main axes of the toy coincide with the x-axes of the measuring coordinate system; fifth, confirm the test position; last, begin noise test.
The microphone tip should keep horizontal or vertical when the testing, the operator should not effect the sound emitting of the tested toy, and also should not reflect the toy sound which will induce the sound pressure up on the testing point.

(1) Close-to-the-ear toys: Mount close to the ear toys in the centre of rotor shaft support at least 100 cm above the reflecting plane, the main sound emitting plane should coaxial with the tip of microphone, move microphone to touch the maximum sound emtting part of the toy. After start the test on the control panel, the automatic orientation device will automatically locate the position of the maximum sound pressure level (L_{PA}) of a close-to-the-ear toy by moving the microphone on the measuring distance is (2.5 ± 0.5) cm from that surface of the toy where the main sound source exists. This position is the microphone position for measurement.

(2) Cap-firing toy: Nip the test specimen on the rotor shaft fixed support, confirm the bump point is coincide with the rotor shaft, gunpoint and the original place of microphone which on the x-axes is coaxial by moving the nipping device, then enter the bump point height which from floor, the original place of microphone is (0,0,0). When testing, the automatic orientation device will move the microphone on the measuring distance is 50cm from the bump point where the same height with the point automatically. After this posision has tested, then let the automatic orientation device rotate the toy for 90°,180°, 270°etc place to continue the test. If the length of the toy exceeds 50 cm, rotate the toy in the xy-plane 45° around the z-axis, after have tested the four places on the x,y plane, then test the top and under surface, the microphone should plumb the plane of the test point. When test the noise from the under surface, change the nipping direction of the toy to achieve.

(3) Rattles and squeeze toys: Operate a toy by grasping and mount it 1.2 m above the floor, keep the forearm essentially horizontal which could put the forearm on a bracket. Operate the microphone at the same height as the toy at a distance of 50 cm from the sound source. Measure the A-weighted single-event emission sound pressure level, $L_{pA,1s}$, and the C-weighted peak emission sound pressure level, $L_{pC\,peak}$ for ten movements (squeezes, beats). Carry out the test three times.

(4) Other hand-held toys: First confirm the size of the measurement box according with the reference box of the toy, then nip the test sample on the rotor shaft fixed support. Confirm the centre point of the reference box is coincide with the rotor shaft and the surface of reference box is parallel or vertical with the original place of microphone by moving the

nipping device, then input the lenth, width, height of the measurement box and the height above the floor. The original place of microphone is (0,0,0). When testing, the automatic orientation device will move the microphone on the position “1” in Figure 2 automatically. After this position has tested, rotate the toy for 90°,180°, 270°etc place to continue the test, after have tested the four places on the x,y plane, then do the test about the top and under surface. When test the noise from the under surface, change the nipping direction of the toy to achieve.

(5) Stationary and self-propelled table-top and floor toys: Nip the test sample on the turnplate center of test plane, confirm the center point of the reference box is coinciding with the turnplate center and the surface of reference box is parallel or vertical with the original place of microphone. Before the test, input the length, width, height of the measurement box and the height of the plane surface above the floor, the original place of microphone is (0,0,0). When testing, the automatic orientation device will move the microphone on the position “1” in Figure 3 automatically. After this posision has tested, rotate the toy for 90°,180°, 270°etc place to test respectively. If the length or the width of the toy exceeds 100 cm, then add 4 point into the measurement box to test, such as the “5”, “6”, “7”, “8” point shown in figure 3.

(6) Pull and push toys and hand-activated spring-propelled toys: put the test sample on the test plane; make sure the midline of the toy is superposition to the midline (x-axis) of the test plane. For toys with a width (w) of 25 cm or less, use two microphones at distances (d) 50 cm from the x-axis of the measuring coordinate system, for toys with a width (w) of more than 25 cm, use two microphones at distances (d) 40 cm plus half the width of the toy from the x-axis (40 + w/2) as shown in Figure 3. The original place of microphone is (0), when being test, the automatic orientation device will move the microphone to the appointed position automatically.

7. CONCLUSIONS

In the noise test of sound toys, the accuracy of locating the test point of the microphone, the rapidity of locating the test point and the methods are crucial to the accurateness and the efficiency of the measurement. This automatic orientation device for noise test can suffice the noise test of sound toys conform to ASTM F963-03, EN71-1:2005, GB6675-2003 and ISO8124. The device can fulfill one-off orientation, multiple tests with great accuracy, good repetition and simple operating. So it can promote the level of the automatization and the efficiency of the test.

REFERENCES

[1] Shao Zhongan, Modern measurement technology of acoustics-noise[M], Xi’an: Xi’an Jiao Tong University Press, 2004.
[2] Wang Shuqing, Technology and application of advanced control[M], Beijing: Chemical Industry Press, 2001,7.
[3] BS EN71-1:2005 British Standard. Safety of Toys-Part 1: Mechanical and Physical Propertics[s].
[4] EN ISO 3746: 1996 Acoustice-Determination of Sound Power Levels of Noise Sources Using Sound Pressure[s].
[5] EN ISO 11201: 1996 Acoustice-Noise Emitted by machinery and Equipment[s].

Yin Lijuan is a Professor of Shenzhen Entry-Exit Inspection and Quarantine Bureau. She graduated from Dalian University of Technology in 1984 with specialty of mechanical manufacturing. Her research interests are in mechanical design and manufacturing.

FATIGUE RESEARCH ON T-BEAM WITH RANDOM VEHICLE LOADING SPECTRUM

Wang Xiangyang[1], Zhao Guanghui[2]
School of Transportation, Wuhan University of Technology
Wuhan, Hubei, China
Email: [1]wangxy2003@163.com, [2]guanghui7@163.com

ABSTRACT

Based on the random vehicle loading spectrum and finite element method, this paper analyses the displacements and the fatigue performances in the tensile region of simply supported concrete T-beams, which are reinforced by the prestressed CFRP sheets. And it is showed that under the random vehicle spectrum the stresses decrease after the T-beam has been reinforced by the prestressed CFRP sheets; the fatigue life of T-beams has doubled to 2×107 compared with that before the reinforcement.

Keywords: CFRP Sheet, Random Vehicle Loading Spectrum, Fatigue Analysis

1. INTRODUCTION

Currently, about Carbon fiber reinforced plastic (CFRP) sheets strengthening researches in bridge are as popular as those in the architectural structure. Prestressed CFRP sheets can improve the mechanical behaviors of the members in usage and bear great significance in promoting the researches of this technology and for the engineering practices. Based on those, this paper uses finite element software ANSYS to research the mechanical behavior and fatigue characteristics of reinforced concrete beams strengthened with CFRP sheets.

2. VEHICLE MODEL

Fatigue characteristic about the road and bridges is a cumulative damage process caused by repeated action of loaded vehicles, and in this paper the fatigue design is adopt through the loading spectrum (as showed in Figure 1) which can represent daily vehicles and the traffic census results in Reference [3] are used to constitute fatigue loaded vehicle model, which are used as the fatigue load of T-beam Bridge.

Vehicle type Numbers	Number Of Axles	Diagrammatic presentation: Axle weight/KN; Wheel base/m	Weight/KN	Number of Vehicles	Rate%
M1	2	30, 60; 4.0	90	3 089	9.81
M2	3	30, 60, 60; 3.5, 5.0	150	2 141	6.80
M3	3	45, 70, 70; 4.0, 1.5	185	761	2.42
M4	4	55, 105, 85, 85; 3.5, 7.0, 1.5	330	229	0.73
M5	5	55, 100, 100, 110, 110; 3.0, 1.5, 7.5, 1.5	475	106	0.33
M6	6	50, 105, 105, 90, 90, 90; 3.0, 1.5, 5.5, 1.5, 1.5	530	24	0.08
Total				6 350	20.19

Figure 1. Frequency value spectrum of loaded vehicle model

3. STRUCTURAL MODEL

The researched structure is a T-beam, the standard span L is 20m and calculated span is 19.5m. The structure is shown in Figure 2 (Reference [4]). Design load: automobile-20, trailer-100 and crowd load 3kN/m^2. The concrete strength is C30; the main steel bar adopts as Class II. CFRP sheets of 0.5mm in depth are used to strengthen the mid-span. The finite element model with CFRP sheets is shown in Figure 3.

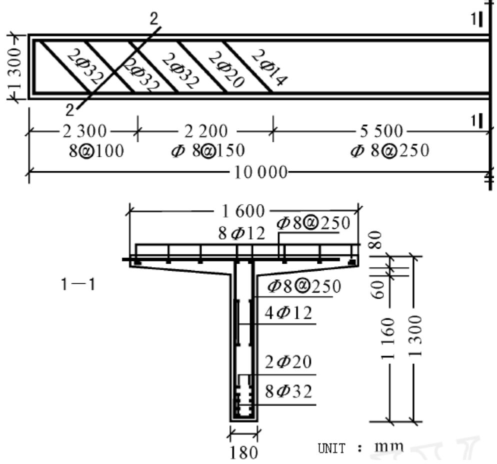


Figure 2. The T-beam size

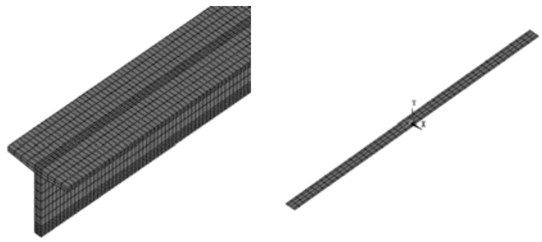


Figure 3. T-beam model with CFRP sheets

4. NON-LINEAR STATIC ANALYSIS OF T-BEAM

This paper considered the non-linear static calculation of the T-beam with the load of automobile-20 and trailer-100 (the impact coefficient not taken into consideration). Before reinforced, loaded with automobile-20, the maximum mid-span displacement is 12. 3mm. compared with the Reference [3], which are 12. 8 mm, the error is 4 %. Loaded with trailer-100, the maximum displacement of is 22.2mm. Compared with the Reference [3], which are 22.6mm, the error is 1.5%.according to that, the finite element model of the T-beam is accurate.

For T-beams strengthened with prestressed CFRP sheets, under the load of automobile-20 and trailer-100, the maximum mid-span displacements are 11.2mm and 20.1mm respectively, a decrease of 8% and 9% compared with that before reinforcement. It is concluded that the prestressed CFRP sheets reinforcement can decrease the deflection of T-beams.

5. DISPLACEMENT UNDER LOADING SPECTRUM

The vehicle loads in Figure 1 are used to work on the T-beam to analyze the mechanical characteristics**.**

When the vehicle model passes the T-beam Bridge at the steady speed of 10km/h, the displacements of the beam under the vehicle dynamic load are greater than that under the automobile-20 and the trailer-100, the increase is 20.1mm and 9.0mm respectively. And strengthened with prestressed CFRP sheets, the maximum displacements of the mid-span are calculated and it is showed that the displacement under dynamic load is greater than that under the automobile-20 and trailer-100, the increase are 15.7mm and 5.7mm respectively.

6. FATIGUE ANALYSIS

Currently it is thought about that the fatigue life of reinforced concrete structure depend on the steel bar. So in this paper for the sake of the fatigue life of T-beam, to consider the loading process of vehicle model, and use the rain flow count method to adopt the frequency of different kinds of vehicles in Figure 1. In this way, we have received the average tensile stresses amplitude of bridges steel bars for 24h before and after the reinforcement (Table 1).

According to the reinforcement broken double logarithmic fatigue formula and Palmgren2Miner's Criterion on the Linear Damage Accumulation in Reference [5], it can be found, before reinforcement and under the average cyclic, the stresses amplitude have been listed in Table 1, circulating 2×106 and 2×107 , the cumulative damage value D is 0.114 and 1.138 respectively, and so the fatigue life of the steel bar is between 2×106 and 2×107, and the fatigue life may decrease as the service time and vehicle load increase. Nevertheless, when the structure has been reinforced by CFRP sheets, under the average cyclic stress amplitude, circulating 2×106 and 2×107 the cumulative damage value D is 0.046 and 0.462 respectively, the fatigue life of the steel bar is between 2×106 and 4×107. So we can conclude that compare with non-reinforced beams the fatigue life of the strengthened T-beams have doubled.

By comparison it can be seen that the concrete beams strengthened with CFRP sheets has changed the dynamic characteristics of concrete beams. Under the action of vehicle spectrum load, the characteristics of CFRP are given full play to. Comparing CFRP strengthened concrete beams with those non-strengthened beams it can be found that under the action of vehicle load, the CFRP sheets play a significant role in decreasing concrete compressive stress range and the stress range of the steel bar and therefore effectively prolong the fatigue life of the concrete beams.

Table 1. Average tensile stresses

Vehicle number	Average cyclic stress amplitude / MPa		Fatigue life /N		Proportion to daily traffic flow / %	daily traffic flow /veh
	Not reinforced	reinforced	Not reinforced	reinforced		
M 1	20.48	19.30	3.00×10^{10}	4.38×10^{10}	48.64	3089
M 2	41.94	37.25	3.09×10^{8}	6.59×10^{8}	33.71	2141
M 3	56.73	51.89	4.50×10^{7}	7.95×10^{7}	12.00	761
M 4	88.42	70.15	4.09×10^{6}	11.07×10^{6}	3.62	229
M 5	128.70	110.66	0.79×10^{6}	1.50×10^{6}	1.64	105
M 6	180.41	138.14	0.17×10^{6}	0.58×10^{6}	0.40	24

7. CONCLUSIONS

Finite element analysis shows that the concrete beams strengthened with CFRP sheets improve the carrying capacity and ductility of the structure. CFRP sheets are effective in strengthening the structure under the random vehicle loading spectrum. The analysis indicates that as CFRP undertakes forces, it decreases the stress of the steel bar in the beams, improves the force condition, and increases the fatigue performance of the structure.

Through the spatial finite element analysis of the concrete beams of different structures and forms it can be concluded that:

(1)Non-linear static finite element analysis of concrete beams shows that after bonding with prestressed CFRP sheets, the compressive stress of the section is increased while the tensile stress is decreased. This indicates that the carrying capacity of the structure is improved.

(2)The fatigue life of CFRP strengthened concrete T-beams shows that concrete beams strengthened with prestressed CFRP sheets effectively decrease the fatigue stress range of the tensile steel. This proves that CFRP sheets are effective in strengthening the structures under fatigue load.

REFERENCES

[1] Chinese National Standards. (2002). Concrete Structure Design Specification (GB50010-2002. Beijing, People's Republic of China Ministry of Construction.

[2] Concrete Structure Design Specification (GB50010-2002. Beijing, People's Republic of China Ministry of Construction

[3] JTJ 023-85, Code for design of highway reinforced concrete and prestressed concrete bridges[S]

[4] Tong L.W, Shen Z. Y. and Chen ZY. (1997). Fatigue Load Spectrum for Urban Road Bridges [J].China Civil Engineering Journal, 30 (5):20-27

[5] Yi J.G. (2001). Simply Supported Girder Bridge (edition 2). Beijing, China Communications Press

[6] Zeng Z. B. and Li Z.R. (1999). Research on Fatigue S-N Curves of Reinforcing Bars in Common Reinforced Concrete Beams. China Civil Engineering Journal, 32 (5):10-14

PROCESSED DATA OF THE INDUCING GRAVITY MEASURED DURING TOTAL SOLAR ECLIPSE

Yisheng Tang [1]   Shouyuan Chen [2]
[1]LinYi Vocational College LinYi, 276017
[2] Shandong Normal University, 250014, Jinan ,China
Email: chensy@beelink.com

ABSTRACT

The temporal variation in the gravity field during total eclipse occurs with the onset of the solar eclipse and the departure. The principle of temporal variation in gravity field maybe similar to Faraday's electromagnet induction, the change of motion induces force field.

Keywords: Gravitation variation, Solar eclipse, Induction gravity

1. IMPORTANT INFORMATION

The temporal variation in the gravity field during total eclipse measuered, in 1995, 1997, 2000, 2001, 2003, and will be 2009. A total solar eclipse on 24 October 1995 starting from sunrise at Iran and ending at sunsetat the Pacific Ocean provides a 46km wide strip for approximately 1800km in India from Nem Ka Thana (Western Rajasthan) to Diamond Harbour (West Bengal) where the total solar eclipse was observed for some time between 7:22 am to 10:30 am (local time). The following figure illustrates this total solar eclipse: [1]. The temporal variation of gravity indicates the onset and departure of solar coverage between approximately 6:30 and 7:00 am local. The horizontal axis is time in hours (5-11 am), and the gravity scale of the vertical axis is each tick mark (as 2x10^-8 cm/s^2).

On Sunday, 1997 March 9, a total eclipse of the Sun was visible from parts of eastern Asia. The path of the Moon's umbra shadow begins in eastern Kazakhstan, and travels through Mongolia and eastern Siberia where it swings northward to end at sunset in the Arctic Ocean. A partial eclipse will be seen within the much broader path of the Moon's penumbral shadow, which includes eastern Asia, the northern Pacific and the northwest corner of North America.

The gravity variation accurs on total eclipse of 2001June 21 and so on .The principle of temporal variation in gravity field cab be problem[2][3].

2. TEMPORAL VARIATION MEASURED IN GRAVITY FIELD DURING SOLAR ECLIPSE

2.1. Gravity Field during Solar Eclipse in 1995

During this period of solar eclipse, Indian scientisiters was at Dhoraji(22 ° 44'N, 70 ° 27'E) .This region falls in approximately 80% of the total eclipse. They recorded temporal variation in the gravity field at this place continuously for approximately 12 h before and after the eclipse.

Figure. 1 observed (top line uncorrected) and tidal corrected gravity field at Dhoraji on 24 October 1995. The temporal variation of gravity indicates the onset and departure of solar coverage between approximately 6:30 and 7:00 am local. The horizontal axis is time in hours (5-11 am), and the gravity scale of the vertical axis is each tick mark (as 2x10^-8 cm/s^2).

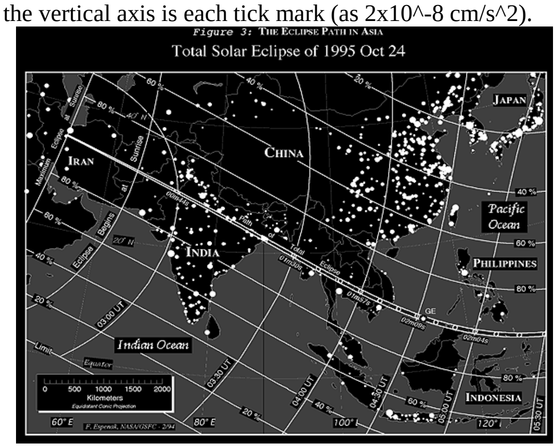


Figure.1 Temporal Variation in Gravity Field during The Solar Eclipse

2.2 Temporal variation in gravity field during Solar Eclipse of 1997 March 9

On Sunday, 1997 March 9, a total eclipse of the Sun was visible from parts of eastern Asia. The path of the Moon's umbra shadow begins in eastern Kazakhstan, and travels through Mongolia and eastern Siberia where it swings northward to end at sunset in the Arctic Ocean. A partial eclipse will be seen within the much broader path of the Moon's penumbral shadow, which includes eastern Asia, the northern Pacific and the northwest corner of North America[1].

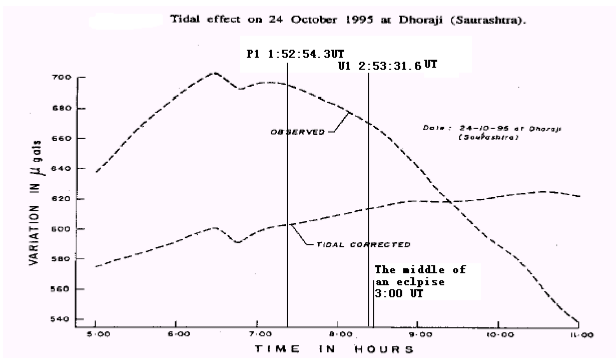


Figure 2. Observed and tidal corrected gravity field at Dhoraji on 24 October 1995

The variations of gravity were measured with a high precision LaCoste-Romberg D gravimeter during the total solar eclipse of 1997 march 9 to investigate the effect of a solar eclipse on the gravitational field. The observed anomaly $(7.0\pm2.7)\times10^{-8}$ m/s^2during the eclipse implies that there may be a shielding property of gravitation. This observation and measurement during the total eclipse were carried out in Moho, China (53°29'N, 122°20'E). we can see that the temporal variation in the gravity field occurs with the onset of the solar eclipse and the departure .

In Figure 3. the horizontal axis is time in hours(Beijng local time);and gravity scal of the vertical axis is each tick mark (as 2x10^-8 cm/s^2);

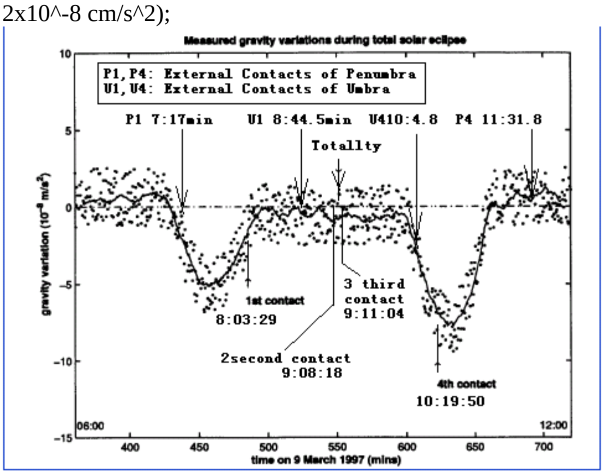


Figure 3. Temporal Variation in Gravity Field during Solar Eclipse on 9 March 1997

P1 - Instant of first external tangency of penumbral shadow cone with Earth's limb. (partial eclipse begins)
P4 - Instant of last external tangency of penumbral shadow cone with Earth's limb. (partial eclipse ends)
U1 - Instant of first external tangency of umbral shadow cone with Earth's limb.
(umbral eclipse begins)
U4 - Instant of last external tangency of umbral shadow cone with Earth's limb.
(umbral eclipse ends)
First Contact - Instant of first external tangency between the Moon and Sun.
(partial eclipse begins)
Second Contact - Instant of first internal tangency between the Moon and Sun.
(central or umbral eclipse begins; total or annular eclipse begins)
Third Contact - Instant of last internal tangency between the Moon and Sun.
(central or umbral eclipse ends; total or annular eclipse ends)
Fourth Contact - Instant of last external tangency between the Moon and Sun.
The gravity variation accurs on total eclipse of 2001June 21, to see figure 4

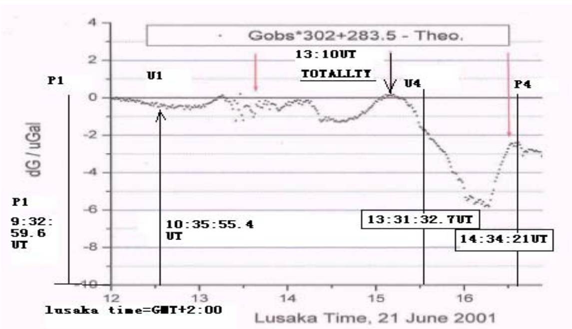


Figure 4. Temporal Variation in Gravity Field during Solar Eclipse on 21 June 2001

The gravity variation will be measured on total eclipse of 2009 July 22. The total eclipse will accur in the Changjing River valley of China.

3. FARADAY'S LAW OF INDUCTION

Faraday made the first step in the discovery of the great law of electromagnetic induction that new bears his name. The electric current was not caused by the magnetic field, but by a change in the magnetic field. We have already seen that faraday associated an electric current with a changing magnetic field; a change in the magnetic flux through a circuit produces an electromotive force around the circuit. Faraday’s law of electromagnetic induction is :

$$emf = -\frac{\phi - \phi_0}{t - t_0} \quad (1)$$

The Faraday's law is the changing magnetic field creates electric field. In this picture an ammeter is connected in the circuit of a conducting loop. When the bar magnet is moved closer to, or farther from, the loop, an electromotive force (emf) is induced in the loop. The ammeter indicates currents in different directions depending on the relative motion of magnet and loop. Notice that, when the magnet stops moving, the current returns to zero as indicated by the ammeter.

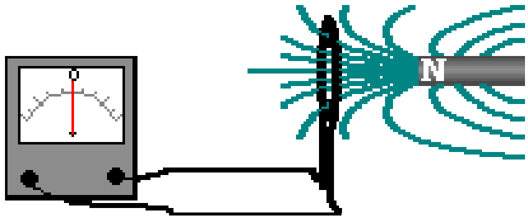


Figure 5. Faraday's Magnetic Field Induction Experiment

A magnet thrust into a coil produces a noticable deviation on a table galvanometer, or the coil can be moved over the magnet.

4. THE TEMPORAL VARIATION OCCURS WITH THE ONSET OF THE SOLAR ECLIPSE AND THE DEPARTURE

A total solar eclipse requires the umbra of the Moon's shadow to touch the surface of the Earth. Because of the relative sizes of the Moon and Sun and their relative distances from Earth, the path of totality is usually very narrow (hundreds of kilometers across). The following figure illustrates the path of totality produced by the umbra of the Moon's shadow.

Two lines or more could be occupied for a long title of paper. However the distance between lines should be zero.

The temporal variation in the gravity field occurs with the onset of the solar eclipse and the departure. As the moon thrusts between Earth and Sun. the temporal variation occurs. Or the moon can be moved out(to see Figure 6).

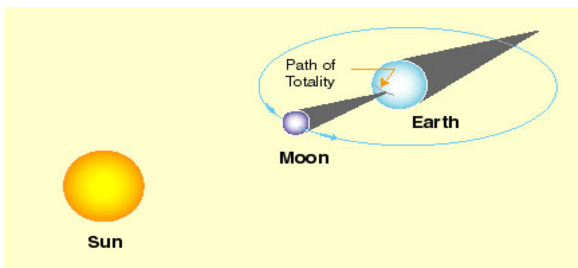


Figure 6. Solar Eclipse (not to scale)

5. PRINCIPLE OF GRAVITY VARIATION

Faraday 's induction law: when the change of electricity is induced induction fields. Newton second law: the change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed. We discuss when the motion state of any object changes, is the induction field inducing?

Any body is moved closer to, or farther from, the gravity loop, an induction force (so called induction gravity) is induced in the loop. To see Fig7.

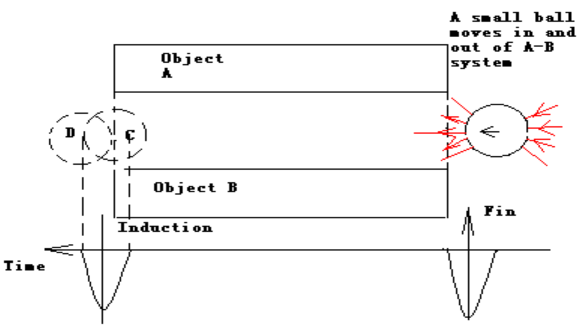


Figure 7. A small ball moves in and out gravity field, inducing gravity occurs

Suppose: gravity system combines A object and B, the moving small ball affects to the gravity system by a force field. This field is called U-field: its field intensity is in direct proportion to its momentum and is in inverse proportion to the square of distance, and its direction is the same as motion direction。, all objects affected by U-field react in return on the moving body, U-field of moving object: U=K（mv/R^2）.Then temporal variation of gravity field is detected in total solar eclipse by Moon moving in or out of the Sun-Earth system induction. This maybe similar to Faraday's Law of Induction. that is a changing U-field field through a closed loop of gravity system induces an induction gravity Fin in the space. The U-field flux through a S is:

$$\Phi(t) = \int_s uds \qquad (2)$$

When the small ball moves in or out of the A-B system, the induction gravity is detected to the change of U-field :

$$F_{IN} = -\frac{d\Phi}{dt} = -K\frac{d}{dt}\int_s U \cdot dS \qquad (3)$$

We think the temporal variation in the gravity field during total eclipse occurs with the onset of the solar eclipse and the departure. the principle of temporal variation in gravity field maybe similar to Faraday's electromagnet induction, the change of motion induces force field.

6. MEASURED DATA PROCESSING

Table 1 measured data on Fig. 2 Temporal variation in gravity field

number	Time on Figure 4 (min)	Gravity on Figure 4	Time measured data (min)	Gravity variation
1	59	136	429、41	-0、22
2	60	133	430、59	-0、54
3	63	129	434、12	-0、98
4	62	127	432、94	-1、20
5	65	122	436、47	-1、74
6	65	118	436、47	-2、17
7	67	116	438、82	-2、39
8	68	110	440、00	-3、04
9	68	104	440、00	-3、70
10	72	100	444、71	-4、13
11	73	95	445、88	-4、67
12	74	96	447、06	-4、57
13	76	93	449、41	-4、89
14	77	93	450、59	-4、89
15	82	93	456、47	-4、89
16	85	95	460、00	-4、67
17	91	99	467、06	-4、24
18	93	102	469、41	-3、91
19	95	104	471、76	-3、70
20	96	109	472、92	-3、15
21	98	111	475、29	-2、93
22	100	115	477、65	-2、50
23	102	117	480、00	-2、28
24	103	121	481、18	-1、85
25	104	122	482、35	-1、74
26	105	127	483、53	-1、20
27	109	133	488、24	-0、54
28	112	137	491、76	-0、11

$$Y_i = AX^2_i + BX_i + C \qquad (4)$$
$$F(A,B,C)=\Sigma(AX^2_i + BX_i + C - Y_i)^2 \qquad (5)$$
$$\Sigma(AX_i^2 + BX_i + C - Y_i)^2 * X_I^2 = 0 \qquad (6)$$
$$A\Sigma X_i^4 + B\Sigma X_i^3 + C\Sigma X_i^2 = \Sigma X_i^2 Y_i \qquad (7)$$
$$A\Sigma X_i^3 + B\Sigma X_i^2 + C\Sigma X_i = \Sigma X_i Y_I \qquad (8)$$
$$A\Sigma X_i^2 + B\Sigma X_i + C = \Sigma X_i \qquad (9)$$

On Figure 4,the left inducing variation in gravity field is commupted:

$$Y = AX^2 + BX + C \qquad (10)$$

A=0.0071; B=-8.9002; C=2791.3

On Figure 4,the right temporal variation in gravity field is computed:

$$Y = AX^2 + BX + C \qquad (11)$$

A=0.0056; B=-5.1272; C=11707.7

Summary in this paper , we discussed induction gravity during solar eclipse. The model and data measured be identical.

REFERENCES

[1] NASA WWW
[2] D.C. Mishra, M.B.S. Rao, National Geophysical Research Institute, Current Science, 72 (11) 1997 (783)
[3] Qian-shen Wang, Xin-she Yang. Precise measurement of gravity variations during a total solar eclipse[J]. Physics Review 2000,D(62):41101～41103

ACTIVE DOCUMENT WAREHOUSE MODEL FOR DISTRIBUTED SOFTWARE DEVELOPMENT BASED ON KNOWLEDGE INTEGRATION AND KNOWLEDGE DIFFUSION

XIA Huo-song
School of Economics and Management, Wuhan University of Science and Engineering, P.R.China, 430073
Email:xiahuosong8@gmail.com

ABSTRACT

This paper presents the concept of active document warehouse. Patterns using active document warehouse for software development based on knowledge integration and knowledge diffusion is descript.

Keywords: Active document warehouse model, knowledge integration, Knowledge diffusion

1. INTRODUCTION

A conservative estimate of the annual cost for failed software projects in the US is $60 to $70 billion. While current literature on marketing/product management focuses on optimizing revenue, the literature on product line engineering focuses on lowering costs through reuse. (Helferich, Schmid & Herzwur, 2006).

We must also analyze what problems typically arise in software development. The following 12 reasons are among the most cited factors contributing to software failure (Charette, 2005):

1. Unrealistic or unarticulated project goals
2. Inaccurate estimates of needed resources
3. Badly defined system requirements
4. Poor reporting of the project's status
5. Unmanaged risks
6. Poor communication among customers, developers, and users
7. Use of immature technology
8. Inability to handle the project's complexity
9. Sloppy development practices
10. Poor project management
11. Stakeholder politics
12. Commercial pressures

For improving distributed software development purpose in 21^{st} century, we should build a good framework for 21^{st} century software development. Especially, active document need to be good controlled and managed.

2. METHODOLOGIES

When searching, analyzing, or transferring documents from the Intranet, users deal with large document collections composed of multiple types documents from various document resources in software development processes. Although some tools such as lotus for document management tool, literature retrieved tools for categorizing document tools, can easily be browsed, sorted and filter them. These documents need to be improved the efficiency and reusability of users' work with semantically-based active document management systems. It is very hard to achieve any significant improvement in software development organizational performance without the integration of people and processes of software development. However, there are not many mature methodologies or models that deal with active documents management in distributed software development. Based on the literature review and interview domain experts and developers in practice, we develop available patterns for software development and a model of active document warehouse model in 21st century software development.

3. PATTERNS FOR SOFTWARE DEVELOPMENT BASED ON ACTIVE DOCUMENT MANAGEMENT

We know that structured data are stored in data warehouses for decision makers to use. However, to distributed software development, unstructured data such as emails, html texts, images, videos, and office documents are increasingly accumulated in software project. Such unstructured data, or what we call multimedia documents, are larger in volume than structured data. So we need a document warehouse for distributed software development where multimedia documents are analyzed and managed for developer wide knowledge sharing and reuse like a data warehouse for structured data. A prototype document warehouse system, which supports management of simple and compound documents, keyword based and content based retrieval, rule based classification, SOM based clustering, and business rules, is described (Ishikawa, H.; Kubota, K.; Noguchi, Y.; Kato, K.; Ono, M.; Yoshizawa, N.; Kanaya, A.(1998)).

Based on the idea of document warehouse system, we can describe active document management, which manages simple and compound documents with time-based. Because time wasted looking for documents is money wasted and opportunity lost. We define that active document is the document with time-based, reused, and revising based content.

Active document warehouse can be defined that it is consist of fact table linking many dimension tables for storing active documents. We can build active document warehouse like data warehouse, but it is more complex than data warehouse. It is need to consider the characters of active document.

In this section, patterns for distributed software development based on active document management are presented. The patterns are based on 6 major dimensions which are the same as the core processes in software development: (i) requirements, (ii) design, (iii) coding, (iv) testing, (v) implementation, (vi) service. The patterns are summarized in table 1.

Table 1. Patterns for Software Development Based on Active Document Management

Major processes	pattern	description
Requirements	Strategies	User-oriented, IT Balance score card (IT BSC), Active document management
	Techniques	DFD,DD, Logic express, output documents, ontology
	Tools	Vision, word, CASE
Design	Strategies	Concurrent engineering, CMMI, Active document management
	Techniques	Robust design, Active document technology
	Tools	Rapid prototyping, Case
Coding	Strategies	Knowledge sharing and diffusion, CMMI, Active document management
	Techniques	Component, reuse, module
	Tools	Middle ware, Mashup, API, JSP, .NET, PB, Delphi, C, Java
Testing	Strategies	TQM,CMMI, Active document management
	Techniques	Internet, Agent
	Tools	Sample data tool, CBR
Implementation	Strategies	Business process management, Field expert-oriented, Active document management
	Techniques	Data management, document management
	Tools	Internet, CBR
Service	Strategies	B2B, B2C,CRM, Active document management
	Techniques	Internet,WWW,Web2.0
	Tools	Internet, groupware, shareware, PDA, messages

Knowledge and active documents transfer and diffusion are the key factors for the success in requirements process. There are several strategies that have been employed for knowledge and active documents management. These include: user-oriented, IT balance score card, and active document management. Basically, these strategies bring the people together including users to participate in the requirements process. Several techniques can be used for knowledge and active document collection, and diffusion which include: DFD, DD, logic expresses, out documents, and ontology. Tools such as Version, word, and CASE can be used to fill the gap between the requirements process and design process.

Now the design process in software development has been transformed into a standard-based function wherein some components have been outsourced. The strategies that could be employed for the design process include: concurrent engineering, CMMI, and active document management. Techniques such as robust design, active document technology can be employed in design process. There are several tools available for this purpose, some of which are Rapid prototyping, and Case.

In coding process, knowledge sharing and diffusion, CMMI, active document management plays a major role in this strategy. Techniques such as component, reuse, and module can be employed. There are several tools such as middle ware, Mashup, API, JSP, .NET, PB, Delphi, C, and Java in coding process.

There are several strategies such as TQM, CMMI, active document management in testing process. Techniques include Internet and Agent can be employed. There are several tools such as sample data tool and CBR.

4. A MODEL OF ACTIVE DOCUMENT WAREHOUSE OF SOFTWARE DEVELOPMENT

Define 1, Let ADW be the set of all active document available for the distributed software development. We define ADW+: ADW+=ADW∪ {ADWi | ADWi is implied by ADW}, and ADW is document can be used by project of distributed software development in document warehouse. It's required to explore the description methods of the active document warehouse.

We can describe active document warehouse in sets as follows:

ADW={Fact table}∪{ DTi |Dimension table} (1)

{Fact table}={document id(c(30)), author id(c(20)), type id(c(6)), press id(c(12)) ,language id(c(2)), link id(c(50)), diffusion id(c(50)), store id(c(50)), time id(c(8)), module id(c(13)), character id(c(13)), file name(c(50))} (2)

{Dimension table of diffusion}={diffusion id(c(50)), download(t(50)), print(t(50)), show(t(50))} (3)

{Dimension table of author}={author id(c(20)), name(c(50))} (4)

{Dimension table of character}={character id(c(13)), topics(t(50)), abstract(t(50)), keywords(t(30)), title(c(50))} (5)

{Dimension table of time}={time id(c(8)), F_name+time(t(50)), year(n(4)), month(n(2)), day(n(2)),), order(c(2))} (6)

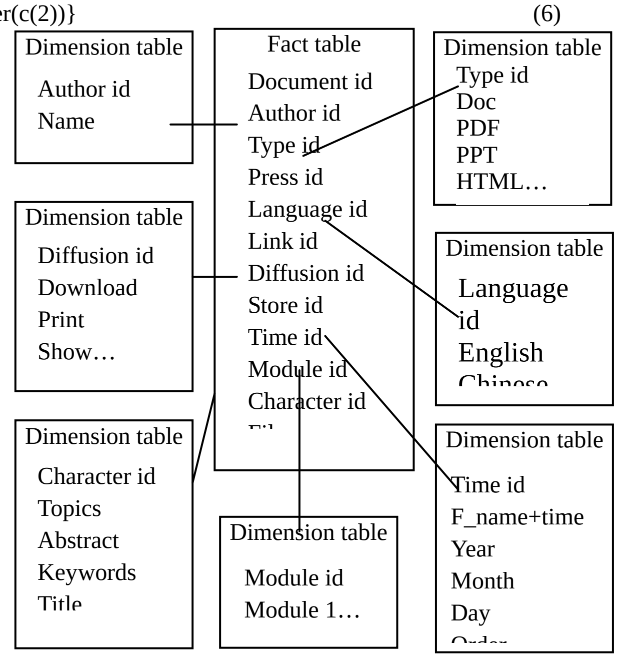


Figure1. Active Document Warehouse Model for Software Development

5. IMPLEMENTATIONS

In recent years, we have developed a MIS project for Wuhan Insurance Center for New Car from 2005. With the new policy developing, the MIS project needs to be developed. We have applied active document model for meeting the new requirements based on virtual folder and document warehouse as the figure 2.

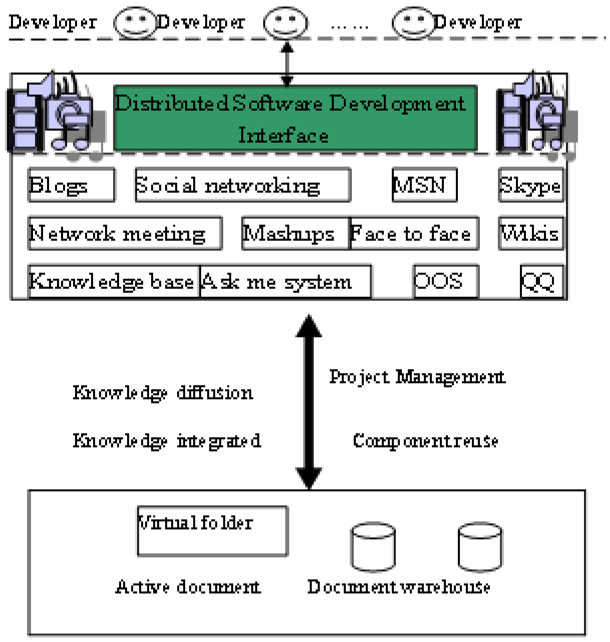


Figure 2. Implementations for Software Development Based on Active Document

All documents are become active document with time in file name and built Intranet for knowledge diffusion and knowledge integration which many electronic compound documents that involve multiple types of media such as images, video clips, word-processed text, graphics, and spreadsheets were shared. And the document management has become joint from all developers based on collaborative distributed software development interface for receiving componets reuse on virtual or document warehouse from developers and Server via Intranet. A prompt response to a request is needed to maintain customers' satisfaction. Therefore, the model based on active document can support developers sharing knowledge and documents.

6. CONCLUSIONS

This model is a benefit to the developer and MIS user. To managing active document, this paper presents good patterns and active document warehouse model. Applying this idea to MIS project for Wuhan Insurance Center for New Car, it is tested well. However, it is necessary to develop a better, more effective and more details model of active document management in distributed software development with trusted environment.

ACKNOWLEDGEMENT

This research is supported by National Philosophy and Social Science Fund (07BTQ010) of China, CNATC(2007082); HuBei Province Fund of China(Z20091701,2008d062,2008244,2007097,20061003106), Wuhan Science and Technology Bureau(200940833384-02),WUSE Fund (20081507). We deeply appreciate the suggestions from fellow members of the 24-Hour Knowledge Factory project team.

REFERENCES

[1] Charette, R.N. (2005). "Why software fails (software failure)." Spectrum IEEE 42(9), 42-49

[2] Helferich, A., Schmid, K. and Herzwur G. (2006). "Product management for software product lines: An unsolved problem." Communications of the ACM. 49 (12), 66-67.

[3] Baresi, L., Nitto, Di E.. and Gbezzi C. (2006). "Toward Open-world software: issues and challenges." Computer Magazine. 10, 36-43

[4] Xia,H.S. and Gupta, A. (2008). A Common Knowledge Sharing Model of 24-Hour Knowledge Factory of Grid Computing based on Case Based Reasoning. International Journal of Knowledge Management, Volume 4(3)1-18

[5] Gupta, A., Seshasai, S., Mukherji, S., Ganguly, A. (2007). "Offshoring: The transition from economic drivers toward strategic global partnership and 24-Hour Knowledge Factory." Journal of electronic commerce in organizations, 5 (2), 1-23.

[6] Gupta, A. and Seshasai, S.(2007). "24-Hour Knowledge Factory: using Internet Technology to leverage spatial and temporal separations." ACM Transactions on Internet Technology. 7 (3), 1-22.

[7] Oshri, I., Kotlarsky, J. and Willcocks, L. (2007). "Global software development: Exploring socialization and face-to-face meetings in distributed strategic projects." Journal of strategic information systems 16, 25-49.

[8]Ishikawa, H.; Kubota, K.; Noguchi, Y.; Kato, K.; Ono, M.; Yoshizawa, N.; Kanaya, A.(1998).A document warehouse: a multimedia database approach. Database and Expert Systems Applications, Proceedings Ninth International Workshop on Volume, Issue, 25-28 Aug 1998 Page(s):90 - 94

Huosong Xia is a professor in the Department of information management and information systems of the Wuhan University of Science and Engineering. He was a visiting scholar at Eller College of Management of the University of Arizona, USA in 2005. He received his Ph.D. degree in management science and engineering, and his M.S. degree in MIS from Management School, Huazhong University of Science and Technology. He received his B.S. degree in mathematics from Huazhong Normal University. His main research interests are: data mining and knowledge management, MIS, DSS, e-Commerce and Logistics Information System. His publications have appeared in refereed international journals such as Journal of Knowledge Management, International Journal of Knowledge Management, Journal of Systems Science and Information, Journal of Grey System. E-mail Box: bxxhs@sina.com; xiahuosong8@gmail.com

THE STUDY OF GEOTECHNICAL ENGINEERING ON DISTRIBUTED PARALLEL COMPUTATION *

He Huang , Zhuoqiu Li
School of Science,Wuhan University of Technology
Wuhan, China
Email:hhshy@whut.edu.cn

ABSTRACT

Computation has been widely used in the geotechnical engineering field, while it is incompetent in real-time computation. In this paper, a System based on distributed parallel computation named DPC is introduced. This system adopts non-overlapped domain decomposition method, and with the method of adopting MPI on the COW platform, a parallel FEM codes can be implemented by the preprocessing conjugate gradient algorithm by using C++ Object-Oriented Programming (OOP) technique. The large-scale geotechnical problems can be successfully solved on PC machine groups by using these techniques. The instance of Longtan high slope is shown to prove the applicability of these methods. The results show the feasibility of the proposed methods for the extensive application of parallel FEM. This study will make it possible for parallel FEM to be successfully applied to geotechnical engineering and will give some useful references to the farther researches.

Keywords:Geotechnical Engineering, Domain Decomposition, MPI, COW, Parallel FEM.

1. INTRODUCTION

Three-dimensional analysis and the nonlinearity of the governing differential system increase the amount of computation [1]. The requirement of real-time computation is beyond the capabilities of any single personal computer. However the distributed parallel computation can successfully achieve the issue.

In this paper, a synthetically system of distributed parallel computation is proposed to satisfy the requirement of the geotechnical engineering, providing a useful and effective analysis platform for slope design and construction. And a large scale integrating analysis software named DPC (distributed parallel computation) with self-contained copyright has been developed, using the programming tools such as Visual C++.

Finally, the DPC is applied to one typical slope project, the Long tan hydropower station. The results indicate that the system has both important academic value and significant applicable value in actual projects.

2. SKELETON OF THE SYSTEM

According to the demand of engineering practice, the visual parallel system mainly includes three parts: the server, workstations, and intranet visiting computers. Connected by switch, the server and workstations make up the intranet/COW (Cluster of Workstation network). And the intranet computer can get access to the data on the intranet server through firewall.

Figure. 1 shows the flow chart of the implementation of the system. In the environment of distributed parallel computation, the server acts as the intranet web server, as well as the role of data distribution and retrieval in the whole process for the COW.

This system is composed of three parts: preprocessing of the input of the finite element, FEM parallel computation and post-processing of the finite element. Preprocessing of the input of the finite element is a process in which the data of the strata information set up through geographic information system are via domain decomposition method mapped to several node machines. FEM parallel computation is the process of implementing distributed computation by the means of message passing through DPC network computing platforms. Post-processing of the finite element is a visualized post-processing in which the data of FEM parallel computation are recycled into the host computer, the boundary data are collected and averaged.

After domain decomposition [2-3], the data is successfully distributed to every local workstation in the intranet. And then, each workstation stores individual data and begins to commutate and communicate with each other as the parallel computation. That is distributed parallel computation. The server retrieves the data after the computation, and then issues it through the COW.

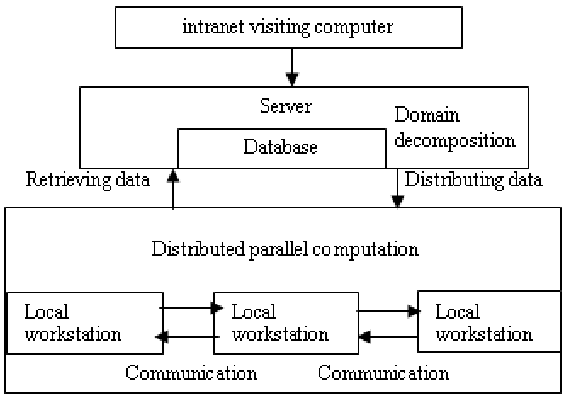


Figure1. The Flow Chart of the System

3. IMPLEMENTATION METHOD OF PARALLEL COMPUTATION

The researchers of rock mechanics implement various mathematical and mechanical methods and endeavor to solve the rock mechanics problems in engineering. Due to the complexity of rock mass structures and engineering factors, quite a number of engineering problems in rock mechanics cannot be solved simply in terms of the analytical method.

Numerical analysis method, therefore, turns out to be the

* The paper is supported by the national natural science foundation of China (10672128) and the major science and technology fund of the Hubei province of China (2007DA111).

effective tool to solve rock mechanics and engineering problems. Geotechnical engineering numerical analysis method is an approach to implementing numerical simulation to the influence the mechanical behaviors and engineering activities of the rock mass and its reinforced structures have on their surrounding environments. Finite element method is capable of dealing with various nonlinear problems, and simulating the complicated construction processes of geotechnical engineering in a flexible manner. It hence becomes the most widely-applied numerical analysis method in the rock mechanics field. The introduction of the parallel method, meanwhile, enlivens the finite element computation.

As for the computation of parallel finite element, it needs various strata data of geology based on the truly three-dimensional Digital Evaluation Model (DEM). In the server, the prototype system uses the database server has been to manage the attribute data and multi-media content (e.g. image files). The data includes various data which meet different needs. It is very complicated of various tasks such as the field investigation of geology, rainfall, excavation conditions, risk assessment of landslides, feedback analysis of design, intelligent recognition of slip plane and mechanical parameters and so on.

As a result of the constant efforts of the research group, we make use of the C++ object oriented mode and successfully interface it with the other systems. This chapter implements object-oriented programming, stands on the basis of integrated rock mechanics parallel computing environment with self-contained copyright （DPC）, and creates a mass geotechnical parallel finite element system. The establishment of this system enables the problems of large-scale finite element computation in PC clusters to be solved in terms of parallel solution, and enhances the computational speed, scale and precision.

This system adopts non-overlapped domain decomposition method. A large computational domain Ω of a finite element can be decomposed into s sub-domains Ω_i（i =1，2，…，s）. The general finite element equation is equation（1）.

In the stiffness matrix, the superscript *i* of A_{II}^{i} denotes the number of the sub-domain, and the subscript *II* stands for the stiffness matrix of freedom of inner degrees. A_{IB}^{1} denotes the stiffness matrix of influence the inner degrees of freedom has on boundary stiffness matrix degrees of freedom. And $A_{IB}=A_{BI}^{\mathrm{T}}$. A_{BB} denotes boundary stiffness matrix degrees of freedom. x_I stands for the displacement of the inner degrees of freedom; x_B stands for the displacement of the boundary degrees of freedom; b_I shows the load of the inner degrees of freedom ; b_B indicates the load of the boundary degrees of freedom;

$$\begin{bmatrix} A_{II}^{1} & & 0 & & A_{IB}^{1} \\ & A_{II}^{2} & & & A_{IB}^{2} \\ 0 & & \ddots & & \vdots \\ & & & A_{II}^{s} & A_{IB}^{s} \\ A_{BI}^{1} & A_{BI}^{2} & \cdots & A_{BI}^{3} & A_{BB} \end{bmatrix} \begin{Bmatrix} x_I^1 \\ x_I^2 \\ \vdots \\ x_I^s \\ x_B \end{Bmatrix} = \begin{Bmatrix} b_I^1 \\ b_I^2 \\ \vdots \\ b_I^j \\ b_B \end{Bmatrix} \tag{1}$$

As for each process, such as the j^th process, the equations (2) should be solved.

$$\begin{bmatrix} A_{II}^{j} & A_{IB}^{j} \\ A_{BI}^{j} & A_{BB}^{j} \end{bmatrix} \begin{bmatrix} x_I^j \\ x_B^j \end{bmatrix} = \begin{bmatrix} b_I^j \\ b_B^j \end{bmatrix} \tag{2}$$

And after the expansion of equations (2), the equations are

$$b_I^j = A_{II}^j x_I^j + A_{IB}^j x_B^j \tag{3}$$

$$b_B^j = A_{BI}^j x_I^j + A_{BB}^j x_B^j \tag{4}$$

In the distributed network parallel programming environment, each item within equation (3) is related exclusively to the information within the individual domain. Therefore, computation within the single computer is enough; communication is not necessary. The computational complexity of this process is, however, maximal. equation (4) is concerned with the nodes on the boundary. Compared with the inner nodes, the computational complexity of this part is relatively minimal.

The computation of each domain finished, it is required to update the data between the adjacent domains. What can be implemented is function MPI_Isend( ) and function MPI_Irecv( ) in the MPI function library.

The flow chart of the preprocessing conjugate gradient algorithm based on domain decomposition method is shown as follows:

1. Construct the Type $b_i = A_{ii}x_i + A_{ib}x_b$ and Type $b_b = A_{bi}x_i + A_{bb}x_b$, and prepare the initialization of vectors and matrixes.
2. b_b=*Update(b_b)* vector update operation
3. $x_i^{(0)} = 0$, $x_b^{(0)} = 0$
4. $r_i^{(0)} = b_i$, $r_b^{(0)} = b_b$
5. Solve $M(z_{\mathrm{i}}^{(0)}, z_{\mathrm{b}}^{(0)}) = (r_{\mathrm{i}}^{(0)}, r_{\mathrm{b}}^{(0)})$ parallel preprocessing technique
6. $p_i^{(0)} = z_i^{(0)}$, $p_b^{(0)} = z_b^{(0)}$
7. $\gamma^{(0)} = \text{Inner Product}(z_i^{(0)}, z_b^{(0)}, r_i^{(0)}, r_b^{(0)})$ inner product operation
8. Retrieve k=0，1，2，…
9. $q_i = A_{ii}p_i + A_{ib}p_b$; $q_b = A_{bi}p_i + A_{bb}p_b$ matrix vector multiplication within separate domains
10. q_b^j =Update(q_b^j) vector update operation
11. $\tau^{(k)} = \text{Inner Pruduct}(p_i^{(k)}, p_b^{(k)}, q_i^{(k)}, q_b^{(k)})$ inner product operation
12. $\alpha^{(k)} = \gamma^{(k)} / \tau^{(k)}$
13. $x_i^{(k+1)} = x_i^{(k)} + \alpha_k p_i^{(k)}$, $x_b^{(k+1)} = x_b^{(k)} + \alpha_k p_b^{(k)}$
14. $r_i^{(k+1)} = r_i^{(k)} + \alpha_k q_i^{(k)}$, $r_b^{(k+1)} = r_b^{(k)} + \alpha_k q_b^{(k)}$
15. Solve $M(z_i^{(k+1)}, z_b^{(k+1)}) = (r_i^{(k+1)}, r_b^{(k+1)})$ parallel preprocessing technique
16. $\gamma^{(k+1)} = \text{Inner Product}(z_i^{(k+1)}, z_b^{(k+1)}, r_i^{(k+1)}, r_b^{(k+1)})$ inner product operation
17. If（$\sqrt{\gamma^{(k+1)}} < \text{tolerance}$） exit
18. $\beta^{(k)} = \gamma^{(k+1)} / \gamma^{(k)}$

19. $p_i^{(k+1)} = r_i^{(k+1)} + \beta^{(k)} p_i^{(k)}$,
$p_b^{(k+1)} = r_b^{(k+1)} + \beta^{(k)} p_b^{(k)}$

From the analysis of the above algorithm steps, it is obvious that the main computational complexities of each iteration loop are twice inner product computation of Step 11 and Step 16; matrix vector multiplication of Step 9. Matrix vector multiplication of Step 9 is of the maximal computational complexity, but it needs not to exchange data and can be finished within separate independent processes. And after the computation of Step 9, a margin vector data update operation is usually required, which needs to exchange data. Yet, the vectors updated during this time are few, compared with those in the domains.

Three are two places where communication is required, that is data update and inner product computation.

(1) Communication is required when data are updated.

The update operation in this paper is Update（b_b）. It issues the vectors of the boundary nodes to the adjacent processing, and receives and adds up the vectors of the boundary nodes through the adjacent processing. When the data are updated, function MPI_Isend( ) and function MPI_Irecv( ) in the MPI function library.

(2) Network communication is also required in vector inner product computation.

For example:

$$\tau^{(k)} = \text{Inner Pruduct}(p_i^{(k)},\ p_b^{(k)},q_i^{(k)},\ q_b^{(k)}).$$

Vector p and Vector q both are divided into two parts: p_i、p_b and q_i、q_b. $p_i \bullet q_i$ is concerned exclusively with internal degrees of freedom of the individual domain, and can be computed directly. $p_b \bullet q_b$ is the degrees of freedom of the nodes on the boundary, and must be divided by the times corresponding degrees of freedom are shared to avoid repeated additions. If corresponding degrees of freedom are shared by 2, 3… sub-domains, the times for corresponding degrees of freedom to be shared respectively are 2, 3… In the end, the summarizations of each processor are collected and added up, and function MPI_Allreduce ( ) of the MPI function library are implemented into communication.

Aiming at the calculation of nonlinear displacement of slope deformation, all the strata data are discretized by mesh generation. And two strategies—MPI and OOP—are adopted in the whole parallel computing.

MPICH is a Free Software and is available for most flavors of Unix (including Linux and Mac OS X) and Microsoft Windows. While MPI is a standard for message-passing for distributed-memory applications used in parallel computing. The finite element implementation of coupled differential systems often results in large codes which might become unmanageable if not well designed. The introduction of Object-Oriented Programming (OOP) enables writing of large and complex codes in an efficient, reusable, and extensible way [4-7].

A 3D elastic-plastic parallel finite element analysis program and its preprocessing and post-processing program have been developed and implemented in this system. Users can pick up boundary-line interactively, draw contour and vector chart of stress, strain and displacement in the system by intranet publication.

The original data, including the three dimensional vector data, image data, strata data, geotechnical data and the results of distributed parallel computation, are stored in the linked server. The server can be used to issue the data.

The services are network-accessible programs that use a standardized messaging protocol to communicate with other programs that want to use their functions.

The flexible and robust system can cater to the increasing needs for data of parallel finite element computation. The heterogeneity of the spatial data puts an extra requirement for parallel system solution to be interoperable so that it can benefit a large scale of users.

4. APPLICATION

The application in Longtan hydropower proves the geographic information system based on distributed parallel computation to be effective and reasonable. The Longtan hydropower project, locating on the Red Water River in Guangxi province, is currently the second largest hydropower station excavated in complicated rock masses in China, with the maximum underground workshop cavern (388.5m×30.3m×74.4m) and the highest rolled-concrete dam (216.5m) in the world. The height of the slope to the left bank of the dam is 420m with the excavated elevation increases from 215m to 635m, and the angle of slope varies from 35º to 90º [8].

Through mesh generation, meshes can be expressed as the formations of the necessary finite element meshes in the database. After that the finite element system will be made. The entire calculation region is divided into 282,157 elements and 51,382 nodes.

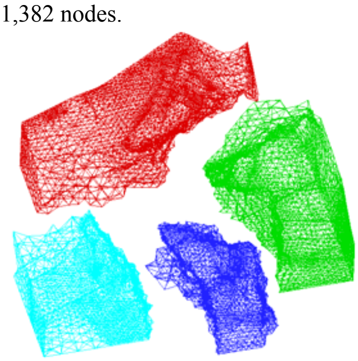


Figure 2. The Multiple Strata Data of the Left Bank of the Dam

The system carries out the domain decomposition of Longtan slopes model into four subdomains, and then distributes them to local workstations for distributed parallel computing. And then, in the environment of distributed parallel computation, each workstation stores individual data and begins to compute and communicate with each other. Fig. 2 shows the meshes data of the 4 subdomains of the left bank of the dam.

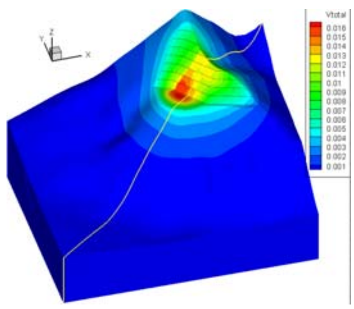


Figure 3. Displacement of Longtan High Slope

After the distributed parallel computing, the server retrieves the data, and then issues it through the intranet. Fig.3 shows the displacement of Longtan high slope in the server.

5. CONCLUSIONS

In this paper, the System based on distributed parallel computation named DCP is introduced. This paper mainly discusses the overall design and algorithm implementation of the integrated system, especially a series of methods to realize distributed parallel computation in intranet.

Finally, the DCP is applied to one typical slope projects，the Longtan hydropower station. The results indicate that the system has both important academic value and significant applicable value in practical projects.

As a result, a new as well as potential approach for the development of DCP has been provided in this paper by adopting the methods of distributed parallel computation and intranet inquiring of the results of intranet computation.

REFERENCES

[1] Feng Xiating, Zhou Hui, Li Shaojun, Sheng Qian, Jiang Quan, “System of intelligent evaluation and prediction in space-time for safety of rock engineering under hazardous environment”, Journal of Rock Mechanics and Engineering, v27, n9, 2008, pp.1741-1756.

[2] James G. Malone, “Automated Mesh Decomposition and Concurrent Finite Element Analysis for Hypercube Multiprocessor Computers”, Computer Methods in Applied Mechanics and Engineering, 70, 1988, pp. 27-58.

[3] Charbel Farhat, “implementation Aspects of concurrent finite element computations in Parallel Computations and their Impact on computational mechanics”, ASME New York, 1987.

[4] D R Rehak， J W Baugh, “Alternative Programming Techniques for Finite Element Program Development”, Proc.， IABSE Colloquium on Expert Systems in Civil Engineering， Bergamo， Italy， 1989

[5] Bruce W R, Forde_Ricardo O Foschi, Siegfried F Stiemer, “Object-Oriented Finite Element Analysis”, Computers & Structures， 1990， 34(3),PP.355-374.

[6] Logozzo F, “Class-level modular analysis for object oriented languages”, In: Proceedings of the 10th static analysis symposium 2003. Lecture notes in computer science, vol. 2694. Berlin: Springer, 2003, pp. 37− 54.

[7] Logozzo F, “Modular static analysis of object-oriented Languages”, PhD thesis, Ecole Polytechnique, June 2004.

[8] Zheng HuiFeng, Wang WeiMing, Chen ShengHong, “Application of block element method to excavation and reinforcement analysis of Longtan slope”, Journal of Dalian University of Technology, v 46, 2006, pp. 178-184.

DEVELOP RFID APPLICATION USING WEBSPHERE RFID SOLUTIONS

Xu JunWu[1], Liang JunLin[2]
[1.] School of Computer Science & Technology, Wuhan Institute of Technology, 430073, China
[2.] School of Automation, Wuhan university of Technology, 430074, China
Email: fiber2001@163.com

ABSTRACT

This article describes research in the use of The IBM Development Kit to develop a Radio frequency identification (RFID) application and. Learn how to integrate the framework with back-end applications and implement business logic. This server allows business-logic integration of the information with back-end EAI systems of connecting disparate systems, such as Supply Chain Management (SCM), Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP).

Keywords: RFID, IBM WebSphere Premises Server, SOA, ERP, CRM, SCM

1. INTRODUCTION

RFID is a wireless system used to identify tags. These tags may be carried by people or animals or mounted on object or vehicles. They may even be embedded under the skin. RFID tags are non-contact and non-line-of-sight. Figure 1 shows a typical diagram of an RFID framework.

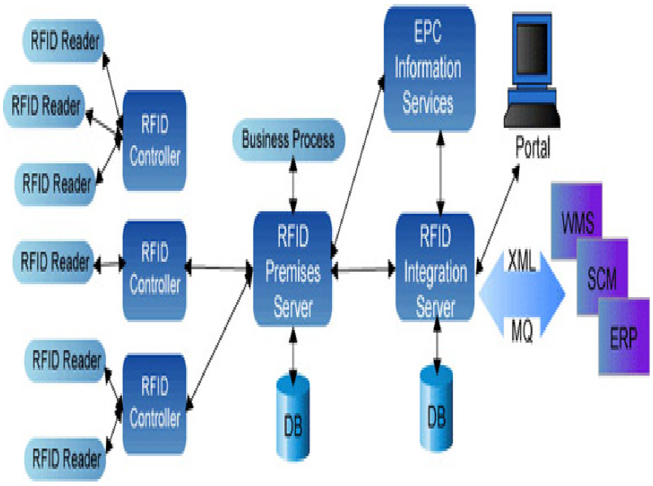


Figure 1. Typical Diagram of an RFID Framework

Many different RFID frameworks exist (including IBM RFID Premises Server and IBM RFID Device Infrastructure) to satisfy different needs, but they all require common functional components:

1) Drivers that can help you understand what the reader devices say and what they want.
2) Adapters that can help you exchange information between your RFID framework and back-end applications.
3) A core component that can manipulate and route the message/data between hardware devices and back-end applications.

The IBM RFID Device Development Kit (hereafter called RFID DDK) is a powerful toolkit for developing RFID reader adapters on IBM RFID middleware solutions. we'll develop several agents based on the project in the Tracking Kit.

2. IBM RFID FRAMEWORK

A RFID solution consists of many kinds of sensor and actuator devices such as RFID tags and readers, visual indicators, switches and sensors.

2.1 WebSphere Premises Server

WebSphere RFID is an on demand solution for RFID applications. WebSphere® Premises Server creates a business context that enables automatic operational decision making and helps ensure reliable delivery of messages between devices and Premises Server, as well as between Premises Server and WebSphere business integration products. Premises Server benefits include:

1) Recognizable impact on asset management, asset tracking, and working capital.
2) Collaborative data sharing enterprise-wide and with partners, allowing trend analysis and traceability.
3) Increased distribution center productivity through automated data capture and verification.
4) Increased data accuracy to improve customer service
5) Improved profit potential with enhanced point-of-decision capabilities and cross-channel selling.
6) Reduced shrinkage, return, and reconciliation costs through improved planning and allocation.
7) Minimized disruption as sensor technology evolves, because of product support.

The Premises Server can filter, correlate, and integrate information obtained from multiple Edge Readers and Controllers for interpretation and correlation of RFID events. It also executes business logic specific to given premises, and it creates and maintains the "system of record" for all RFID events that occur on given premises. (see Figure 2)

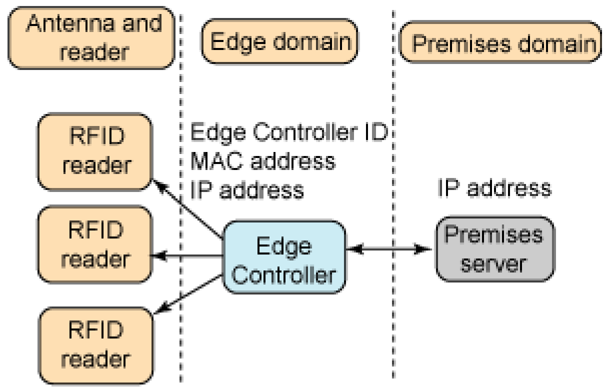


Figure 2. RFID Framework

WebSphere Premises Server:

1) Builds upon IBM SOA process integration through integration among sensor data and enterprise business processes and SOA business applications allowing clients to rapidly deploy secure, reliable solutions.
2) Provides a sophisticated, intelligent business rules engine to develop complex sensor business

processes, speeding time to deployment and increasing process flexibility and control.
3) Enables real-time location tracking services including device interfaces for active tags.
4) Includes starter kits for commonly needed tasks, including Dock Door Receiving and Print-Verify-Ship.
5) Provides an Eclipse-based open platform for integrating sensor devices.
6) Provides an LLRP device agent that enables WebSphere Premises Server to work with any LLRP compliant RFID reader.
7) Supports EPCglobal standards through compliant filtering and correlation, including Application Level Events (ALE) processing.
8) Integrates with IBM InfoSphere Traceability Server to enable solutions to strategically manage and integrate sensor information with enterprise applications as well as securely share sensor information and events with selected trading partners with this scalable EPCglobal standards-based repository.

2.2 RFID Device Development Kit

If you want to write programs to interact with these devices, you need to write code to do the following:

1) Submit requests to the devices.
2) Listen to response data streams from the devices.
3) Enable communication channels such as TCP/IP or RS232 COM port or USB.
4) Aggregate and parse response data streams and then fire the corresponding triggers for the specific response.

RFID Device Development Kit is a good toolkit for modeling the communication channels. We call this type of XML Control Markup Language (CML). you can define the parameters of communication channels with the devices, commands to control the devices, and responses from the devices using CML. So,we can define request and response data streams (device data stream layer), and triggers corresponding to the response data streams into XML.

We just need to focus on studying and understanding the programming specifications for how to control and operate the devices by sending requests to the devices and receiving and parsing the responses from the devices. Don't need to worry about how to write code to enable the communication channels, send the commands to the devices, receive and parse the response data stream from the devices, and drive the triggers corresponding to the response data stream. Once you become familiar with the specifications, you can start to define the CML file. After that, you can use the RFID DDK, an Eclipse plug-in of WebSphere Studio Device Developer to generate the necessary code.Figure 3 shows how this works:

1) The application gets a command and signal from the device and registers interest in the signal.
2) The application tells the command to execute.
3) The command writes a message to the transport.
4) The transport translates the message into a format that the hardware device can understand and writes the translated message to the connection.
5) The connection writes bytes to the class library and OS layer.
6) The class library and OS layer write bytes to the hardware.
7) The class library and OS layer read bytes to the hardware.
8) The connection reads bytes from the class library and OS layer.

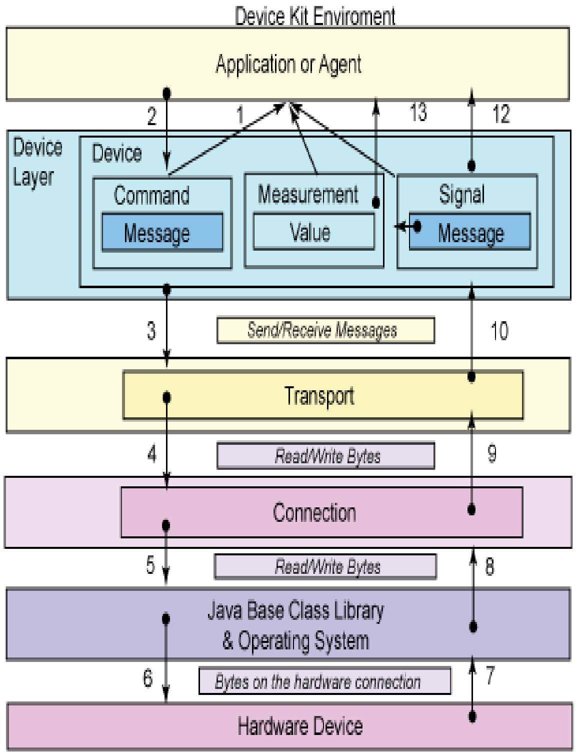


Figure 3. RFID DDK Device Adapter Data Flow

9) The transport monitors the byte stream being sent by the connection and parses the bytes back into messages that the transport understands.
10) The transport notifies the devices that are listening of the messages received.
11) When a signal matches the messages received from the transport, the measurement is notified that the signal was received.
12) When a signal matches the messages received from the transport, the application is notified that the signal was received.

3. DEVELOP APPLICATION WITH THE RFID DDK

3.1 Using MBAF in a Tracking Kit Agent

The MBAF provides an abstract class BaseMicroBrokerAgent from which subclasses derive their own agent classes. MBAF defines an agent simply as an abstraction of an object that uses the MicroBroker. The BaseMicroBrokerAgent class encapsulates many of the functions of using the MicroBroker, including:

1) Creating a MicroBroker client.
2) Connecting the client to the MicroBroker.
3) Maintaining the client's connection to the MicroBroker.
4) Automatically subscribing to topic publications.
5) Automatically decoding the data of a subscribed topic as it arrives.
6) Performing publication requests on a private thread. This is important because it allows an agent to publish a topic directly from the MicroBroker thread that delivered a subscribed topic publication.
7) Automatically encoding data when it's published.

8) Disconnecting the client from the MicroBroker.

Listing 1. Shows a Simple Agent

```
public class Webagent    implements ClientAPIConstants
{
// ClientAPIConstants interface contains the constants used in
the code.
public static void main(String[] argList) {
try {

//Create Proxies
JobProxy proxy = new JobProxy();
DeviceProxy device_proxy = new DeviceProxy();
SoftwareProxy sw_proxy = new SoftwareProxy();

//Get device id from name
Device device =
device_proxy.getDeviceFromName("RFIDClientDevice ",
"Win32");

//create Software object
Software sw = new Software();
sw.setSoftwareName("RFIDswdistributiontest");
sw.setDeviceClassName("Win32");

//create Software on DM Server
long swid = sw_proxy.createSoftware(sw);

//    Job constructor
Job job1 = new Job();

// Set device class name of job.
job1.setDeviceClassName("Win32");

Set the job type.
job1.setJobType(DMS__SOFTWARE_DISTRIBUTION_JOB
_TYPE);

//Set the job priority.
job1.setJobPriority(1);

// Set the job description.
job1.setJobDescription("RFID Software Distribution Job");

//Set job parm - identifier of the software package
Hashtable cfg_parms = new Hashtable();
cfg_parms.put(DMS__ID_OF_SOFTWARE_PACKAGE,
Long.toString(swid));
job1.setJobParms(cfg_parms);

//Set target device
long target_devices[] ={device.getDeviceID()};
job1.setDevicesForJob(target_devices);

//Create the job and return the job_id
long job_id = proxy.createJob(job1);
System.out.println("Job created, id = " + job_id);
return;
}catch (Exception e) {
System.out.println("Exception in api - " + e);
}
}
}
```

3.2 Pass the RFID Data to the Back-end Application with JMS

When the information we want has been retrieved from thousands of tags, we need to pass the tags to the back-end application, which will be the real consumer of this data. But how do you do that? Many RFID frameworks retrieve data from tags and use a Java Message Service (JMS)-compliant messaging queue as the data repository. Listing 2 shows an MDB that monitors the JMS queue and then retrieves the message when it arrives, transforms its format, and routes it to another JMS queue.

Listing 2. A Message-driven Bean in Action

```
public class MessageBean
   implements javax.ejb.MessageDrivenBean,
javax.jms.MessageListener
  {

   private javax.ejb.MessageDrivenContext
fMessageDrivenCtx;

   public String generateXMLStr(String result, String date,
String location)
  {
                StringBuffer strBuf = new StringBuffer("<?xml
version=\"1.0\"

encoding=\"UTF-8\"?><dock-door-receiving-response><respo
nse result=");
                strBuf.append("\"" + result + "\"");
                strBuf.append(" location=");
                strBuf.append("\""+ location + "\"");
                strBuf.append("
original_timestamp=\"1085558889990\" date=");
                strBuf.append("\"" + date + "\"");
                strBuf.append("
status=\"2\"/></dock-door-receiving-response>");
                return strBuf.toString();

   }

  public void putXMLToMQ(String xmlStr)
  {
                  // send message to KIMONO.RESPONSE.Q
                  //System.out.println("Start to send message");

                  try {
                         QueueConnectionFactory
connectionFactory =

(QueueConnectionFactory)getRemoteObject("jms/ibm.rfid.qm"
);
                         Queue queue =
(Queue)getRemoteObject("jms/kimono.response.q");
                         QueueConnection queueConnection =
connectionFactory.createQueueConnection();
                         QueueSession queueSession =
queueConnection.createQueueSession(false, 1);
                         QueueSender queueSender =
queueSession.createSender(queue);
                         TextMessage sendMessage =
queueSession.createTextMessage(xmlStr);
                         queueSender.send(sendMessage);
                         queueSender.close();
                         queueSession.close();
                         queueConnection.close();
                  }    catch (ParserConfigurationException e) {
                         e.printStackTrace();
                  } catch (FactoryConfigurationError e) {
                         e.printStackTrace();
                  } catch (JMSException e) {
```

```
                        System.out.println(e.getMessage());
                } catch (Exception e) {
                        System.out.println("===
MDB1Bean.onMessage.Exception ===");
                        System.out.println(e.getMessage());
                        System.out.println("===
MDB1Bean.onMessage.Exception ===");
                }
                    System.out.println("Finish send to queue");
            }

   protected InitialContext getInitialContext()
                                throws NamingException
{
                if(context == null)
            {
                    context = new InitialContext();
                }
                        return context;
}
}
```

4. CONCLUSIONS

In this article, we can use the RFID Tracking Kit to easily develop a check-in check-out RFID application, including detailed steps showing how to create the customized agents, as well as steps for configuring and debugging the runtimes for Edge controller and the Premises simulator for the project. The WebSphere RFID DDK proves the powerful tools you need to develop and test the glue code easily and efficiently.

REFERENCES

[1] A. Hetzroni, D. Shapira, I. Esquira and Y. Edan, Data capture in an agricultural setting for traceability and management, EFITA/WCCA Joint Congress on IT in Agriculture VilaReal, Portugal (2005), pp. 922–927.

[2] Y.G. Ampatzidis, S.G. Vougioukas, D.D. Bochtis and C.A. Tsatsarelis, A yield mapping system for hand-harvested fruits based on RFID and GPS location technologies: field testing, Precis. Agric. 10 (1) (2009), pp. 63–72.

[3] J.D. Walter and L.F. Backer, Sugarbeet yield monitoring for site-specific farming. Part I. Laboratory tests and preliminary field tests, Precis. Agric. 4 (4) (2003), pp. 421–431.

[4] R. Bishop, Intelligent Vechicle Technology and Trends. Artech House Publishers, 2005.

[5] K. Finkenzeller, RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification. New York, NY, USA: John Wiley & Sons, Inc., 2003.

[6] EPC Radio-Frequency Identity Protocols Class-1 generation-2 UHF RFID Protocol for Communicaitons at 860 MHz – 960 MHz, Version1.1.0, 2005.

[7] J. Myung and W. Lee, “Adaptive binary splitting: A RFID tag collision arbitration protocol for tag identification,” in Proc. 2005 2nd Int. Conf.Broadband Networks (Broadnets), Boston, MA, Oct. 3–7, 2005, vol. 1,pp. 347–355.

[8] P. Hernández, J. D. Sandoval, F. Puente, and F. Pérez, “Mathematical model for a multiread anticollision protocol,” in Proc. 2001 IEEE Pacific Rim Conf. Communications, Comput. Signal Process., Victoria, BC, Canada, Aug. 26–28, 2001, vol. 2, pp. 647–650.

[9] B. Zhen, M. Kobayashi, and M. Shimizu, “To read transmitter-only RFID tags with confidence,” in Proc. 2004 IEEE 15th Int. Symp. Personal, Indoor Mobile Radio Commun., Barcelona, Spain, Sep. 5–8, 2004, vol. 1, pp. 396–400.

[10] G. Mazurek, “RFID system with DS-CDMA transmission,” in Proc.KKRRiT-2006 Conf., Poznan, Poland, Jun. 7–9, 2006, pp. 313–316.

[11] G. Mazurek, “Jamming protection of spread spectrum RFID system,”Proc. SPIE, vol. 6347, pp. 1–6, Oct. 2006.

[12] U. Meyer-Baese, Digital Signal Processing with Field Programmable Gate Arrays. New York: Springer-Verlag, 2001, ch. 3.4.

[13] T. Nakanishi and T. Ikegami, “Throughput performance of CDMA-ALOHA in s-band land mobile satellite and stratospheric platform channels,” in Proc. 11th Int. Symp. Personal, Indoor Mobile Radio Commun., London, U.K., Sep. 2000, vol. 2, pp. 1085–1089.

[14] J.-W. So and D.-H. Cho, “Performance analysis of a DS/SSMA unslotted ALOHA system with two user classes,” IEEE Trans. Veh. Technol., vol. 51, pp. 1628–1639, Nov. 2002.

[15] K. A. Hamdi, “A new accurate packet error rate analysis of unslotted CDMA wireless networks,” in Proc. 2006 IEEE Int. Conf. Commun.,Jun. 2006, vol. 11, pp. 5160–5165.

EQUIPMENT MANAGEMENT SYSTEM BASED ON DISTRIBUTED DATABASE

Lijun Bai, Gongxin Wu, Yujia Ge
College of Computer Science and Information Engineering, Zhejiang Gongshang University
Hangzhou, Zhejiang 310018, China
Email: blj@mail.zjgsu.edu.cn; ywwgx@netease.com; yge@mail.zjgsu.edu.cn

ABSTRACT

To overcome the deficiency of current centralized Equipment Management Systems, this paper presents a system based on distributed database, and discusses the architecture, design of the system, implementation of distributed transaction of the Equipment Management System. It can not only solve the contradiction between data distribution and centralized management, but also obtain query result fast and conveniently.

Keywords: Distributed Database, Equipment Management, Distributed Query.

1. INTRODUCTION

With computer networks and digital communication technology's rapid development, although the centralized database has been made great achievements from theory to practice, the research and development of Distributed database systems, which are mainly in the distribution of data storage and distributed processing, is receiving increasing attention [1]. For example, in government departments, financial sectors, logistics management and other sector-specific applications, taking into account the logic and geographic factors, the distribution of computer information systems is an inevitable trend [2].

Technology of distributed database takes many advantages, but it also takes some difficulties on data consistency, data remote-transfer, transaction concurrencies, which make the research of Distributed Database System more complex. Presently Equipment Management System of national inspection and quarantine bureau is operated separately by each provincial branch and the national bureau can’t acquire real-time data in time. With the increase of equipment, the distributed data management schema must be adopted. This paper describes the a management model based on Distributed Database System for Equipment Management System of national inspection and quarantine bureau.

2. ANALYSIS OF EQUIPMENT MANAGEMENT DATA

In instrument management systems, data can be divided into 2 classes: one is equipment data, including basic equipment information - utility rate, delivery of equipment, equipment disuse and lab information, etc; the other is basic data which is identified information of equipment attributes, such as equipment category, equipment status, means of equipment delivery, means of appending, etc, and these data usually are fixed. Equipment data has a few features:

(1) Physical distribution – collections of data (e.g. in a database) can be distributed across multiple physical locations.
(2) Logical integrity – although data is distributed in different computers, they are not irrelevant and related to each other in logic.
(3) Location transparency – users must be able to interact with the system as if it were one logical system.
(4) Site coordination and autonomy – every site in the network is independent and can perform local application. At the same time, each site is one part of the whole system and can deal with global application.
(5) Redundancy of data – in centralized database, one of goals is to reduce redundant data. In distributed system, to increase redundant data is one means to enhance reliability and usability.

According to the distributed features of equipment management system, we adopt client/server architecture and build a system based on distributed database, which can satisfy the need of provincial branch and realize real-time monitor of national bureau. Thus it can improve equipment management efficiency for inspection and quarantine section.

3. ARCHITECTURE OF THE SYSTEM

The system is a multi-layer network topology and is composed of local area networks of provincial branches. Each provincial branch network is composed of database server and several clients. Provincial branch databases are inter-connected by private commerce and inspection network. The architecture of the system is shown in Fig. 1.

A provincial branch database server includes data from its different districts and provincial branch itself. It takes all equipment management function. National bureau database server includes equipment data from all provincial branches, but it does not have functions such as equipment update, delivery of equipment. So it does not need to collect provincial branch data and only monitor and backup provincial branch data. Cross-regional delivery of equipment is operated through provincial branch database server.

4. DESIGN OF SYSTEM

This is involved with design of distributed database and catalogue structure based on it for inspection and quarantine section [3].

4.1 Design of Database

There are 2 methods for the design of a database: top-down and bottom-up. Top-down designs distributed database from beginning, while bottom-up designs distributed database by clustering existed data collection. With present equipment management database, we employ bottom-up. First, we build a global data schema according to the existed data collection. Then this global relation is divided into several segmentations not overlapped. and finally the physical allocation of segmentations are decided. Segmentation and allocation of data is mainly discussed.

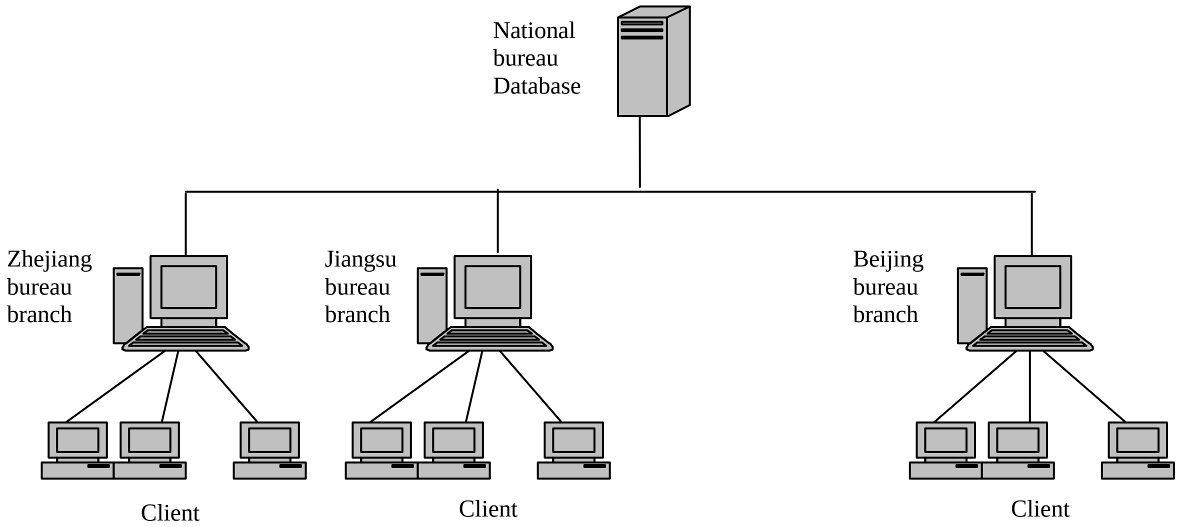


Figure. 1 Architecture of System

Each provincial branch will employ the same equipment management system, and the relational schema of data table they used is the same. For many operations, the provincial branch is responsible for the process of its equipment management. Operations, such as appending, disuse of equipment, can be handled at provincial side. Delivery of equipment among provincial branches is operated through national bureau database. Thus, data distribution of the system employs horizontal segmentation divided by identifier of provincial branch.

Because of the local distribution of provincial branch, all data relation schema concerning with concrete operations of equipment management is divided into several segments that are not overlapped. These segmentations are allocated into corresponding provincial branch database server. For these data segmentations, each has just one mapping copy at the corresponding website, and no replication copy. But some relational schemas, such as equipment category, equipment status, delivery of equipment, appending of equipment, must reside on every provincial branch. For these relational schemas, our model treats them as horizontal ones; each relation is divided into one segment, and then each segment has one copy on each provincial branch server.

The segmentation and allocation of two relational schemas is shown in Fig. 2. SBKPK is table name of basic equipment information, while SBFL is table name of equipment category. SBKPK (Zero_Id='33') is a segmentation divided by the identifier of Zhejiang Province. SBKPK $(\text{Zero\_Id}='33')^{zj}$ is a physical mapping of segmentation allocated into Zhejiang Province. SBFLALL is segmentation with the whole relation; $SBFL_{ALL}^{zj}$ is a physical mapping of segmentation SBFLALL allocated into Zhejiang Province.

With segmentation and allocation schema mentioned above, national bureau database server stores these tables, equipment information, equipment utility rate, budget, equipment disuse, delivery of equipment, operator information. Each provincial branch server stores tables, including equipment information, equipment utility rate, budget, equipment disuse, delivery of equipment, operator information, and all these tables are related to the provincial branch itself.

4.2 Design of Catalogue Structure

Catalogue of distributed database is called data dictionary, which is the container stores objects and control information of distributed database. It not only contains databases, domains, relations, attributes, users, but also contains description of segmentation, location of segmentation, mapping of native name, access permissions, etc [4]. We only concern with the required catalogue information that supports the distributed database, including basic information of the site, global logic structure, segment structure, mapping structure, data source information structure, etc.

Basic site information includes required information of provincial branches of the distributed database. It includes identifier of site, site name, IP address, port number, connection password, description of site. When system starts up, it automatically searches other related sites according to IP address and port number of basic information and determines their status. When the system detects connection request from its listening port, it will verify the validity according to site identifier and connection password. When system starts up, a site can set up its listening port according to its port number.

Global logic schema provides a conceptual description of global relation, and the committed SQL statement handling global database will be checked and transformed with it. Segmentation structure will provide related information, specific segmentations that a global SQL statement will access can be determined with it. Mapping structure describes the allocation of segmentation, when specific segmentations a global SQL statement will access are determined, the copy at which site will be accessed can be determined. Native mapping structure describes the relation of physical mapping of a site with native database. Data source information table provides login information required when native database is accessed.

In a distributed database, catalogue distribution can be 3 types:
(1) Centralized catalog – global catalog will be stored at one specific site called central site.
(2) Full replicated catalog – catalog will be replicated at each site.
(3) Local catalog – each site will store one catalog, i.e., local catalog, for native site.

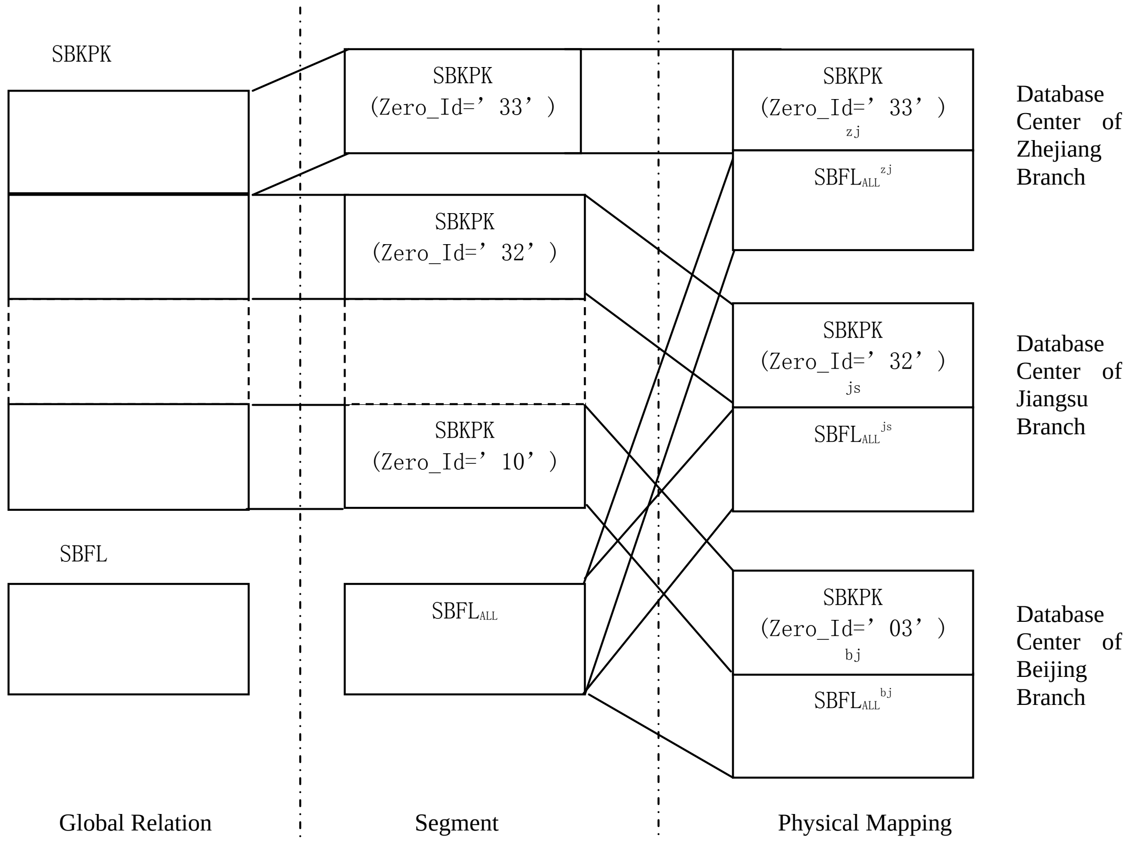


Figure. 2 Segments and Physical Mapping of Global Relation

Once basic information structure of the site, global logic schema, segmentation structure, and mapping structure is determined, they are little changed. In order that each provincial branch can accomplish distributed scheme independently without relying on other site, we adopt full replication scheme. This can reduce the cost of accessing catalog among sites. Native mapping structure and data information structure only concerned with native DBMS employs native catalog distribution.

5. IMPLEMENTATION OF DISTRIBUTION OF THE SYSTEM

5.1 Design of Catalogue Structure

There are 4 methods of distributed database synchronization, including handwork, synchronization based on e-mail, synchronization based on ftp service and independence data synchronization module [6]. Data replication synchronization of SQL Server is based on "publish/subscribe" mode. In replication, servers for SQL Server include publishing server, distributing server, subscribing server [5]. A computer can be one or two server role. Publishing server stores and provides source data, that is, it is the server that provides sample data for replication. Distributing server is a transfer between publishing server and distributing server, and all distributive data are sent to subscribing server by way of distributing server. Subscribing server is destination server in replication, and it receives data from distributing server. With different situation, different replication synchronization schema may be adopted. Distinctions between different schemas are location and role of the server.

In equipment management system, two replication schemas are employed:

(1) Centralized publishing server with remote distributing server

In this schema, an extra server takes on the distributive role, and its mode is publishing, distributing, subscribing. A server fulfilling distributive task independently may alleviate the burden of publishing server. Replication of basic data, replication of statistical parameters employs this schema. When they are changed, national bureau modifies and distributing them.

(2) Centralized Subscribing Server

A physical server is deployed as centralized subscribing server. It may receive data from several publishing server. In equipment management system, basic equipment information, equipment disuse, delivery of equipment may employ this replication schema, with national bureau database server as centralized publishing server and provincial branch server as distributing server.

5.2 Distributed Query

In a large scale distributed system, it is usually very difficult to find an optimal plan for a query [7]. When a distributed query is executed, global query committed by a user is interpreted as several local queries that related nodes can recognize. Then every local query is executed, and summarization of query results from each node is returned to users.

When tables or views on remote database are accessed, data is operated by Select, INSERT, UPDATE, DELETE statement [8]. The statement has the following form,

Select column expression <, column expression> from tablename@Date_link <, tablename@Date_link> [where logic expression];

If the accessed data is from the same data table, the table or its view (snapshot) can be accessed directly. If accessed data is from different tables, they may be accessed by “join” operator. Data in a view come from 2 sources, one case is from one or more tables in the same database, and the other case is from 2 or more tables from databases at different site. If global data are queried, views from different sites may be built up, and then views are queried.

Basic strategies of distributed queries include simple query decomposition strategy, improved query strategy, and schedule query strategy. But the common problem among these strategies is the conversion of global query to local queries. There are six steps to convert a global query to local queries:
(1) Conversion of global SQL to global query tree: when global processor receives a global SQL statement, it begins lexical analysis and grammar analysis. When this is done, global table that stores required data is obtained. At the same time, SQL statement is conversed into a grammar tree. Then grammar tree is analyzed. Finally the query tree is generated.
(2) Conversion of global query tree to segment tree.
(3) Algebraic optimization of segment query tree.
(4) Disposition and location optimization of segment query tree.
(5) Semi-join optimization of segment query tree.
(6) Conversion of segment query tree to local SQL statement.

6. CONCLUSIONS

With distributed system being widely used, distributed database becomes an important field of information processing. We propose our equipment management system based on distributed database. The system architecture, design, and implementation of distributive functions in the paper can resolve the conflicts from distributed data and centralized data management of the system. Supervisors in national bureau may monitor the equipment utility of provincial branches in real-time and obtain query results from global database to meet the need of management. This catches up with the trend of the latest technique.

REFERENCES

[1] Sha Zhouyang, Zhang Zhonglei, “Synchronization in Distributed Database”, Science Information, No.3, 2007, pp.43~44.
[2] Lin Qingguo, Zhang Qiang, Yang Xiujian, “Chain-store Distribution Management System Based on Distributed Database”, Logistics Sci-tech, Vol.29, No.128, 2005, pp.28~30.
[3] Sa Shixuan, Wang Shan, Conspectus on Database System, Bei Jing: Higher Education Press. Pub., 2002.
[4] Shao Peiying, Distributed Database System and Its Applications, Bei Jing: Science and Technology Press. Pub., 2005.
[5] Gai Jiuyu, Zhang Zhongneng, Xiao He, “Data Replication Technique Analysis and Application in Distributed Database System”, Computer Applications and Software, Vol.22, No.7, July 2005, pp.36~41.
[6] Jing Shenyan, “Synchronization of distributed database technology and its applications”, Electronic Engineer, Vol.34, No.12, December 2008, pp.48~50.
[7] Yongluan Zhou, Beng Chi Ooi, Kian-lee Tan, Wee Hyong Tok, “An Adaptable Distributed Query Processing Architecture”, Data & Knowledge Engineering, Vol.53, 2005, pp. 283~309.
[8] Jing Zheng, Tian Yongchao, Advanced Development and Professional Applications of SQL Server, Bei Jing: National Defense Industry Press. Pub., 2002.

Lijun Bai is an assistant professor in College of Computer Science and Information Engineering of Zhejiang Gongshang Univesity. She received her master’s degree from Computer Science department of Shan Xi University in 2003. She has published more than 10 journal papers and two books. Her research area is data mining and distributed database.

Gongxin Wu is an assistant professor in College of Computer Science and Information Engineering of Zhejiang Gongshang Univesity. He received his master’s degree from Computer Science department of Hehai University in 2003. He has published more than 10 journal papers and attended several national and provisional research projects as a major member. His research area is distributed database systems, design and implementation of information systems.

APPLICATION RESEARCH OF WORKFLOW TECHNOLOGY IN MANAGEMENT SYSTEM

Ying Yang, Hong Liu, Yongqiang Fan, Lei Xu, Jianliang Zhou
Computer Science and Information Engineering College, Zhejiang GongShang University
Hangzhou, Zhejiang, China
Email: yangying0602@126.com

ABSTRACT

Workflow technology is a key technology to improve business process efficiency and productivity. It can monitor, control and coordinate the business processes and programs and can track the flow of work and information as well as the utility and inputs of resource in advance. The paper introduces the workflow technology into an online meal-ordering system, System application result shows that，such integration can not only improve system development efficiency, but also help to realize automation of ordering meal process.

Keywords: Workflow Technology, Online Meal-ordering System, JBPM

1. INTRODUCTION

Workflow management systems had a significant impact on organizations in the past decade [1] and are a promising technology that automates business processes to improve the efficiency and manageability of an enterprise's daily teamwork [2]. How to rapidly and cost-effectively satisfy the various dynamic demands of the market adapt to an unpredictable business environment and increasing rates of new production methods are the major challenges for enterprises. Achieving the combination of internal and external flows on heterogeneous platforms in a distributed environment is an important research area. As the core of workflow technology is now considered as the key technology to improve business process efficiency and productivity. It monitors, controls and coordinates the business processes and programs and tracks the flow of work and information as well as the utility and inputs of resource in advance.

Workflow is the automation of a business process. It has now been adopted as a way to implement the cross organization management needed to carry out businesses. With the development and popularization of network technology, convenient, fast and personalized online meal-ordering is also gradually coming into people's daily life. And e-commerce is the most appropriate business mode to reflect the diversity and personalization of such meal-ordering service. In this paper, workflow technology is introduced into online meal-ordering system. Paper firstly introduces the workflow technology, and then it puts forward a model of online meal-ordering system based on workflow technology, lastly we do a summarization and provide suggestions for future work.

2. ONLINE MEAL-ORDERING SYSTEM

The system is primarily divided into two separate subsystems: foreground meal-ordering subsystem and background management subsystem. Of which, foreground meal ordering subsystem aims to provide perfect and convenient services of meal ordering manipulation for consumers, that includes hotel introduction, ranking list of meal ordering, specialty recommendation, fast track of meal ordering, meal ordering trolley, article information, online consultation and so on. Its' class diagram is shown as figure 2.And background management subsystem aims to maintain system data and system process flow, including menu information management, meal ordering data management, user information management, meal sequence management, distributor management, article management etcetera, and so on. Its' class diagram is shown as figure 1.

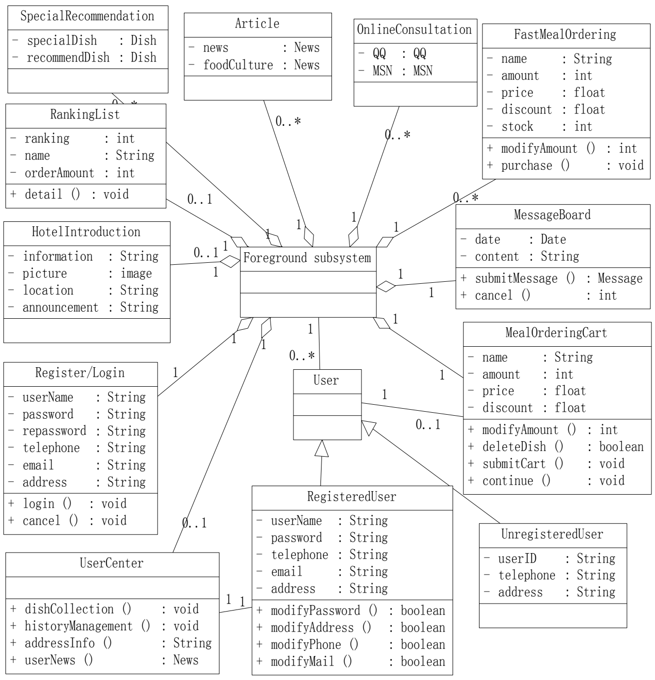


Figure1. Class Diagram of Foreground Meal-ordering Subsystem

(1) Hotel introduction. Including fundamental information of hotel, hotel picture, geographical location of hotel, hotel announcement etc.
(2) Ranking list of meal-ordering. It displays the meals which ordered mostly, and it can be classified according to weekly, monthly, seasonally ranking information.
(3) Special recommendation. Classify into shopkeeper recommendation, e-pal recommendation, latest meal recommendation, and such classification exhibits special meal and latest meal.
(4) Fast track of meal ordering. User can browse entire meals or specific one, also could search by name or price, either can skim the detail information, which contains meal images, ingredients, remark from e-pals etcetera. And selected meal could be distributed into meal trolley. Then if user demand to book meal, the meal ordering submodule will be invoked automatically, and this module shows latest meal sequence information list (including user amount, fee criterion and so on), statement (booked or not), user is allowed to book empty meal seat in accordance with filtering book time and

book the seat by filling the meal booking menu concurrently.

(5) Meal-ordering cart. If user booked the meal via fast track, it'll be added the meal-ordering cart. The meal-ordering cart displays the booked meals of the user. User is permitted to do such operations: modification of the amount of the ordered meals, deletion of ordered meals, also is able to continue the ordering, or make cancellation of this ordering.

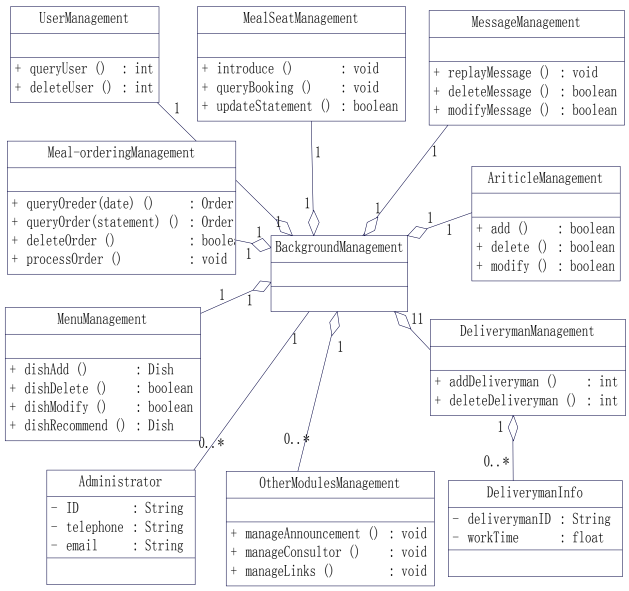


Figure2. Class Diagram of Background Management Subsystem

(1) Menu module: Meal addition, deletion, modification and update, statement (provided or not currently), special recommendation.
(2) Meal-ordering module: By date and statement, Search and manage orders, delete expired orders.
(3) User management module: User information search, delete users, user's history orders, user's rank, news reported by users.
(4) Message module: Reply user's message, delete message.
(5) Meal-seat management module: meal-seat information, search booking situation by information of everyday and every meal, update the available state of meal-seat.
(6) Deliveryman management: each order is combined with each deliveryman, work hours will be accumulated.
(7) Article management: manage the articles in cooking culture, such as addition, deletion and modification etc.
(8) Modules management: including hotel announcement management, consulter management, and friendly link etc.

3. ONLINE MEAL-ORDERING SYSTEM ARCHITECTURE BASED ON WORKFLOW

Conventionally, business processes were implemented by hard-coding business process related aspects, such as control and data flow, into the organization's software systems. This leads, however, to inflexible systems that were hard to modify and maintain. Workflow is a technology that addresses such problems by separating and abstracting business processes from the software systems [8]. It is an automation of business process, during which documents, information, or tasks are passed from one participant to another for action, according to a set of procedural rules.

In the online meal-ordering system, system functions as well as business processes and business activities involved are relatively stable. Based on the data controlling model described in workflow, all the business processes of ordering meal system are under collective control. System architecture integrated with workflow technology is shown as Figure 3.

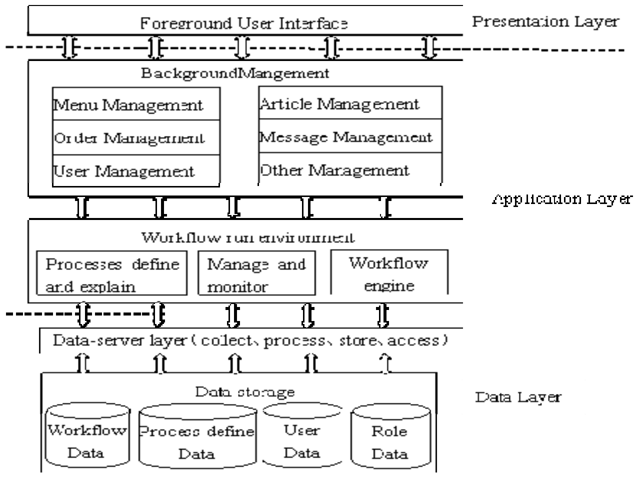


Figure 3. System Architecture

According to process logic, the system can be divided into following three layers: presentation layer, application layer and data layer.

1) Presentation layer. It provides unified user interface and processes interaction between users and system in order to facilitate the completion of some human-computer interaction tasks.

2) Application layer. It means all the functions of business management are implemented under workflow management mechanism. As the core of workflow management system, workflow engine is a runtime software service for workflow instance. This engine is responsible for online meal-ordering scheduling and resource allocation. It provides concurrent services for a number of external requests, responsible for the definition of meal-ordering business process, creates business instance and manages its data-item, calls external procedures and maintains workflow data.

3) Data layer. It provides unified data acquisition, processing, storage and access services. Data storage includes Workflow process definition data, organization role data, workflow data and user data.

4. WORKFLOW DESIGN OF ONLINE MEAL-ORDERING SYSTEM

4.1 Order Processing Flow Design

In the system, unregistered users can order meal and reserve table, but personal information is needed when confirmation is submitted. Registered users are entitled to do other operations besides what unregistered users can do, such as inquiring their historical reservation information, member discount and so on. Such the flow is shown as figure 4.

In generic order processing flow, there are three roles, and they have their own different tasks. User's tasks are to choice goods and submit order; checker will affirm the legal order and hand it to deliveryman. For orders of exception, this is likely to be caused by insufficient stock, system will send the exception information to user and when the exception is handled it will be handed to deliveryman. The flow is shown as Figure 5.

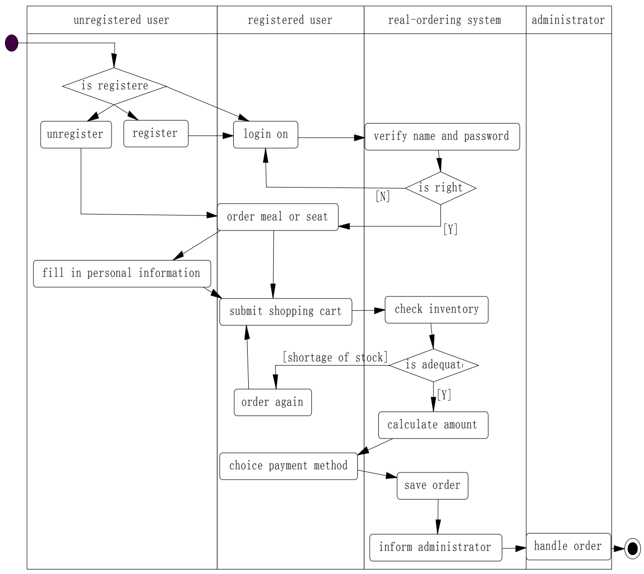


Figure 4. Activity Diagram of Ordering Meal

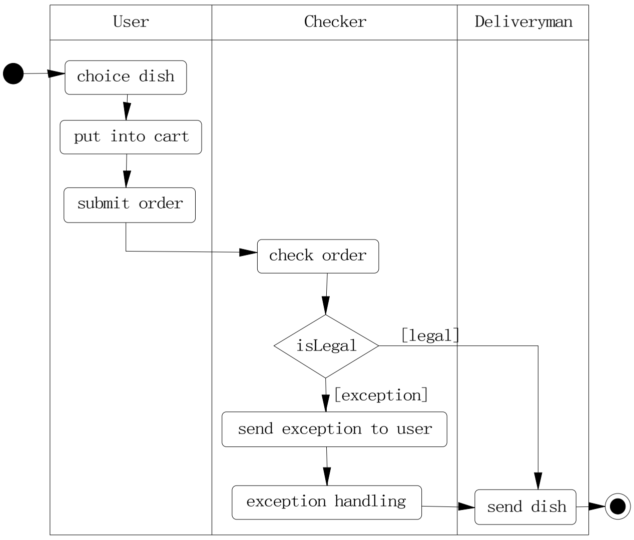


Figure5. Generic Order Processing Flow

4.2 Special Flow Design

After user submits the order, he may want to modify the order, such as add, delete, or modify the goods in this order, so system setups a temporary buffer. This buffer will expire after checker audits the order. The flow of user changes order is shown as Figure 6.

Figure6. Flow of Users' Modifying Order

Of course, user can cancel order in this period. The flow of user cancel order is shown as Figure 7.

Figure7. Flow of Users' Deleting Order

4.3 Flow's changeability Design

This system has three roles, and the whole flow may be changed in the future. For example, goods may have no use for sending out, so deliveryman is an unwanted role and should be deleted from the flow. For the sake of realizing this foundation, the flow should be changed. Intercommunion of each role is no longer a top-down. All the roles will join to the same node and intercommunion each other through this node.

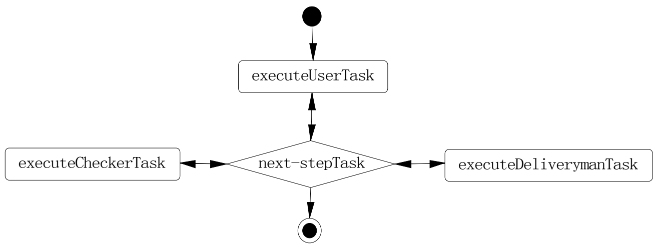


Figure 8. Activity Diagram of Flow Changing

After every role executes its own task, it needs to inform the middle node, and the next role or end the flow will be chosen by the middle node.

The disadvantage of the process design is that the role can not be added or modified in the flow. But there is another method to realize hot-deployment for new or modified role. We could define roles and confirm the relations between each role, and then distribute existent tasks to every role. This scheme will take the most workload in the workflow system, because many operations in the workflow are based on this, such as flow modification, flow query and so on. Its workload mostly focuses on the edit-workload that web page uses Workflow Definition Language to edit the workflow definition document and dealing-workload of relationship between definition document and tasks.

5. SYSTEM IMPLEMENTATION WITH JBPM

Business Process Management involves the graphical modeling of a business process, from which workflow software can be generated, which in turn will control the live operation of the process, interacting with both humans and other applications.

Business Process Management is the natural evolution and convergence of several powerful forces within the fields of software development methodology, enterprise application technology, and management theory. These underlying forces have all matured and converged at the right time for a productive fusion, which we know as business process management.

5.1 Traditional Software Development Methodology

Developing enterprise application software is about delivering value to a business, and the business expresses that value as a set of business requirements. Business requirements are subject to change. Businesses cannot afford to stay still: if they don't adapt to the marketplace then they will not survive. But traditional software development methodologies are hard to satisfy the changeable requirements.

Unfortunately, this is not the only problem with the traditional software development methodologies. There is also the problem of business requirements "dissonance". This is where the layers of end users, analysts, and developers create a chain and each link in the chain puts its own interpretation on the requirement, until the end result is horribly different from what the business originally needed.

5.2 Agile Development Methodology

Agile development is more adaptable and can attempt to break down the requirements dissonance by taking out the middle man as much as possible, and by creating prototypes early on, and then iterating them towards the final version. This allows for an iterative approach to software development.

Nevertheless, agile development does have serious drawbacks and limitations. The first and most obvious limitation is that the agile development methodology does away with the Business Analyst. This is an important drawback, because often the BA's interpretation of the requirements is more logical and more far-sighted than that of the end user who specifies the original requirement. This can mean that the developer can be led up blind alleys by an end user who doesn't have the necessary perspective.

Otherwise, some agile developments have turned into one extremely long prototyping process, with an end result never being reached. This is an expensive way to develop software.

5.3 BPM Approach To Agile Development

For some idealists, the best situation would be one where the business users can build the software tools they need for themselves, without having to rely on developers or analysts. Unfortunately, although programming languages are becoming simpler all the time, but for an end user to build their own software is still hard.

Nevertheless, BPM does go some way towards this ideal, and given the right scenario, it can successfully deliver valuable software in extremely short time scales. In a similar fashion to Agile, BPM relies on cutting out the middle man as much as possible, only this time the emphasis is on a strong partnership between the end user and the BA working on iterations towards the final software.

The reinstatement of the Business Analyst has several advantages:

Firstly, the BA is skilled in the interpretation of requirements, and so their business process models are likely to be close to the original requirement.

Secondly, a BA's models are far easier for an end user to understand than code or even prototype software, allowing for closer collaboration and faster development.

Thirdly, the BA has the long term view and business skills to steer the end user's expression of their requirements in the most beneficial direction.

Last, and by no means least, models can usually be produced much more quickly than working software. As the working software produced by a BPM system is initially generated from the BA's process model, this is an extraordinarily fast method of software development.

In this online real-ordering system development, we use JBPM, an open-source workflow engine, to build the meal-ordering system model and the workflow process model and succeed develop this system with the BPM approach.

6. CONCLUSIONS

We have shown the potential of applying workflow technology with online meal-ordering system. The paper discusses the construct of the system. By adopting Unified Modeling Language, we utilized class diagram, use case diagram, activity diagram and sequence diagram to cope with analysis and design for this system from each mentioned perspective, then adopt the BPM approach based on JBPM workflow engine to construct the system model of Workflow Meal-Ordering System.

This system used workflow technique to implement process automation of the online meal-ordering. This technique is not merely beneficial for frugality of manpower, also it contributes to better transaction management, improvement of our work efficiency, frugality of costs, and furthermore it has better utility value and will realize the optimized economic benefit.

REFERENCES

[1] Jianxun Liu, Shensheng Zhang, Jinming Hu. A case study of an inter-enterprise workflow-supported supply chain management system. Information & Management, Volume 42, Issue 3, March 2005, Pages 441-454.

[2] Charles V. Trappey, Amy J.C. Trappey, Ching-Jen Huang, C.C. Ku. The design of a JADE-based autonomous workflow management system for collaborative SoC design. Expert Systems with Applications, Volume 36, Issue 2, Part 2, March 2009, Pages 2659-2669.

[3] YuYue Du, ChangJun Jiang, MengChu Zhou, You Fu. Modeling and monitoring of E-commerce workflows. Information Sciences, Volume 179, Issue 7, 15 March 2009, Pages 995-1006.

[4] Cardoso, J., Sheth, A., Miller, J., Arnold, J., & Kochut, K. (2004). Quality of service for workflows and web service processes. Journal of Web Semantic: Science, Service and Agents on the World Wide Web, 1, 281–308.

[5] Wang, M., Wang, H., & Xu, D. (2005). The design of intelligent workflow monitoring with agent technology. Knowledge-Based Systems, 18, 257–266.

[6] Xu Dahua, Dai Fang, Design of Workflow System Based on Components, Computer and Communications,2005, (23)

[7] Wang Hong, Design and Implementation of Online Meal-ordering system [J], Computer and Information Technology, 2008, (06)

[8] Yu Shi-hua, Du Fu-tao, Developing Online Meal-ordering system with .NET, Journal of Hulunbeier College, 2007,(03)

[9] Matt Cumberlidge , Business Process Management with JBoss jBPM, Packt Publishing Ltd, 2007.

COMBINED FORECASTING MODEL FOR TAX REVENUE BASED ON METHODOLOGY OF SELF-ORGANIZATION

Cungen Shen [1, 2] Wenzhen Zhang [2]
[1] School of Business, Hohai University,Nanjing 210098, China
Email: yzscg@sohu.com
[2] Party School of State Administration of Taxation, China

ABSTRACT

The methodology of self-organization is adopted to select data as basis for establishing a linear regression model of forecasting and a forecasting model based on BP neural network, the self-organization model will then be adopted to establish a combined forecasting model with forecasting model of time series being taken into consideration. Comparison between different models proves high accuracy of the combined forecasting model.

Keywords: Combined forecasting, Methodology of Self-organization, Neural network, Linear regression, Time series.

1. INTRODUCTION

Changes of tax revenue are under influences imposed by several factors beside tax policy which is the major factor: uncertain factors triggering fluctuations, economic growth, resident consumption, total volume of financial expenditure, retail sales of social consumption, etc. Statistical techniques consisting of time series model and regression model could be used in tax forecasting. However, time series model can not take full use of economic factors concerning tax revenue, leading inaccuracy and uncertainty, while regression model requires functional relations between economic factors and tax revenue before substantial mathematical calculation. Experiments have proved that BP neural network works in forecasting economic indexes like tax revenue thanks to its simulated multiple variables and its advantages in using historical data to extract or approximate hidden input/output nonlinear relations rather than making relevant assumptions. But BP neural network is prone to be subjective when selecting input variables during the course of modeling. To make full use of advantages of time series model, regression model and neural network model, the self-organization methodology is adopted for establishing a combined model. In other words, these methodologies will be combined in proper way to maximize accuracy based on useful information that we have access to. Currently, two types of combined forecasting are available:
One is weight coefficient combined forecasting which consists of optimized combined forecasting and variable weight combined forecasting. This type is featured by a precondition of linear relationship between all the forecasting models involved. But linear combination of all the methodologies is not optimal when a single methodology is sourced from non-linear model or the precondition is nonlinear functions of the information set.

The other type is non-linear combined forecasting, and artificial neural network is most commonly used. Experiments show improvements of both fitting precision and reliability of forecasting [1].
Document [1] introduces the combined forecasting based on self-organization methodology. Experiments witness obvious improvements of accuracy of forecasting. Theory of self-organization has been successfully applied into simulation of complicated systems, forecasting, pattern recognition and sample clustering, etc. Document [2] proposes an optimized forecasting model based on a combination of self-organization and BP neural network and proves that self-organization methodology based on selection of relevant variables is more effective in forecasting. This paper proposes that selection of self-organization variables and combined forecasting method mentioned above could be adopted in forecasting, namely, combined forecasting based on self-organization could be achieved by combining results produced from multiple linear regression model, time series model and neural network model before the self-organization methodology is adopted.

2. COMBINED FORECASTING BASED ON SELF-ORGANIZATION METHODOLOGY

2.1 Theory of self-organization

Core technique of this theory is GMDH [3] (Group method of Data Handing) which is originated from development of neural network and computer science. It is a logical combination of Black Box Theory, Biological Neuron Methodology, Induction Method, Probability Theory and Mathematical Logic. Its basic idea is: combination of simple initial input (local variables) generates the first generation of intermediate candidate model from which the best will be selected for the second generation, after several rounds of such cycles involving generation, selection and evolution, models are increasingly complicated until the best one is selected.

GMDH is featured by classification of samples into learning sets and testing sets. Gradually complicated models will be generated in learning sets and the optimized one will be selected in testing set based on external criterion (for selection). Termination principle of such a selection course is the theory of self-organization: while the models become more and more complicated, value of the external criterion, which has a nature of external supplement, will reach to its minimum value demonstrating that the model with optimized complexity comes into exist.

2.2 Obtain data for forecasting

Based on document [4] and prioritization of influences made, comparability of information and requirements of forecasting model, this paper selects the following 9 indexes as input variables for neural network: total output value of agriculture, total retail sales of consumer goods, total export-import volume, total volume of employees' salaries, total volume of investment in fixed assets, consumption of residents, year end balance of savings deposit in rural and urban area, gross domestic product and fiscal expenditure.

Table 1 shows data from 1981—2001 selected from China Statistical Yearbook, measured by 100 million yuan.

2.3 Pretreatment for data through self-organization methodology

Based on self-organization methodology, tax revenue is set as dependent variable and others as independent variables, so that we will establish a model and get the model equation of tax revenue and relevant variables through KnowledgeMiner 5.0, an implementation tool of self-organization GMDH (Group method of Data handing), the model equation is:

Y = -190.456894 -0.033271X7 + 0.105339X2 + 0.736523X9

Coefficient Of Determination (R-squared) R^2=0.9934,Prediction Error Sum Of Squares (PESS) is 0.0079, Mean Absolute Percentage Error is 4.53 %, and Approximation Error Variance is 0.0066.

Fitting of the model proves establishment of a regression model since most of the fitting errors between fitted data and actual tax revenue are below 5% except that of year 1985.

Based on this equation, the variables best reflecting changes of tax revenue are: fiscal expenditure, year end balance of savings deposit in rural and urban area as well as consumption of residents. Study on self-organization methodology shows that the most proper combination should be selected although tax revenue is related to many economic indexes including GDP, because it is not necessarily effective to have more and more indexes or select indexes closely related.

2.4 Establish a combined forecasting model.

1) Establish three single models based on the variables mentioned above.

a) Multiple linear regression model Y1

Research[5] shows that equation acquired from data based on self-organized theory is better in forecasting in a noisy environment compared with the equation of linear regression based on the principle that the sum of square of deviations of observed value and regressive estimation value is the minimum, so the former is used as the equation of linear regression:

Y1 = -190.4569 -0.03327X7 + 0.10534X2 + 0.73652X9

b) Time series model Y2

There are various time sequence models; here we use the Auto-regression (AG) time sequence model which is established through the self-organization data mining tools.

Y2(t)= - 58.077019 + 1.182588Y2(t-1), note: t=2,3, ⋯ 21,Y2(1)=629.89

Test of fitting proves establishment of a regression model since most of the fitting errors between fitted data and actual tax revenue are below 3% except that of year 1985.

c) BP neural network model Y3

Y3=f(X2,X7,X9),note: X2、X7、X9 are variables showed in table 1, three-layered neural network with 5 neurons in hidden layer is selected after training and testing.

Weight matrix of hidden layer is

$$W_1=\begin{bmatrix} 6.2033 & -4.1562 & 9.6011 \\ -21.9743 & 22.1488 & 27.4704 \\ -0.4098 & 5.9621 & 12.7808 \\ -16.1871 & 9.5551 & 3.7225 \\ 6.9283 & -4.0009 & -2.3748 \end{bmatrix}$$

, Threshold value

$$\theta_1=\begin{bmatrix} 4.4801 & 0.0708 & -0.4386 & 0.0831 & -0.7964 \end{bmatrix}^T,$$

Weight matrix of output layer is

W_2= [-11.7519 18.6932 -27.8054 -31.1400 -7.1336],

Threshold value θ_2 = [-11.0320].

Testing shows that the error between its fitting value and actual value is extremely small.

2) Combined forecasting model based on self-organization methodology

"Combined Forecasting" theory was proposed by J.M Bates and C.W.J Granger in 1969 which has been attracting significant attention form the forecasting field [6-8]. This theory, which has been widely adopted, is well-known for its accuracy in forecasting thanks to its nature. According to this theory, various individual forecasting are treated as sections representing different information, and the integration of information is better in reducing uncertainty of both individual forecasting and the whole forecasting, achieving further accuracy. The combined forecasting model is:

Y4=f(Y1,Y2,Y3),the combined forecasting model acquired based on KnowledgeMiner 5.0, using data fitted value and historical tax data:

Y4=47.07665+0.7708Y1-0.22236Y2+0.42303Y3

2.5 Results from forecasting

Table 2 shows data from Chinese statistical yearbook during 2002—2004 concerning fiscal expenditure, year end balance of savings deposit in rural and urban area, consumption of residents and tax revenue. Table 3 is the comparison between forecasted value of tax revenue obtained from the previous 4 models and the actual value. Formula of forecasting sum squares error (FSSE) of is $\sum_{t=1}^{n}(Y_t-\hat{Y}_t)^2$, Y_t is the actual value while $\hat{Y}_t$ is obtained from the model.

3. CONCLUSIONS

Modeling based on self-organization provides access to internal relations between tax revenue and various economic indexes. Linear regression model, time series model and neural network model are reliable in forecasting and combined forecasting methodology can improve forecasting. Tax policy should be taken into consideration during both data selection and treatment since it has substantial impact on tax revenue.

REFERENCES

[1] He Changzheng, Yu Hai. Self-organizing Combination Forecasting Method and Its Application. The Journal of Quantitative & Technical Economics, 2002 (2).

[2] Li Xiaofeng. The Establishment of Economy Forecasting Model Based on GMDH and Artificial Neural Network. Forecasting, 2002 (6).

[3] Madala H.R,Ivakhnenko A.G. Inductive learning algorithms for complex systems modeling [M].Boca Raton,London,Tokyo：CRC Press Inc,1994.

[4] Qi Lili Statistical model of Tax Income and Economic Analysis, Taxation Research Journal, 2005,1

[5] Zheng Mingcui , He Changzheng Comparison between the self-organizing data mining and regression analysis, Systems Engineering and Electronics, 2005.10:1748-1751.

[6] Bates J.M. Granger C.W.J. The combination of forecasts [J].Operational Research Quarterly, 1969.20:451-468.

[7] Granger C.W.J. Ramanathan R. Improved methods of combining forecasting [J].Journal of Forecasting, 1984.3:197-204.

[8] Granger C.W.J. Invited review, combining forecasts-twenty years later, J. of Forecasting.1989, 8.

Table1 Tax Revenue And Relevant Data From 1981—2001

Year	Tax revenue	total output value of agricult -ure	total retail sales of consum -er goods	total export-import volume	total volume of employees' salaries	total volume of investment in fixed assets	Consump- -tion of residents	year end balance of savings deposit in rural and urban area	gross domest -ic product	fiscal expend -iture
	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9
1981	629.89	2181	2350	735	820	961	249	524	4862	1138
1982	700.02	2483	2570	771	882	1230	267	675	5295	1230
1983	775.59	2750	2849	860	935	1430	289	893	5935	1410
1984	947.35	3214	3376	1201	1133	1833	329	1215	7171	1701
1985	2040.79	3619	4305	2067	1383	2543	437	1623	8964	2004
1986	2090.73	4013	4950	2850	1660	2121	485	2238	10202	2205
1987	2140.36	4676	5820	3084	1881	3792	550	3073	11963	2262
1988	2390.47	5865	7440	3822	2316	4754	693	3802	14928	2491
1989	2727.4	6536	8101	4156	2619	4410	762	5147	16909	2824
1990	2821.86	7662	8300	5560	2951	4517	803	7120	18548	3084
1991	2990.17	8157	9416	7256	3324	5595	896	9242	21618	3387
1992	3296.91	9085	10994	9120	3939	8080	1070	11759	26638	3742
1993	4255.3	10996	14270	11271	4916	13072	1331	15204	34634	4642
1994	5126.88	15751	18623	20382	6656	17042	1746	21519	46759	5793
1995	6038.04	20341	23614	23500	8100	20019	2236	29662	58478	6824
1996	6909.82	22354	28360	24134	9080	22914	2641	38521	67885	7938
1997	8234.04	23788	31253	26967	9405	24941	2834	46280	74463	9234
1998	9262.80	24542	33378	26850	9297	28406	2972	53406	78345	10798
1999	10682.6	24519	35648	29896	9875	29855	3138	59622	82068	13188
2000	12581.5	24916	39106	39273	10656	32918	3397	64332	89468	15887
2001	15301.4	26810	43055	42184	11831	37214	3609	73762	97315	18903

Table 2 TAX revenue and relevant data from 2002—2004

Year	Tax revenue (million thousand yuan)	year end balance of savings deposit in rural and urban area（million thousand yuan）	Fiscal expenditure(million thousand yuan)	Consumption of residents(million thousand yuan)
2002	17636	86911	22053	48136
2003	20017	103617	24650	52516
2004	24166	119555	28487	59501

Table 3 DATA from forecasting methodologies

Year		2002	2003	2004	FSSE
Tax revenue（million thousand yuan）		17636	20017	24166	
Y1	Forecasted Value	18231	20049	23081	1530105
	Relative error	3.37%	0.16%	4.49%	
Y2	Forecasted Value	18037	20799	23614	1075926
	Relative error	2.27%	3.90%	2.28%	
Y3	Forecasted Value	18358	19037	24424	1548765
	Relative error	4.09%	4.90%	1.07%	
Y4	Forecasted Value	17971	20119	23524	534296
	Relative error	1.90%	0.50%	2.66%	

A DISTRIBUTED DATABASE MANAGEMENT SYSTEM FOR COLLEGE TEACHING

Xia Weiwei
College of Information, Linyi Normal University
Linyi, Shandong, China
Email: xiaweiwei1999@126.com

ABSTRACT

Oracle databases management system is used as the distributed database system for college teaching in most universities. However, Oracle can implement distributed database technology only after complex data management and configuration. And what's more, it is disadvantage for students to understand basic concepts of distributed database. Aimed at this condition, distributed databases management system (DDBMS) for teaching is designed based on Client /Server(C/S). System architecture and system syntax are presented, and the design of Database Dictionary, Parsing Module, Command Processing Module and Localization Module are stated in detail. The system can meet the requirements of college teaching of distributed database, and has certain significance for college teaching.

Keywords: Distributed Databases Management System, C/S, College Teaching, Homogeny

1. INTRODUCTION

Distributed database environment is difficult to set up because of the extreme complexity of DDBMS, and the prototype systems have not been used widely. For these reasons Oracle database management system is used as distributed database system (DDBS) for teaching in most universities. Oracle supports application of distributed database, but its implementation of distributed database technology needs large investment of hardware and software, complex database link, data management and configuration, such as configuration of TNSNAMES.ORA, LISTENER.ORA, SQLNET.ORA, etc. [1] What's more, it is disadvantage for students to understand basic concepts of distributed database. Therefore it is not suitable to be used for distributed database teaching in colleges and universities. Aimed at this condition, the DDBMS for teaching is designed in this paper. This system is based on C/S, and uses Access as background database system, and supports horizontal fragment, vertical fragment and mixed fragment. Students can understand basic concepts and the process of fragment and allocation in distributed database.

2. SYSTEM ARCHITECTURE

Table 1 shows a comparison of adaptive occasion between C/S mode and Browser/Server (B/S) mode. Because this system is a homogeneous DDBMS used in local area network (LAN), it adopts C/S mode.

The potential advantages of using C/S are [2]: (1) It raises the efficiency of investment, and can use high-performance system platform. (2) It greatly reduces network cost, response time, and requirement of network bandwidth. (3) Friendly interface, good interoperability between platform and software.

The limitation of this mode is: data query efficiency is easy to reduce because of bottleneck on the server. But in multi-server distributed database system, such limitation will be less affected. System architecture is shown in Fig. 1

Table 1. The comparison of adaptive occasion between C/S mode and B/S mode

Comparative contents	C/S	B/S
Used in LAN	very fit	fit
Used in WAN	not fit	very fit
Develop timely	fit	not fit
High-performance database operations requirement	very fit	not fit
High security requirement	very fit	not fit
Heterogeneous requirement	not fit	very fit

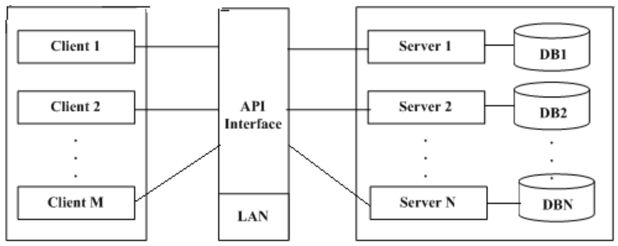


Figure1. System architecture

3. SYSTEM SYNTAX

This system supports 4 kinds of statements, such as CREATE TABLE, INSERT, DELETE and SELECT statements. Although functions of INSERT, DELETE and SELECT statements are not same with its in SQL, the grammars are the same. CREATE TABLE statement is extended here, it not only can realize the definition of tables and views, but also can realize the definition of fragments. For example, create a table student, split table into two horizontal fragments, and store fragment student1 on two sites S1 and S2. Statements as following:

```
CREATE TABLE student
(sno CHAR(5) KEY,
sname CHAR(20),
sdept VARCHAR(15);
SPLIT student INTO
student1 WHERE sdept='Information school'
    student2 WHERE sdept =' Foreign Language School'
ALLOCATE student1 TO S1
ALLOCATE student1 TO S2
ALLOCATE student2 TO S2
```

4. SYSTEM DESIGN AND DEVELOPMENT

This system adopts C++ language, and includes User Interface, Database Dictionary, Localization Module, Parsing Module and Command Processing Module. Fig.2 shows the total framework.

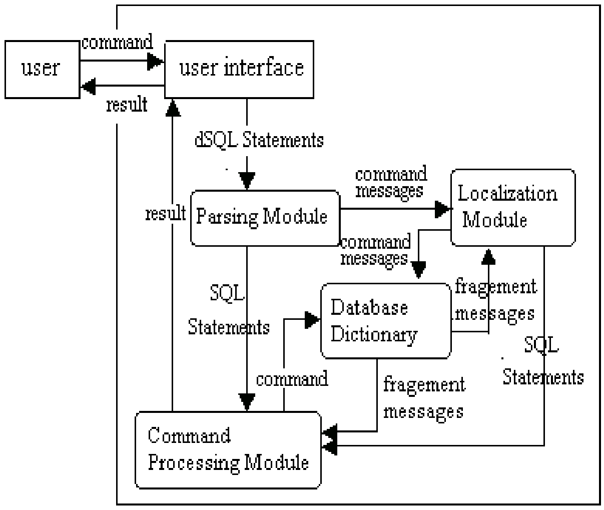


Figure2. System Total Framework

4.1 Database Dictionary

Database Dictionary is a group of tables and views which contain database information, that is, the system directory. It plays a decisive role for the performance of DDBS. The distribution of the database dictionary has three ways: centralized management, fully duplicate management and partition management.

System directory subsystem of this system adopts a compromise approach, in which global data dictionary (GDD) is stored in the database of one server (called super- server), and in databases of other servers (called sub-server) only store local data dictionary (LDD),and LDD only stores data object or register on the local website. This way can prevent the super-server becoming bottleneck, enhance autonomy and reduce the storage cost.

GDD describes the global structure and information of a distributed database. A detailed and accurate GDD is very important for the database. Through the establishment of global data dictionary, get the static properties of the sites and the tables [3], In order to realize fragment and allocation [4], GDD mainly consists of three parts: the first part is a Hash table, recording information of all logic tables; the second part is a Hash table too, recording information of all fragments; the third part records mappings tables of tables and fragments. Because of GDD is stored in the disk, and the accessing frequency to the mapping tables is high, in order to reduce the number of disk read-write, improve efficiency, we adopt B-tree to implement the images' storage. Each tree's root node represents a logical table, a leaf node represents a fragment, a middle node records the domain of the fragment. Codes as following:

```
typedef enum
{
   unknown_type, string_type, number_type;
} data_type;

struct SchemaInfo
{    typedef struct FieldInfo
          {
           string filedName,
           data_type type;
             } filed_info_t;
     vector<field_info_t> fieldInfoList;
     vector<int> keyList;
     vector<SchemaInfo *> foreignKeyList;
     vector<FragmentInfo> * pFragmentList;
   FragmentAction * pFragmentTree;
};

struct FragmentAction
{    typedef enum ActionType { horz, vert, derived, leaf }
     type_t;
     string intermediateName;
     FragmentAction * parent;
     FragmentAction * next;
     int nKid;
     type_t type;
     union {
            predicate_t predicate;
            vector<field_t> fields;
            FragmentAction * derive;
} v;
     FragmentInfo * pFragmentInfo;
};

FragmentNode =
{ FragmentAction * pLeaf;
   SchemaInfo * pSchema;
   vector<int> fileds;
   AllocationSite site;
     card(A), min(A), max(A) [ | Card(A), min(A), max(A) ]
     int nCard ;
}
```

4.2 Parsing Nodule

Parsing Module is one of the main modules of the system, and its processor uses bison. It mainly analyses the grammars which users input, and determines whether they are legal, and calls the correct modules by statements' type. Grammar as following:

```
SQLstatement: select\_statement { call selection processor }
| create\_table\_statement { call Table creation processor }
| insert\_statement { call Insertion processor }
| delete\_statement { call Deletion processor }
```

4.3 Localization Module

The main role of Localization Module is allocating tasks to involved servers that based on information of LDD and GDD.

If the data operations only relates to the local database, then it equates with a centralized database operations; if more than one site are involved in, then GDD is needed to localize servers which involved in the operation. Workflow is shown in Fig. 3.

4.4 Command Processing Module

Command Processing Module allocates tasks to correlative servers, and returns operating summary of the servers to User Interface. If GDD or LDD is involved in the operations, then the dictionaries are operated correspondingly.

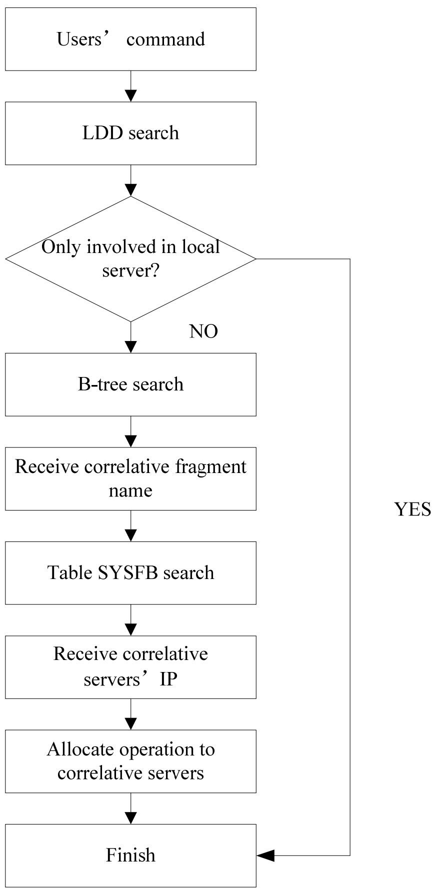


Figure3. Workflow of localization module

5. CONCLUSIONS

Comparing with Oracle DBMS, this DDBMS uses Access as background database, hardware cost is reduced, data management is easier, and does not need complicated configuration. Practice has proved that the system can meet the requirements of college teaching.

REFERENCES

[1] YAO Wenlin, WANG Cungang, LIU Shidong, QIU Like, Design and Technologies of Distributed Database Based On Oracle, Computer Engineering, Vol.32, No.20, October 2006, pp.89~91.

[2] SHAO Peiying, Distributed Database System and Application (second edition)，Science Press，2005.

[3] HE Xin, Cost model of query plan database system, Computer Engineering and Applications, Vol.44, No.13, May 2008, pp.144~146.

[4] Syam Menon, Allocating Fragments in Distributed Databases, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, Vol.16, No.7, JULY 2005, pp.577-585.

ANALYSIS AND OPTIMAL DESIGN OF NETWORKED CONTROL SYSTEM WITH VARYING SAMPLING RATE*

J.R Fan[1], H.J Fang[2]
[1]College of Computer Science, South-Central University for Nationalities
Wuhan 430074, China
[2]Department of Control Science and Engineering, Huazhong University of Science and Technology
Wuhan 430074,China

ABSTRACT

This paper focuses on the state observer design for networked control system with short delay and less noise by varying sampling period, where the sampling rate belongs to some known set of time intervals. The static observer gain can be achieved by solving a set of linear matrix inequalities (LMIs). Moreover, the method guarantees that the covariance of the estimation error will be asymptotically bounded by an upper bound, which can be optimized. The simulation results illustrate the effectiveness of the proposed method.

Keywords: Networked Control System, Time-delay,Varying Sampling Rate, Linear Matrix Inequalities (LMIs).

1. INTRODUCTION

The networked control system (NCS) is a distributed control system wherein (part of) the control loop is closed through a real-time network. However, the insertion of the communication network and share of the limited communication source, will inevitably lead to time delay. The past few decades, delay has drawn much attention from experts and scholars throughout the world. A lot of research results concerning time-delay have existed in the literature, such as the modeling [1], simulation [2], system analysis and design [3], communication protocol and network scheduling [4], etc. For the sake of simplicity, the constant sampling rate is usually adopted and the sampling period usually large enough to avoid network congestion when the network is occupied by the most users. But in NCSs with a limited bandwidth, a nonuniform sampling strategy could achieve better performance than a uniform sampling strategy.

Recently, there are a number of papers considering varying sampling rates of NCSs, Such as Stability[5,6], controller design[7,8,9]. Some control theory have been used in NCS with varying sampling period: H_∞ controller[6,7], neural network control[8,9], Switched control theory[10]，etc.

However, all above papers have ignored the process and measurement noise, which is inevitable for NCS controlled plant. Therefore, in this paper, a static observer gain is chosen to guarantee the stability of the estimation error under any sequence of the allowed inter-sample times. Moreover, the covariance of the estimation error will be asymptotically bounded by an upper bound.

This paper was organized as follow. Section 2 presents the models of NCS with time varying sampling rates. A state observer is set up and its condition for robust stability is proved in section 3. The gain of the observer is obtained by using LMI-based method in section 4. In section 5, an simulation is given. Conclusions are stated in section 6.

*This work was supported in part by the National Natural Science Foundation of P.R.China (60274041)

2. PRELIMINARIES AND PROBLEM FORMULATION

Considering the following continuous linear time-invariant system:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Gw(t), \\ y(t) = Cx(t) + v(t). \end{cases} \tag{1}$$

Where $x(t)$, $u(t)$ and $y(t)$ are the state vector, control input vector, controlled output, respectively. The process and measurement noise, denoted by $w(t)$ and $v(t)$, are assumed to be mutually uncorrelated, Gaussian and have covariances $E[w(t)w(t-s)] = Q\delta(t-s)$ and $E[v(t)v(t-s)] = R\delta(t-s)$. A, B, C and G are known constant matrices. Throughout this paper, matrices, if not explicitly stated, are assumed to have appropriate dimensions. The system (1) is sampled only at discrete sampling instants $t_k \in R, k = 0,1,\cdots,$ with $t_{k+1} > t_k$, $t_k = 0$. The time-varying sampling period will be denoted by $T_k = t_{k+1} - t_k$,and $T_k \in \{T_1, T_2, \cdots, T_i, \cdots\}, i = 1, 2, \cdots, n$, the compact set is known. Network and computation delays can be lumped into one value τ_k denoting the time delay from the instant t_k when sensor samples data from plant to the instant when the data reaches the actuator. Here, it is assumed that $\tau_k \le T_k$ and there is no packet dropout in the transmission of networked control system. Also suppose both controller and actuator are event-driven. During sampling intervals, τ_k keeps a constant.

Given the state at time t_k, input is kept constant in the inter-sampling interval T_k (i.e., a zero-order hold device, ZOH, is present). And during that interval, the ZOH at the actuator is still presenting $u(t_{k-1})$. The state at time $t_k + \tau_k$ is

$$\begin{aligned} x(t_k+\tau_k) &= e^{A\tau_k}x(t_k) + \int_{t_k}^{t_k+\tau_k} e^{A(t_k+\tau_k-\xi)}Bd\xi u(t_{k-1}) + w_k \\ &= e^{A\tau_k}x_k + \int_0^{\tau_k} e^{A(\tau_k-\xi)}Bd\xi u_{k-1} + w_k \end{aligned} \tag{2}$$

So state at time t_{k+1} is

$$\begin{aligned} x(t_{k+1}) &= e^{A(t_{k+1}-t_k-\tau_k)}x(t_k+\tau_k) \\ &\quad + \int_{t_{k+\tau_k}}^{t_{k+1}} e^{A(t_{k+1}-\xi)}Bd\xi u(t_k) + w_k \\ &= e^{A(T_k-\tau_k)}x(t_k+\tau_k) + \int_0^{T_k-\tau_k} e^{A(T_k-\tau_k-\xi)}Bd\xi u_k + w_k \end{aligned} \tag{3}$$

Introducing notations $\Psi(T_k) = e^{AT_k}$, $\Tau(\tau_k) = \int_0^{\tau_k} e^{A(\tau_k-\xi)}Bd\xi$,

$N(T_k,\tau_k)=\Psi(T_k-\tau_k)\mathrm{T}(\tau_k)$, $x_k=x(t_k)$, $u_k=u(t_k)$

Combine (2) and (3)

$$x_{k+1}=e^{AT_k}x_k+e^{A(T_k-\tau_k)}\int_0^{\tau_k}e^{A(\tau_k-\xi)}Bd\xi u_{k-1}+\int_0^{T_k-\tau_k}e^{A(T_k-\tau_k-\xi)}Bd\xi u_k+w_k \tag{4}$$
$$=\Psi(T_k)x_k+N(T_k,\tau_k)u_{k-1}+\Gamma(T_k-\tau_k)u_k+w_k$$

Denoting $\psi_k=[x_k^T\ u_{k-1}^T]^T$ as the augmented state vector, so the discrete time representation of (1) can be described as follows

$$\begin{cases}\psi_{k+1}=\Phi(T_k,\tau_k)\psi_k+\overline{\mathrm{T}}(T_k,\tau_k)u_k+w_k,\\ y_k=\bar{C}\psi_k+v_k.\end{cases} \tag{5}$$

Where the notations:

$\Phi(T_k,\tau_k)=\begin{bmatrix}\Psi(T_k) & N(T_k,\tau_k)\\ 0 & 0\end{bmatrix}$,

$\overline{\mathrm{T}}(T_k,\tau_k)=\begin{bmatrix}\mathrm{T}(T_k-\tau_k)\\ I\end{bmatrix}$ and $\bar{C}=[C\ 0]$ have been introduced.

The variables w_k and v_k denote discrete time equivalents of process and measurement noise, which can be taken to be white, zero mean and uncorrelated. Note that w_k can be taken as $w_k:=\int_{t_{k-1}}^{t_k}e^{A(t_k-\xi)}Gw(\xi)d\xi$. But a similar expression for v_k is difficult to find. So their covariance matrices can be given as follows:

$$Q_i'=\int_0^{T_i}e^{A\xi}GQG^Te^{A^T\xi}d\xi,\ R_i'\cong R/T_i \tag{6}$$

Note that Q_i' is exact, R_i is an approximation. But that is generally accepted [11] and used [12].

For augmented state vector $\psi_k=[x_k^T\ u_{k-1}^T]^T$, the process and measurement noise is only related with state vector x_k. so the augmented covariance matrices can be given as follows:

$$Q_i=\begin{bmatrix}I_{2\times2} & 0\\ 0 & 0\end{bmatrix}Q_i',\ R_i=R_i' \tag{7}$$

3. STATE OBSERVER DESIGN

The equations of observer can been given by:

$$\begin{cases}\tilde{\psi}_{k+1}=\Phi(T_k,\tau_k)\tilde{\psi}_k+\overline{\mathrm{T}}(T_k,\tau_k)u_k+L_k(y_k-\tilde{y}_k),\\ \tilde{y}_k=\bar{C}\tilde{\psi}_k.\end{cases} \tag{8}$$

L_k is a time varying observer gain.

The state estimation error is defined as $e_k=\psi_k-\tilde{\psi}_k$, combine (5) and (8),

$$e_{k+1}=[\Phi(T_k,\tau_k)-L_k\bar{C}]e_k \tag{9}$$

Because of $T_k\in\{T_1,T_2,\cdots,T_i,\cdots\}, i=1,2,\cdots,n$,

So denoting $\Phi_i:=\Phi(T_i,\tau_i)$

Theorem 1: assumed that $Q_i+L_iR_iL_i^T$ is positive definite for all $i=1,2,\cdots,n$. If there exists $P=P^T>0$ and $L_i, i\in\{1,2,\cdots,n\}$ such that:

$$(\Phi_i-L_i\bar{C})P(\Phi_i-L_i\bar{C})^T-P+Q_i+L_iR_iL_i^T\le 0 \tag{10}$$

then 1) In the noise free case ($w=0, v=0$)the estimation error dynamics $e_{k+1}=[\Phi_i-L_i\bar{C}]e_k$ is asymptotically stable under all sequence of admissible inter-sample values taken form $\{T_1,T_2,\cdots,T_i,\cdots\}, i=1,2,\cdots,n$.

2) The covariance of the estimation error $P_k=E[e_ke_k^T]$ is asymptotically bounded by P, i.e., $\exists\varepsilon>0,\exists k_0,\forall k>k_0,P_k<P+\varepsilon I$. Moreover, if $P_{k0}\le P$ for some k_0 then $P_k\le P$ for all $k>k_0$.

Proof (1): $Q_i+L_iR_iL_i^T=[I\quad L_i]\begin{bmatrix}Q_i & 0\\ 0 & R_i\end{bmatrix}[I\quad L_i]^T$ is positive define, $(\Phi_i-L_i\bar{C})P(\Phi_i-L_i\bar{C})^T-P<0$ can be gained. Using Schur complements, for $-P=-P^T<0$, $-P-(\Phi_i-L_i\bar{C})(-P)(\Phi_i-L_i\bar{C})^T<0$ can be obtained.

$$\begin{bmatrix}-P & (\Phi_i-L_i\bar{C})\\ (\Phi_i-L_i\bar{C})^T & (-P)^{-1}\end{bmatrix}<0 \tag{11}$$

By schur complements twice, we can gain $(\Phi_i-L_i\bar{C})^TP^{-1}(\Phi_i-L_i\bar{C})-P^{-1}<0$, which implies that the Lyapunov function $V(e)=e^TP^{-1}e$ for all possible error dynamics as $V(e_{k+1})<V(e_k)$. Hence, the estimation error dynamics is guaranteed to be asymptotically stable under any sampling sequence.

Proof 2): for time-varying observer gain, L_k can be found by using the time varying Kalman filter:

$$L_k=(\Phi_iP_k\bar{C}^T)(R_i+\bar{C}P_k\bar{C}^T)^{-1} \tag{12}$$

Where the state estimation error at each time step is obtained by the well known Riccati recursion formula:

$$P_k=(\Phi_i-L_{k-1}\bar{C})P_{k-1}(\Phi_i-L_{k-1}\bar{C})^T+Q_i+L_{k-1}R_iL_{k-1}^T \tag{13}$$

Combine (9) and (12), the following can be derived:

$$P_k-P\le(\Phi_i-L_{k-1}\bar{C})(P_{k-1}-P)(\Phi_i-L_{k-1}\bar{C})^T \tag{14}$$

If $P_{k-1}\le P$, then it follows that $P_k\le P$ as well. This proves the second statement of the theorem.

Now how to find P and L_i? The problem can be cast into LMIs, such that P and L_i can be obtained.

4. SYNTHESIS USING LMIS

The continuous system (1) and its discretization (5) are robustly stable for the varying sampling rates T_i and its performances is bounded by $\psi^TP\psi$ have been proved. The remaining problem is to obtain the P and L_i. Here the problem is synthesized into a set of LMIs.Taking (10) $(\Phi_i-L_i\bar{C})P(\Phi_i-L_i\bar{C})^T-P+Q_i+L_iR_iL_i^T\le 0$ can be written into

$$\begin{bmatrix}(\Phi_i-L_i\bar{C})^T\\ I\\ L_i^T\end{bmatrix}^T\begin{bmatrix}P & 0 & 0\\ 0 & Q_i & 0\\ 0 & 0 & R_i\end{bmatrix}\begin{bmatrix}(\Phi_i-L_i\bar{C})^T\\ I\\ L_i^T\end{bmatrix}-P\le 0 \tag{15}$$

$\forall i\in\{1,2,\cdots,n\}$

Appling Schur's complement to (15), the following can be obtained:

$$\begin{bmatrix} P & (\Phi_i - L_i\bar{C}) & [I \quad L_i] \\ (\Phi_i - L_i\bar{C})^T & P^{-1} & 0 \\ \begin{bmatrix} I \\ L_i^T \end{bmatrix} & 0 & \begin{bmatrix} Q_i & 0 \\ 0 & R_i \end{bmatrix}^{-1} \end{bmatrix} \geq 0 \qquad (16)$$

$\forall i \in \{1, 2, \cdots, n\}$

Multiplying the above inequality from left and right with $\begin{bmatrix} P^{-1} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$ And setting $W = P^{-1}$, $Y_i = P^{-1}L_i$,the state observer synthesis LMIs as follow:

$$\begin{bmatrix} W & (W\Phi_i - Y_i\bar{C}) & [W \quad Y_i] \\ (W\Phi_i - Y_i\bar{C})^T & W & 0 \\ \begin{bmatrix} W \\ Y_i^T \end{bmatrix} & 0 & \begin{bmatrix} Q_i & 0 \\ 0 & R_i \end{bmatrix}^{-1} \end{bmatrix} \geq 0 \qquad (17)$$

$\forall i \in \{1, 2, \cdots, n\}$

5. EXAMPLE AND SIMULATION

To illustrate the presented the observer design procedure which is robust against the variations in sampling rate, the following system is considered.

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, Q = \begin{bmatrix} 2000 & 0 \\ 0 & 2000 \end{bmatrix}, R = 50$$

Measurements are taken with two different sampling rates and constant short delay $T_1 = 1s$, $\tau_1 = 0.2s$ and $T_2 = 0.5s$ $\tau_2 = 0.1s$. Solving the linear matrix inequalities, the following observer gains can be obtained:

$L_1 = \begin{bmatrix} 1.1684 & 1.0316 & 0 \end{bmatrix}^T$,

$L_2 = \begin{bmatrix} 1.0968 & 1.0381 & 0 \end{bmatrix}^T$.

An asymptotic upper bound of the covariance by $P = W^{-1}$

$$P = 10^2 \times \begin{bmatrix} 3.6927 & 0.4838 & -0.0008 \\ 0.4838 & 1.1877 & -0.0059 \\ -0.0008 & -0.0059 & 1.1118 \end{bmatrix}$$

6. CONCLUTIONS

The paper has described a networked control system with a continuous LTI plant which is sampled at non-equidistant sampling rates. The varying sampling rate belongs to a known set. An observer design which is robust against the variations in the varying sampling rate has been proposed. The synthesis procedure is then formulated in terms of LMIs. An example the synthesis procedure using the proposed LMIs is carried out. The observer gains, which are the solutions to the LMIs can be obtained, and in addition bounds the covariance of the estimation error. Future work may be in the following areas: 1) on the control problem of the data packet dropout. 2) on the stabilization of the networked-induced variable delay that is longer than sampling periods.3) fault-tolerant control for the varying sampling rates.

REFERENCES

[1] Wei Zhang, Michael S.Brannicky, and Stephen M.Philips. Stability of networked control systems. IEEE Trans. Control Systems Magazine, 21(2) (2001) 84-99.

[2] Shawn Hu, Wei-Yong Yan. Stability robustness of networked control systems with respect to packet loss[J]. Automatica, 43(7)(2007)1243-1248.

[3] De-ming Kong, Huajing Fang. Stability Design for a Class of Networked Control system. Transaction of Electric Machines and Control, 1(4)(2007)398-401.

[4] Yan Wang, Qing-wei Chen, Xiao-bei Wu, Wei-li Hu. Co-design of Dynamic Scheduling Strategy and Controller in Networked Control Systems. Control and Decision, 22(6)(2007)680–684.

[5] Fuchun Liu, Yao Yu, Fenghua HE, Songlin Chen. Stability analysis of networked control systems with time-varying sampling periods. J Control Theory Application. 6(1) (2008)22-25.

[6] Yulong Wang, Guanghong Yang. H_∞ controller design for networked control systems via active-varying sampling period method. Acta Automatica Sinica, 34(7) (2008) 814-818.

[7] Zhanlei Shang,Xunlin Zhu，Yiqun Deng. Controller design for networked control systems with passive varying sampling period. Microcomputer Information，23(8)(2007)128-130.

[8] Yan Xue, Ke Liu. Analysis of variable-sampling networked control system based on neural network prediction. Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition,Beijing, China, 2-4 Nov. 2007.

[9] Jiangang Liu, Biyu Liu, Ruifang Zhang, Meilan Li. The new variable-period sampling scheme for networked control systems with random time delay based on BP neural network prediction. Proceedings of the 26th Chinese control conference July 26-31, 2007,zhangjiajie,Hunan, China.

[10] Hongli Zhang, Yuanwei Jing, Siying Zhang. Parameter uncertainty-based controller design for networked control systems with active varying sampling period.

[11] A. Alessandri and P. Coletta. Switching observers for continuous-time and discrete-time linear systems.In Proc. of the American Control Conference. (2001)2516-2521.

[12] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. Linear Matrix Inequalities in Control Theory, volume 15 of Studies in Applied Mathematics. SIAM, 1994.

Jinrong Fan is with College of Computer Science, South-Central University for Nationalities, Wuhan, Hubei 430074, China,e-mail: fanjrock@ hotmail.com).

Huajing Fang is with Dept. of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, 430074, P.R.China (e-mail:hjfang@hust.edu.cn

THE DESIGN OF LARGE SCREEN DISPLAY CONTROL SYSTEM BASED ON FPGA

Shihong Qin[1], Huanhuan Li[2], Kai Chen[1]
[1]Electrical and information Engineer Department, Wuhan Institute of technology
Wuhan, 430205 P.R.China
[2]Electrical and information Engineer Department, Wuhan Polytechnic University
Wuhan, 430023 P.R.China
Email: qinsh@mail.wit.edu.cn,123ilylcf@163.com

ABSTRACT

This paper presents a scan control method of large screen display, which is based on the design platform of programmable logic devices (FPGA). First according to the system general design requirements, the structure and principle of system are analyzed. The methods of the top-down designing and modularizing are adopted .Using a modular approach to design the FPGA chip's internal functions , the LED screen display function is achieved.

Keywords: FPGA, Large Screen, Scan Controller

1. INTRODUCTION

In the past, the large screen display system is realized by medium-small scale integrated circuit. The system has many features including large volume, debugging modifying difficulty. With the development of the technology of lager scale IC and programmable logic circuit, the device performance is greatly improved. The programmable logic device has good characteristics of high integration, high speed, high reliability, system-on-board programmable (SOP). It is not only satisfied with the large screen system high speed data transmission requirement, but also improved the circuit performance and increased flexibility in circuit design. According to the actual applications, some functions of hardware can be obtained. By using flexible modified hardware description language programming without modifying circuit hardware designing. Thus, design to cycle is shorten and cost is reduced. Based on the above advantages, programmable logic devices are widely used in the complex circuits design .In this paper, FPAG is used to achieve the large screen display system's main functional module, so that the performance of the system can be improved.

2. THE STRUCTURE AND PRINCIPLE OF SYSTEM

At present, LED display system is made of three parts, which are the PC, display control circuit and LED display screen. PC is used as upper computer to control and to manage lower computer in control. It completes the display date sending to display control circuit and the display effect setting to LED display screen .The focus of this paper is on the research of the display control circuit. The display control circuit is composed of input-output circuit, data converter circuit, signal control circuit and scan drive circuit.

ARM [1] and FPGA are the core of the display control circuit. They jointly complete the data conversion, the sending of control signal, the control of the LED display screen dynamic scanning [2]. And the part of data conversion control signal is controlled by ARM. The scan drive circuit of LED display is controlled by FPGA. The data signal received by ARM is stored in the two different external extended SRAM. It converts the data signal which reading from SRAM to LED on-screen display by serial transmitting. The structure of system is shown as Figure.1

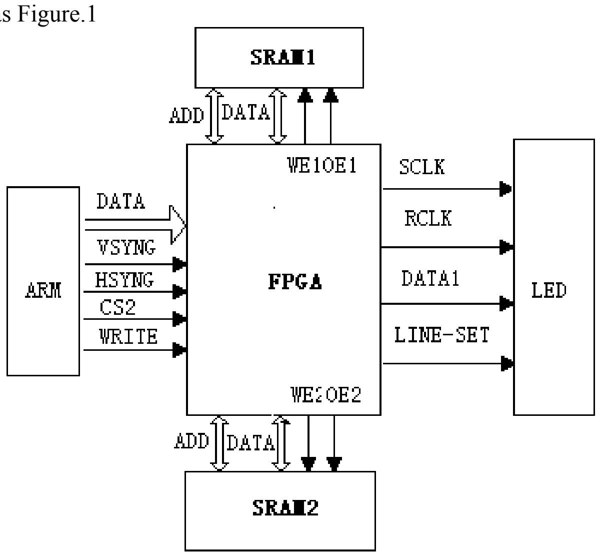


Figure 1. the structure of system

Data storage uses dual addresses memory [3]. The memory capacity is divided into data-storage date address and scan-data address. The scan drive controller reads data from scan-data address. Data-storage date address and scan-data address are switched dynamically. After one frame image completely stored, the data-storage date address is converted to scan-data address automatically [4]. Then the next frame image is stored in the original scan-data address. Using dual addresses memory can effectively prevent the incompletely scan image data caused by the failure of data sending.

3. DESIGN OF FPGA INTERNAL FUNCTION

FPGA internal function module is shower as in Fig.2.FPGA series as the scan drive. Its internal solidification digital logic is responsible for producing the screen display control signal, serial shift clock, line latch signal, line selection signal, etc. Two SRAMS of its external configuration are used to read display data by time-sharing. At the same time, FPGA reads display data signal from SRAM, and converts it into on-screen data .Through serial output to the corresponding signal Data Bus

FPGA internal circuit is mainly composed of read/write Address Bus, read/write address controller, read/write data controller, line scan Address Bus circuit, etc. FPGA sends the signals which include display data(DATA),system clock(CLK),frame synchronization signal(VSYNC), Linesynchronization signal (HSYNC)to the read/write

Address.Bus and read/write data controller, the sends data in SRAM1 and SRAM2 to storage[5]. SRAM1 and SRAM2 reads and writes alternately .The static storage's state of reading and writing controlled by system clock, frame synchronization signal, Line synchronization signal and chip select signal .The reading Address Bus is used to calculate the stored address, which is needed data information stored, to ensure the large-screen LED display correctly.

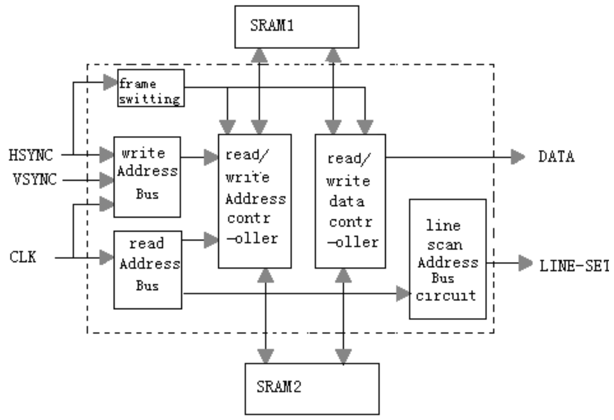


Figure 2. FPGA internal function module

4. THE ANALYSIS OF SOFTWARE SYSTEMS

The design method that is from the top-down and modularized is adopted .According to the structure of the module and overall module, the top-level design document is designed. Top-level design document is to put various modules together, and to form an easy-to-read graphic mode.

The whole realization of the digital control logic contains three sub-modules: signals from ARM controller, the signals connected with two SRAM, the control signals to the LED display screen.

(1) The signals from ARM controller: system clock (CLK), frame synchronization signal (VSYNC), line synchronization signal (HSYNC), display data (DATA).
(2) The signals connected with two SRAMS: 8 bits bidirectional data bus of SRAM1(SRAM1--DATA), 8 bits bidirectional data bus of SRAM2 (SRAM2—DATA), 16 address bus of SRAM1 (ADD1), 16 address bus of SRAM2 (ADD2) , reading enable signal of SRAM2 (OE2), writing enable signal of SRAM1 (OE2), writing enable signal of SRAM1 (WE1), reading enable signal of SRAM1(WE2).
(3) The control signals to the LED display screen: the line selection control signal of LED screen drive circuit (LINE--SET), the serial clock of LED screen circuit (SCLK), the latch signal of LED screen circuit (RCLK), display data signal (DATA1).

As shown in that Figure 2 FPGA internal circuit mainly includes read/write Address Bus、read/write data controller, line scan Address Bus circuit. Each module has its realizing circuit. According to that requirement, the scan program of the LED display screen is divided into the following process:

(1) Clock process. Produce kinds of frequency information.
(2) Reading RAM process. Sending the display date to the LED display screen.
(3) Data latch process. Latching the data which send to the LED display screen to the data latch in screen.
(4) Line selection process. Selecting scan control signal of every line.
(5) Controlling the control signal of microcontroller's transmit data.

5. CONCLUSIONS

Through the analysis and design of the whole design, the all features of display control module are integrated into an FPGA chip. Using a modular approach, the FPGA chip's internal functions are designed in the paper. The LED screen's display function is achieved by using the methods of the top-down and modularizing. The large and more complex display control is divided into some simple functional modules

REFERENCES

[1] ARM Limited, ARM Development Suite Assembler Guide.
[2] Deng Chunjian, "Design and application of LED large screen display universal offline control card based on FPGA and ARM-μClinux", Chinese Journal of Electron Devices, Vol.28, No.4, Dec 2005, pp710-713.
[3] Liang Zhiming, Li Bin, "Design of scan controller in large LED screen based on FPGA circuit", Chinese Journal of Liquid Crystals and Displays, Vol.22, No.4, Aug 2007, pp 477-481.
[4] Holm, Paige, Rhyne, William, "Two-Dimensional LED Arrays for Virtual Display Image sources", IEEE Transactions on Electron Devices,Vol.46,No.5,May 1999,pp 897-904
[5] W. Kurdthongmee, "Design and implementation of an FPGA-based multiple-color LED display board", Journal of Microprocessors and Microsystems,Vol.29,No.7,Sep 2005,pp327-336

THE PENALTY-FUNCTION METHOD FOR FEM ANALYSIS OF A LONG-SPAN REINFORCED CONCRETE CABLE-STAGED BRIDGE *

Wang Yingjun, Li Zhuoqiu, Zhu Sirong
School of Science, Wuhan University of Technology
No.122 Luoshi Road,Wuhan430070,P.R.China
Email: wyj95950@126.com

ABSTRACT

When establishing the finite element model (FEM) for a large reinforced concrete (RC) cable-staged bridge, inevitably, there will be a difference between the FEM and the real structure because of its complexity. To make the model accurate, the FEM should be updated according to the result of dynamic response of bridge under ambient condition. The model includes 4 types of elements such as SOLID45 element that represents reinforced concrete beam, SOLID65 element that represents the pre-stress steel-reinforced concrete towers, LINK10 element that represents high stress cables and LINK8 element represents pre-stress rebar. By using the penalty-function optimization method and studying the dynamic mechanical property of Wuhan No.2 Yangtze River Bridge (Cable-staged Bridge), this paper creates the bridge's 3D finite element model, calculates the sensitivity of model parameters and updates the FEM model parameters. The updated model provides a basis for simulating dynamic response of the bridge

Keywords: Penalty-Function, Dynamic FEM, Model updating, Cable-Staged bridge.

1. INTRODUCTION

One of the important aspects to monitor and evaluate the health condition of large bridges is to understand its dynamic properties which are determined by their modal properties, such as natural frequency, mode shape, damping and so on [1-3]. The analytical and real values of these parameters can be derived by FEM analysis and mode test. However, as for structure nature frequency, there is a large difference value between analytical and real results because of the following reasons: (1) error of discrete continuous system; (2) uncertainty of structural geometry and boundary conditions; (3) inhomogeneity of material properties; (4) error of test data. The former three causes are relevant with the assumptions of the structural finite element model; the fourth can be attributed to the test methods. Assuming that the measured value of modal characteristics is very close to the real one, the issue to be resolved is to amend the structure FEM to make the difference between analytical and measured values of the structural modal characteristics close to zero.

Matrix updating and parameter updating are the two widely used methods for FEM updating [4-5]. These methods have been verified in simple structures such as simple supported beam and cantilever truss. However, when come to long-span cable-staged bridges, suspended bridges, it becomes more difficult to FEM updating because of the uncertainty properties of its materials, geography and joint conditions [6-8].

2. A DESCRIPTION OF WUHAN NO.2 YANGTZE RIVER BRIDGE

The Wuhan No.2 Yangtze River Bridge is located in Wuhan, Hubei Province, China over the Yangtze River. The bridge is a 5-span RC cable-stayed bridge with an overall length of 770 m, main span of 400 m, edge span of 180 m and cantilever length of 5 m. The bridge, having 6 lanes, carries two roadways with 30 m wide. The two H-type towers are made of reinforced concrete with a height of 150 m and a clear navigation of 40 m. The cable is a fan type in both planes. There are totally 24*8=192 stay cables. The cables are composed of a number of strands. Strand number varies from 27 to 85 per cable in eight groups. One strand includes 7 high strength wires with the diameter of 5 mm.

3. FINITE ELEMENT MODELING OF THE BRIDGE

To perform the analytical modal analysis of the Wuhan No.2 Yangtze River Bridge, a 3D finite element model has been constructed using ANSYS finite element software. The model represents the bridge in its current as-built configuration with the geometry and structural properties estimated from the design drawings. Four types of finite elements are employed to model the different structural members in Fig.1.

All stay cables are modeled by 3-D tension-only truss element (LINK10) since they are primarily designed to sustain tension forces. LINK10 is a 3-D spar element having the unique feature of a bilinear stiffness matrix resulting in a uniaxial tension-only (or compression-only) element. With the tension-only option, the stiffness is removed if the element goes into compression (simulating a slack cable or slack chain condition). This feature is useful for static guy-wire applications where the entire guy wire is modeled with one element. It may also be used in dynamic analyses (with inertia or damping effects) where slack element capability is desired but the motion of the slack elements is not of primary interest.

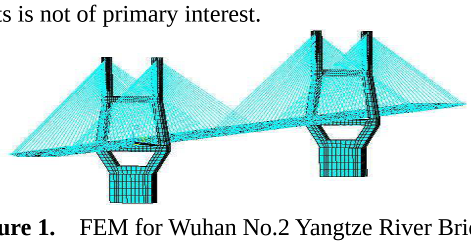


Figure 1. FEM for Wuhan No.2 Yangtze River Bridge

The Pre-stressed rebar is modeled by LINK8 of which is similar with that of LINK10. LINK8 is a spar which may be used in a variety of engineering applications. This element can be used to model trusses, sagging cables, links, springs, etc. The 3-D spar element is a uniaxial tension-compression element with three degrees of freedom at each node: translations in the nodal x, y, and z directions. As in a pin-jointed structure, no bending of the element is considered. Plasticity, creep, swelling, stress stiffening, and large deflection capabilities are included.
The common rebar is also modeled by LINK8.

The two concrete towers are modeled by Solid65 which is a pre-stressed concrete element with 8 nodes and 3D structure. Each node has 3 translational freedoms. It can not only

* The work was supported by the National Natural Science Foundation of China (No. 50878169)

simulate the fracture under tension and the deformation when applied pressure of the fracture occurred, the pressure on the circumstances of deformation, but also the flow and creep occurred in the concrete.

The main beam is modeled by Solid45. Similarly, it is used for the 3-D model of solid structure. The element is defined by 8 nodes having 3 degrees of freedom at each node: translations in the nodal x, y, and z directions. The element has plasticity, creep, swelling, stress stiffening, large deflection and large strain capabilities.
According to this division, it is a big scale model with a total node of 780,208 and a total element of 321,016.

4. UPDATING METHOD OF THE MODEL

4.1 Sensitivity Analysis

The error of the FEM parameters generally results from the uncertainty of material, geometric parameters and boundary conditions. All of the structure parameters can be modified to update the model. However, it is unpractical and infeasible for the Wuhan No.2 Yangtze River Bridge, which is a large cable-staged bridge with complex structures. Moreover, if a parameter that is incorrect has not been selected to the parameter sets to be modified, it may come to an unreasonable updating result in Fig.2.

Considering the good control of the construction process of the Wuhan No.2 Yangtze River Bridge, we can assume that the geometric parameters, material Poisson ratio and linking boundary of the beam, cable and tower in the initial FEM are accurate while physical parameters of these structures, such as Young's modulus and density is of uncertainty and discontinuity which need to be modified. Taking into account the tiny variation of the mass density of stay cables, pre-stressed rebar and common rebar, only the Young's modulus of these three kinds of material are modified to reduce the work. Therefore, the updating parameters and their initial estimate values, upper and lower limit values of the modification are shown in table 1.

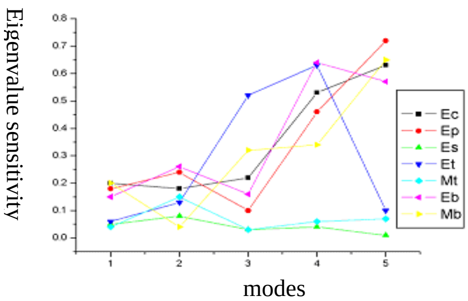


Figure2. Eigenvalue sensitivity of the initial FEM

To determine the 7 updating parameters above, some sensitivity analyses are carried as follows:
Letting E equals to E_0 in the FEM, we obtain the corresponding frequency $\lambda_i^{'}$ (i=1,2,…,m) by the finite element program; Letting $E=E_0+\Delta E$, while keeping other parameters constant, we can obtain another set of frequencies $\lambda_i^{''}$ (i=1,2,…,m).
For small ΔE, the first order sensitivity can be estimated as:

$$\frac{\partial \lambda_i}{\partial E}=\frac{\lambda_i^{''}-\lambda_i^{'}}{\Delta E} \tag{1}$$

Fig.2 shows the sensitivity results of the 7 parameters. It reveals that the Young's modulus of common rebar and the mass density of the main tower relatively have a minor impact on the structural frequencies. Therefore, they can be disregarded and the left 5 parameters need to be updated.

4.2 Optimized Calculation

In order to obtain the accurate FEM, the relative error of natural frequency of the analytical was selected as the objective function and the first five intrinsic frequencies were selected as the state variables [9]. The iterated operation was conducted by using ANSYS software optimization module.
This paper chooses the first order optimization method as the basic method while other methods play the auxiliary role. A group of optimal solution was yielded after each iterating, which was summarized in the following form:

$$\begin{aligned}&\underset{x}{Min}\|E(x)\|_2^2,\ E(x)=\{f_e\}-\{f_a(x)\}\\&s.t\qquad x_l\le x\le x_u\end{aligned} \tag{2}$$

Where f_e、f_a represent test mode and analysis mode of the structure mode respectively, $E(x)$ is the residual and x is the variable parameters to be updated, of which the upper and lower limit value are x_l, x_u respectively.

Apparently, the residual E, $\{f_a\}$ and physical matrix are the implicit function of the updating parameters. Based on sequential unconstrained minimization technique (SUMT), it can be transferred into a one-dimension and unconstrained single-objective optimization problem. The penalty function is as follows:

$$Q(X)=\|E(x)\|_2^2+w\sum_{j=1}^{n}p(x) \tag{3}$$

Where, w is the penalty factor of the constraint-based design variable and status variable. Employing gradient method for the unconstrained optimization and the iterative formula is

$$X(j+1)=X(j)+s_j d(j) \tag{4}$$

Where, s_j is the optimal step length. The convergence condition of iteration is as follows:

$$\left|f^{(j)}-f^{(j-1)}\right|\le\tau \tag{5}$$

Where τ is the tolerance of the objective function. Here we set 5% as its value.

4.3 Model Updating Results

By conducting four iterations, the error of analytical frequency and measured frequency can converge to the set error range which was shown in table 3. From the updated results, elastic modulus of high-strength steel wire cable and pre-stressed reinforcement and mass density of the main beam reduce while

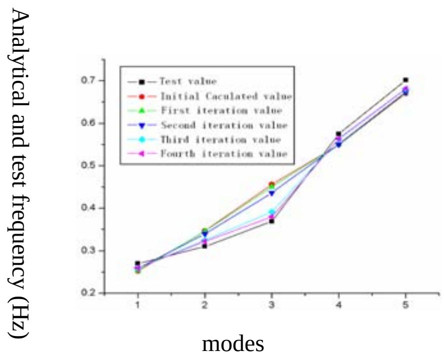


Figure3. Analytical and test frequencies of the initial and updated FEM

elastic modulus of the main tower and main beam go to the opposite trend. The greatest updated parameter is elastic modulus of the main tower the value is up to 15.29%, which might due to the low presuming initial value.

5. CONCLUSIONS

In this paper, a 3D FEM model was established for the Wuhan No.2 Yangtze River Bridge. Optimization method was employed to update the FEM model based on the vibration test results. The updated model has a better reflection to the dynamic characteristics of the bridge structure and provides a solid basis for the dynamic response analysis.

Table1 The updating parameters and their initial estimate values, upper and lower limit values

Location	Element	Material	Structural parameter	Symbol	Initial estimate values	Upper limit value	Lower limit value
Stay cable	Link10	high tensile steel wire	Young's odulus	E_C	1.98E11 Pa	1.8E11 Pa	2.05E11 Pa
Pre-stressed rebar	Link10	high tensile steel wire	Young's odulus	E_P	2.1E11 Pa	1.9E11 Pa	2.15E11 Pa
Common rebar	Link8	Steel wire	Young's odulus	E_S	2.1E11 Pa	1.9E11 Pa	2.15E11 Pa
Main tower	Solid65	RC	Young's odulus	E_T	3.4E10 Pa	3.0E10 Pa	4.0E10 Pa
			mass density	M_T	2600 kg/m^3	2340 kg/m^3	2860 kg/m^3
Main beam	Solid45	RC	Young's odulus	E_B	3.5E10 Pa	3.4E10 Pa	4.0E10 Pa
			mass density	M_B	2600 kg/m^3	2340 kg/m^3	2860 kg/m^3

Table 2 Analytical and test frequency comparison of the iteration model

Vibration mode	Frequency (Hz)						Error (%)
	Test value	Initial analytical value	First iteration	Second iteration	Third iteration	Fourth iteration	
Vertical bending (Symmetric)	0.2698	0.25204	0.2531	0.2583	0.2591	0.2595	-3.8
Lateral bending (Symmetric)	0.3097	0.34663	0.3461	0.3398	0.3248	0.3216	3.8
Vertical bending (Antisymmetric)	0.3687	0.45575	0.4498	0.4365	0.3912	0.3801	3.1
Vertical bending (Symmetric)	0.5750	0.54950	0.5501	0.5503	0.5625	0.5638	-1.9
Torsion (Symmetric)	0.7012	0.6699	0.6721	0.6736	0.6802	0.6812	-2.9

Table 3 Correction Value of Structure Parameters

Location	Symbol	Initial estimate values	Iterated value	Variation (%)
Cable	E_C	1.98E11 Pa	1.92E11 Pa	-3.03
Pre-stressed rebar	E_P	2.1E11 Pa	2.03E11 Pa	-3.93
Tower	E_T	3.4E10 Pa	3.92E10 Pa	15.29
Beam	E_B	3.5E10 Pa	3.74E10 Pa	6.86
	M_B	2600 kg/m3	2583 kg/m3	-0.65

REFERENCES

[1] SALANE H J. Identification of model properties of bridge. [J] Journal of structural engineering.1990,116(7)

[2] Sanayei M, Saletnik M J. Parameter estimation of structure from static strain measurements I: Formulation, [J] Journal of Structrual Engineering, ASCE, 1996, 122(5), 555-562

[3] Ren,W.X. and Peng,X.L. Baseline Finite Element Modeling of a large Span Cable-Staged through Field Ambient Vibration Tests. Computers and Structures,2005,Vol.83, No.8-9,pp536-550

[4] Yuan Wancheng, Cui Fei, Zhang Qiwei. Current Research and Development of Structural Health Monitoring and Condition Assessment for Bridges. [J]. Journal of Tongji University. China.1999,27(5)

[5] Curran P, Tilly G. Design and monitoring of the Flint shire Bridge, UK [J]. Structural Engineering International, 1999,3:225-228

[6] Friswell, M.I , Mottershead, J.E. Finite Element Modal Updating in Structural Dynamics. Kluwer Academic Publishers, 1995.

[7] Fan Lichu, Yuan Wancheng,Zhang Qiwei. Sensitivity Based-FE Model Updating of A Suspension Bridge. [J] Journal of Civil Engineering,China.2000,33(1)

[8] Zhang Qiwei, Fan Lichu. Damage Detection for Bridge Structure Based on Dynamic and Static Measurements. [J]. Journal of Tongji University. China.1998,26(5):528-532

[9] Doebling,S.W. Minimum-Rank Optimal Update of Element Stiffness Parameters for Structural Damage Identification. [J] Journal of American Institute of Aeronauties and Astronauties, 34(1):2615-2621, 1996.

Wang Yingjun is a Associate Professor and the head of Computing Mechanics Teaching and Research Section, School of Science, Wuhan University of Technology. He graduated from Wuhan University in 1989; master from Dalian University of Science and Technology in 1994. Ph.d from Wuhan University of Technology in 2006. His research interests are in Parallel 3-D Finte Element Analysis & computing mechanics.

THE INTEGRATION AND APPLICATION OF RESOURCES FOR DISTANCE EDUCATION TEACHING BASED ON DISTRIBUTED TECHNOLOGY *

Cheng Jiejing, Liu Xiaoxiao
Education Department, Nanchang University
Nanchang, Jiangxi, China
Email: chengjj2001@yahoo.com.cn

ABSTRACT

Firstly, the impact of distributed computing technology on distance education, the systematic model structure and the process of distributed computing technology are introduced. Secondly, the sharing platform of distributed distance education resources and the program of distributed distance learning service system are discussed. Thirdly, we analyzed many difficulties and problems such as how to integrate resources for distance education teaching based on distributed technology, how to formulate plans and measures for education resources, how to install the Virtual server and interface server, and so on.

Keywords: Teaching resources, Distributed technology, education, Education resources.

The distributed sharing of Internet resources and computing power is a significant subject at home and abroad in recent years. The utilization of computing resources has been in a state of imbalance on the Internet. In addition, with the rapid development of digital technology and the Internet, especially Web 2.0, the amount of data on the Internet grows in high speed, which leads to the relative lack of data processing capacity. How to realize the distributed sharing of resources and computing capacity and how to deal with the current high-speed growth of data on the Internet, are problems needed to be tackled as soon as quickly at present. In such a developmental context, it is imperative that the application of distributed technology in the integration of the distance education teaching resources is discussed.

1. INTRODUCTION

Distributed computing technology is a kind of applied technology which involves many areas and practical problems. Distributed technology is the development of Parallel Computing, Distributed Computing and Grid Computing or the realization of these concepts of computer sciences. Distributed technology is result of the mixed evolution and rise of such concepts as Virtualization, Utility Computing, IaaS, PaaS, SaaS, etc. In short, complex data-processing (such as running software), which is originally completed on a local PC or independent remote server, is completed by a large number of distributed computers in synergy now. And this process is delivered to our desktop as a service through the Internet. Any user all over the world can obtain such services in the local via the Internet only by one simple access equipment, realizing that equipments with low allocation can enjoy high-performance computing services.

It is the basic principle of distributed technology that the operation of distributed data centers will be more similar to the Internet by making the calculation distribute in a large number of distributed computers, rather than the local computer or remote servers. Applications required by users do not need to run on terminal equipments such as user's personal computers, mobile Phones, etc, but running on large-scale server clusters on the Internet. The data processed by users is not stored in the local, but stored in the data center on the Internet. The enterprises, which provide distributed computing services, are in charge of managing and maintaining the normal operation of the data center, and ensuring that strong enough computing power and large enough storage space are available for users. At any time and any place, users can connect to the Internet terminal equipments at will. At the same time, the function of user terminals will be greatly simplified, and many complex functions will be transferred to the network behind the terminal for completion. All the said above allows companies to switch resources to applications which are in need of these resources, and access to computers and storage systems, according to the demand. To understand distributed technology, it is necessary to understand the developmental history of application at first. For a computer connected to the Internet, the applications it can use are traditionally divided into two kinds: One is the local application which refers to the applications which compute and store on the local by using local CPU and hard disk resources, such as MsWord，excel, the other is the web application which refers to the applications which compute and store by using servers on the network, such as Web-based mail system and search engine. However, with the development of web technology, web applications become increasingly complex, the traditional browser and html language has been gradually difficult to meet the needs of all applications. The browser added the support for scripting language (javascript or vbscript), and then there appeared so-called fat client, namely web client applications which are not existing browser-based or are further developed by making the existing browser the core. The complexity of local applications and the emergence of fat client make the requirement for the performance of personal computers increasingly demanding, and this growth rate is much higher than the out rate of personal computers, which leads to two results: one is that some old, still using computers receive more and more limitations, even they are eliminated in advance, The other is that some professional applications are so complex that the highest performance personal computers are needed to deal with them, correspondingly additional costs have to be paid. The concept of distributed technology focuses computing and storage on the network, and simplifies local applications and the fat client to the browser with only one supported script, so that minimize the performance of personal computers and maximize the function of personal computers [1]. The structure of distributed technology is shown in fig.1.

As shown in Fig. 1, in the blueprint of the applications of distributed technology, users only need one terminal, such as a monitor or a terminal platform, to realize all the functions and operations needed by users via Web services. For users,

* Supported by Major Subject of Educational Science of Jiangxi "Eleventh Five-Year Plan " (08ZD069):Theory Research of Jiangxi Digital Resources for Distance Education Teaching.

distributed technology puts all the possible power and resources together to provide for each user to utilize.

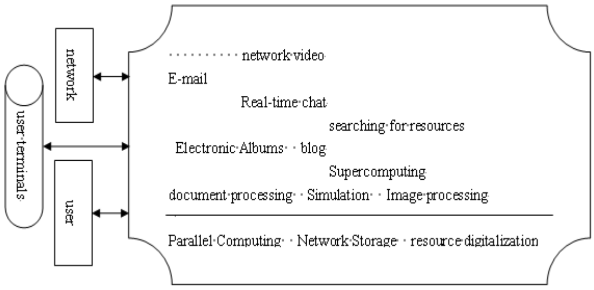


Figure 1. Structure of distributed technology.

2. IMPACT

Distributed computing model can greatly decline the cost of construction of educational information system. For distance education, it will cost much to build computing center and it's hard to patch with the fast growing of educational information system and the various requirements of the service. The distributed computing model offers the distance education a suitable reference plan, the respecting mission of the educational department data centre and Internet centre can choose the distributed computing service to fulfill them and we can save cost without purchasing expensive hard devices and affording frequent maintain and up-grated. Meanwhile, the distributed computing technology will effectively dismiss the phenomenon of the "lonely island" in the educational information system. The influences that the distributed computing technology affecting on distance educational area mainly manifest at the following aspects:

2.1 The Transformation of the Sharing Ways of Teaching Resources

Although traditional Internet teaching system and the educational teaching system based on the Internet both put emphasis on the sharing. This kind of sharing still remains in the lower level. From the teaching resources, it's so diverse that hard to coordinate and manage universally and effectively, from the organization of resources, the degree of resources redundancy is high and the using ratio of valid resources is low, from the hardware and operating system, various different construction systems exist together that many distance teaching system can't processing at different platforms, from the ways of realization, most of them are slack or dynamic Web pages without showing the concept of service. However, the Internet distance education offers users the universal service to realize higher level sharing which not only embodying on the teaching resources but also including every kinds of hardware resources [2]. It's unnecessary to have a knowledge of who and how to provide the studying service for learners whereas the developers of teaching resources can realize more quick development through the unify joint. The distributed long-distance educational system takes the colleges and scientific studying institutions as its key knots and the companies and groups that specializing in Internet education could join into the system of the distributed long-distance education.

2.2 The Transformation of the Organizing Ways of Teaching Resources

It will be changed hugely when the distributed computing technology enters into the distance educational field. As we mentioned before, the main problem that the distributed technology will resolve is the resources sharing issue in the huge number, the great power and the large scale. Actually the distributed technology is the realization of the management of the huge number distributed resources, and the resources sharing we talk here is no longer the simple resources connection and the singular using but to solve the problems that the users want to tackle through the mutual connection and coordination to fulfill the fresh requirements of the users.

In the distributed studying environment, learners don't need to consider where the resources what they learned came from nor think which university they belong to, and the only fact is that what he got is the best education resources that selected strictly by the distributed educational system. The positioning of the studying resources would be fair accurate with the studying within the environment of the distributed educational system. Teacher can provide resources according to students' features while students can easily get what they need and are different from others and these resources all are directly expressing to the most suitable students with the most proper way.

The resources in the distributed educational environment is distribute, separate construction and dynamic and was possessed by different organization or personal, each of them has their own resources management strategy and distinct visiting expense model, consequently managing distributed distance education resources is quite complicate, and the resources finding, resources attempering and deploying is the stress point of current distributed technological application studying. The distributed educational system essentially is a kind of fundamental infrastructure which permits the resources and service accessing without relation to the position and they are provided by the machine and Internet lined geographically, the basic operation that supports the non-related position computing is resources finding. In order to finish the mission that customers submit and the requirements that customers' proposal, we should match all the usable resources of the distributed education and figure out the best and the most reasonable resources deploying way and resources attempering strategy [3].

2.3 The Transformation of Teaching Ideology

1) Traditional Internet resources mainly focus on the teaching and offers teachers the teaching materials, PPT that being used as teaching. However, the distributed education transfer the design and development of the resources of the teaching materials, PPT that focus on teaching into the structure category resources design of Internet curriculums, researching study column and certain degree of organization while the construction of resources content into supporting learning and is able to suit for the educational ideology posed recently.

2) The large scale quick computing competence that the distributing distance educational system possesses also offer the technical basics of making simulate experimental operating PPT. This will enhance the practice of study largely, as a result, the Internet distance education enters into a new stage on the side of experimental courses. Besides, the issues that the distributed distance education resources sharing have to resolve includes: ① the issue of lacking of the data transferring ability of Internet, ② the issue of intellectual property in the distributed distance education resources, ③ the issue of the security of the distributed computing technology.

2.4 Reduce the Purchasing and Maintaining Cost of the Hard Devices of Computers and Internet Transition

Currently, each grade school has equipped with large amount of computer and Internet devices and has to often purchase new computer and Internet devices in order to satisfy more and more computing needs. The intrinsic characteristics of the distributed computing technology make it being accepted more easily than other new technology. If we use the distributed computing service, the major computing mission hand over to the computer server to finish and the distance educational model only make the computer access into the Internet. The distributed computing technology can put the memories, storages and computing competence that distribute at the great deal of the distributed computers all together into a visual resources pool and provide the service of Utility Computing to the Internet users. The characteristic that the distributed computing technology has lower requirements for the device of users decides that the distributed computing technology will be great popular in the field of distance education and reduce the purchasing and maintaining cost of the hard devices of computers and Internet transition for distance education.

2.5 Offer Economic Application Software Customization

Software as a service (SaaS) is a type of service provided by a distributed computing technology, which provides software as an online service, and Google Enterprise Application Suite (EAS) is this type of application. It no longer need to spend a lot of funds for the purchase of commercial Software authorization after remote terminal accessing to the education system which provides such services, such as office software, e-mail systems because all the above distributed services are provided in low-cost, some or even free. The local computer as client just to run the graphical interface operating system and Firefox browser to enjoy the distributed services, without worrying about whether the latest version, which also greatly reduce cost for school to maintain and upgrade the operating system, software. And Linux and Firefox are open source software, which we can use with free license [4].

2.6 Provide the Reliable and Safe Data Storage Center for the Distance Education

The investment to enlarge the education information resources becomes more and more large; each service organization accumulates the massive educational information. In viral and hacker's rampant Internet time, the most important is how to store data on the safe side. This pose a challenge to the school, and it is more serious in some schools which is short of professional. The distributed computing can provide safe and reliable data storage center, store the data in the distributed end, the computing service gives specialty, highly effective and the safe data storage. Therefore does worry that viral, the hacker's attack as well as the data missing question causes by hardware's damage.

2.7 Cause the Educational Information Resources' Build and Sharing to be More Convenient

At present all levels of educational administration organization, school and the education enterprise in our country have already constructed the massive educational information resources, and are constructing the more educational information resources. Store the education information resources on distributed platform, and then the sharing of education information resources became more convenient and quickly. Each educational institution or the information resources construct persons also is able to use the distributed computing moldable joint operation ability and realize educational information resources building.

In the SIMtone Corporation's education distributed computing project, the ordinary old computer visits the "WebSNAP" entrance via browser, and inters "the general distributed computing service" platform. While the new computer (can move Windows XP or above) enters the distributional service via terminal virtual machine software "SoftSNAP". This distributional service transmits for each teacher and student hypothesized personal computing tabletop. No matter in school or family, use desktop computers, notebook computers or mobile Phones, after entering the distributional end, everyone can gain to the same identical hypothesized tabletop, likely using the identical computer's system. Therefore you may continue your work in any place, share your data on many kinds of equipment.

What is more, the distributed computing service provides the sharing mechanism, which makes it possible to share the documents with other people. For example, use the Google EAS, you may bring the writing not finished to your home, complete the teaching design text and the demonstration documents together with other partners, share them with other teachers. The distributed computing causes the educational information resources' build and sharing to be more convenient and quickly [5].

3. THE MODEL AND PROCESS

The distance education resources which compose distributed distance education system are decentralized and decentralized. They are no longer limited to a single computer or small-scale local area network. The ultimate goal of distributed computing technology is to form a suppositional super computer using many computers on the net. Thus the architecture of distributed distance education resources system is a primary problem that we must settle.

The core of distributed distance education system is the sharing of teaching resources and the establishment of educational service system. At present, however, this kind of sharing is located on a low level. First of all, the styles of distributed educational resources are various and their formats are not unified. So we can't make a consolidate or effective coordination. In the next place, so many isomeric systems of hardware and OS exist at one time, which makes it impossible to run many different distributed education systems on different platforms. In addition, most of the styles of realization are static or dynamic web page frames, which do not show the concept of service. The distance education system based on distributed technology is a kind of virtual distance educational circumstance which is based on distributed net technology. Its leading aim is to make all users share a compositive setting which is provided by net. To learners, it is unnecessary to know who and how the educational service is provided. To developers, fast-speed development can be realized through the agreed interfaces [6].

The distance educational model based on distributed technology is basically composed by following parts. From Fig. 2 we know that distance educational model based on distributed technology is composed by following parts:

（1）Network Platform:

Distance education system based on distributed education system uses traditional network transport protocol and, it is based on present network educational transport platform rather than replace it. So the existing computer networks will become the base of distance education system based on distributed technology in the future.

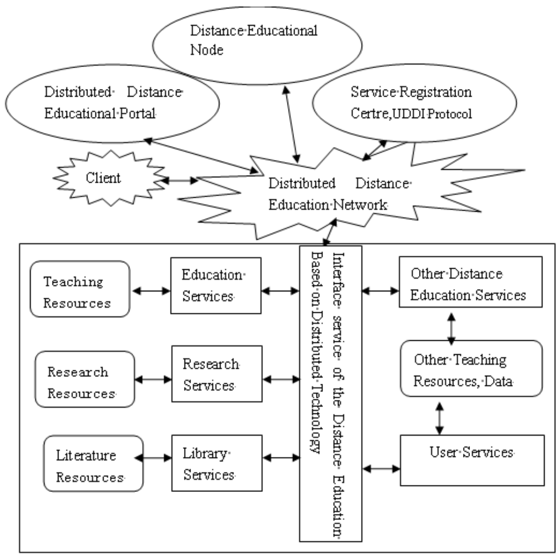


Figure 2. The distance educational model based on distributed technology.

（2）Distance Educational Distributed Technological Node:
The node is the most important part of distance education system based on distributed technology. The realization of distributed technology is radically relied on distance educational distributed technological nodes. Distance educational distributed technological nodes contain distributed educational resources and their realization. No matter how complex the realization of distributed educational service is, the distributed service provided by distance educational distributed technological nodes shields its complexity, making users use this distributed service pellucidly. The main object of distance educational service based on distributed technology is learners and developers of network education system. So its inner resources is mainly distance education resources.

（3）Service Registration Centre:
Its function is to provide large-scale query of distributed educational service. The way to share network educational service developed by schools and scientific researchers is relied on service registration centre. So does the way how users realize the existence of distance education system based on distributed technology. The process of registration is finished through UDDI (Universal Description, Discovery, and Integration).

（4）Distributed Distance Educational Portal:
The portals can be treated as web page frames of distributed service for the users. Apparently, distributed distance educational portal is no different from traditional web pages. But its connotation is far more different. It is very necessary to provide such distributed distance educational portals in distance education system based on distributed technology. It makes getting services from distributed distance educational portals more convenient. Most users and learners will enter distributed education system through distributed distance educational portals. So that they will get their needful services and learn.

（5）Client:
In distance education system based on distributed technology, client represents the object of service, learners, in another word. Client can not only use distance education resources and enjoy its service, at the same time, but also construct its own application system using distance education resources based on distributed technology. Distance education systems have familiar characteristics. So the core of construction of distance education system is the development of distance education service based on distributed technology.

Apparently, distance education system based on distributed technology is very familiar with traditional Client/Server structure. But its way to satisfy users is different. The former is to finish the requirements through calling distributed services. Distance education system using distributed technology can provide services for the legal users on the Internet. Fig. 3 shows its flow of work.

The client sends a request that calls the grid services of Turbo C compiler system to the server through the browser. Upon receiving the request, the server hands over the task to Servlet or Javabean.

Servlet sends SOAP messages and then calls the grid services of Turbo C compiler system.

On receiving messages, the grid services of Turbo C compiler system hand the request over to the executive programs of grid services. The executive programs carry corresponding operations, compile the C source code users submit and return the results to the Web server.

After receiving SOAP messages, the Web server sends messages to the corresponding Servlet.

At last, the Web server returns the messages, which are the running results after Web pages and the C source code are compiled, to the client so that the client could see program-running results.

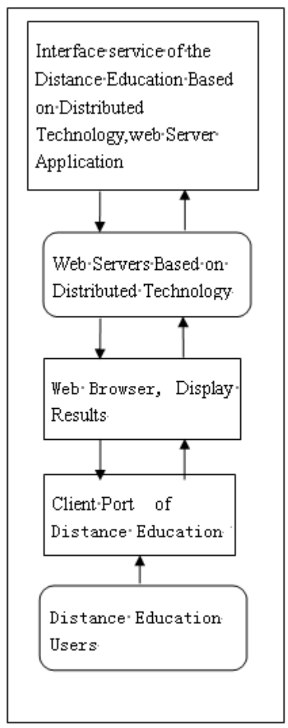


Figure 3. The system processes of distance education based on the distributed technology.

4. INTEGRATION AND APPLICATION OF RESOURCES

The main purpose of modern distance education teaching and distributed computing technology is to build a massive database of distance education resources and provide a mechanism for efficient resources sharing and memory management [7]. Distributed computing technology, which can bring in a great deal storage capacity and other IT resources, is preferred to be the first choice to replace the traditional technology. The integration and application of the distance education teaching resources are primarily seen in the following areas:

4.1 The Current Condition of distance education teaching resources

As the basis of teaching, teaching resources include infrastructure and teachers. The integration of teaching resources is conducive to a more rational and efficient use of teaching resources so as to promote the improvement of the teaching quality. The integration of resources carrying out among colleges and universities can be used to deal with deficiency of education input compared with the lasting increase of demand for teaching resources due to college expansion and to improve the utilization rate of high-quality teaching resources, so that more students have the access to good educational opportunities. Now, however, the domestic higher education in general exist a number of problems: uneven distribution of educational resources: well-known universities have better resources, including teaching, laboratory and computer equipment. However, most of the local institutions, as a result of limited sources of funds, often are faced with aging equipment, lack of laboratory resources, etc. The high cost and low speed of updating teaching resources: rapid progress in technology often requires schools to provide students with up-to-date laboratories and computer equipment which need more input and bring lots of pressure to colleges. Low level of teaching resources to share: In recent years the development of distance education and Internet education, as well as the expansion of colleges and universities in the scope of joint courses, made possible the integration of curriculum resources to some extent. Hardware resources, however, are still unable to achieve integration [8].

On the one hand, these problems resulted in duplication of resources into education, on the other hand, it expand the difference between the qualities of teaching. So how to integrate the teaching resources in the largest degree and scope has become the primary problem of reducing the cost of teaching, improving teaching quality, raising the level of high education.

4.2 The Platform of Distance Education for Connecting Heterogeneous Storage

In general most of distance teaching system includes real-time interactive video, voice, text, whiteboard, document sharing, etc. Teaching for real-time interactive can build a good communication platform for internal or among different organizations. The application of file upload and download technology occupied a larger proportion in the entire teaching and learning environment. So the client is often disturbed by bottleneck problem of the server-side storage.

Most of the traditional teaching resources networks have established different sizes and types of storage systems. The majority of storage systems are not compatible due to the fact that each of the storage matrixes uses different software and data access agreements. To access multiple storage systems, applications must use various methods to retrieve the required data, which is obviously inefficient. In specific applications, many of the geographical distributed applications or users need access to large amounts of data, which requires various storage systems to transfer large amounts of data quickly and efficiently.

Distance Education System with Distributed Grid File Transfer Protocol (Grid FTP), can be used to realize distance education system with large-scale service capabilities of distributed resources for education. Such distance education system does not need to store all the education resources for services in the server, but to combine different storage server resources by using Grid FTP, which can realize large-scale deployment of the platform of distance education.

The “University courses online” project put forward by Li Xiaoming who is the professor of Peking University is grid network infrastructure based on the CER Net. Its goal is to build a large-scale video-on-demand platform for the realization of the various colleges and universities nationwide education resources, in particular, the video sharing of education resources, by using distributed computing technology. At present, 15 well-known colleges and universities have provided nearly 3 000 hours of video of the best course, which have become the country's largest public university courses Video Network Service Platform. In this service system, video education resources are distributed in a number of CER Net grid nodes, rather than focus on a single server. Courses on-demand system can determine automatically which server can be used, according to the viewer's location and establish the corresponding connections during which the users do not need to consider this process.

4.3 Remote Immersion Technology Based on the Distributed Technology

Remote immersion (Tele-immersion) is a special kind of network-based virtual reality environment which is a faithful reflection of reality or history, and also can be a purely imaginary space which makes one immerse into it completely and create the sense of feeling. Participants around the network in the same virtual space, not only can roam, but also can communicate with each other, or even interact with the virtual environment so as to make it change. Practically, although there was no grid support, we can still undergo the development of remote immersion application. However, with the support of the grid, to achieve this type of application would be easier. In a sense, I believe that the emergence of the grid makes the distributed distance education technology truly operable.

In the field of distance education, remote immersion technology has the potential application prospects. NICE (Narrative Immersive Constructionist/ Collaborative Environments) - is produced jointly by Chicago EVL Electronic Visualization Laboratory and Interactive Computing Environments Laboratory of the University of Illinois. NICE is a collaborative learning environment with an explorative nature, specifically designed for children 6-10 years old. When they enter the CAVE virtual reality environment, wearing special glasses, children from different places greet, talk, and learn from each other in this virtual garden. They can observe, touch the flowers and trees of the garden and water them. Since it is a virtual environment, they can even touch the sun, or zoom in or out the plant at random. NICE is interesting and useful. The children can gain knowledge in the process of entertainment.

4.4 Construction of Distributed Distance Education Services System

The core of Distributed Construction of Distance Education is the establishment of the system of teaching resources sharing and teaching service. It gathers together the mass teaching information distributed in the field of education, such as, digital library resources, digital museums, papers, multimedia course ware and digital video of colleges so as to build a distributed information grid covering the various regions to provide a unified and efficient information service, active learning methods and interactive teaching methods. It enables learners to learn and communicate at home. At the same time, distributed distance education grid is also available to learners to carry out mimic experiments on the network.

4.5 The Integration of Resources for Distance Education Teaching Based on Distributed Technology

With the development of information technology information education has changed greatly. Thanks to the national education network and the high speed education network international interface, the distance learning organizations have more opportunities to exchange with others. The network hubs in each big educational institution develop, performance servers used in the teaching and the research help many educational institutions have the very strong information-handling capacity. Meanwhile the open source software, the opening standard, the opening data accessing and the opening scientific research idea have developed into the open education content, these course contain the curriculum data (courseware, reflection) and the interactive teaching community. The good network infrastructure together with the sharing of preliminary resources has built the foundation for the application of distributed computing in the distance education.

At present, the distributed computing, restricted by the size of the network and the resources, is mainly initiated by several big companies. This is mostly being applied to commerce, some application to education project is limited to the computing resources. If uses the existing distributed computing plans directly, no doubt the short-term cost is low, the implementation is simple. But how to transplant the original distance learning resources and dill with the original computing power of educational institution will become the main question. Therefore deploying the distributed computing technology in the existing education network and integrating distance learning teaching resources, will be one kind of feasible choice. This deployment way contains the following step approximately:

To associate or reform the existing distances learning teaching resources in a service-oriented way, this is a work many educational institutions are carrying on. And construct new centers providing the basic application service computation and the data. Such as distance teaching system, new hypothesized lab system, application procedure server and so on.

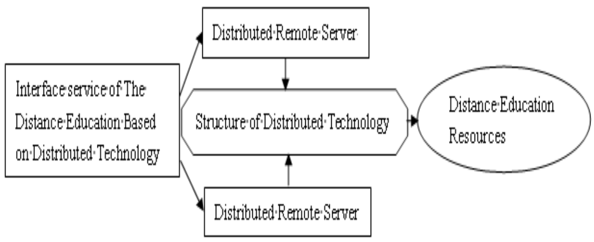


Figure 4. The platform in distance education based on distributed technology.

On the basis of fig. 4, construct the infrastructure, and use of virtual technology for the distribution. Simultaneously constructs the hypothesized distributional junction point (cloud portals), ensure the different distance learning organization be able to register the distributional technology the distance learning resources conformity platform quickly.

Distance teaching which takes the network as the carrier takes the resources environment as the main support, these foundation resources contain the text, the graph and image, the animation, the audio frequency storehouse, the video frequency storehouse, the discipline basic symbol and the discipline fundamental figure storehouse used in the network curriculum teaching. But in some part of the branch institution distance education digital resources are deficient, different branch institution sources redundant construction are fairly common problems.

In order to realize the integration of distance learning resources, propose education resources integration program based on the distributional technology. Firstly, integrate disperse education resources in distance education, seal network original education service logic, establish distance Web Service registration center, thus the distance education application system's development may integrate these Web Service, realize the distance education resources sharing. Based on this establishes distance Web Service registration center, realize the distance education resources sharing, finally realize the integration of distance education Web Service and develop distance education resources application system [10].

Based on the above solution, this article selects the integration of education resources in “the wireless access distance education system”. “The wireless access distance learning system” includes two kind of typical applications, take the wireless teaching, discusses alternately as the center wireless study (i.e. W-Leaming) and take the conformity campus information, the wireless mass organization as the center wireless campus (i.e. W-Campus). Through providing the short news, two WAP wireless access methods and the traditional Web open-door policy, meets needs which the user studies anytime and anywhere. The architecture of distance education resources system will integrate distance learning resources in an Object Oriented way. From the top downward respectively is the database level, the XML unification description level, Web the Service unification visit level, the registration level, the application logic level and the equipment visit level, the following carries on the explanation separately to each.

The database level contains the existing disperse education resources stored in different ways. Many education resources such as the library database, the educational administration resources database, the student achievement database, on-line choose education resources and courseware database, distributes in each place of school, and the memory way varies. For example, Bei Hang University's graduate student educational administration information store in the SQL Server database, while the library information and the undergraduate student educational administration information saves in the Oracle database, BBS based on the MySQL database.

The XML Universally describe courseware resources based on layer specification, provides the unitive description documents. Regarding the courseware database, use XML to realize the processing, the organization, the description, provide the uniform and standard courseware description documents.

Web Service level packages the original education application logic, the isomerous database and the courseware documents, provides Web Service run environment. This is not to replace the original education application system, but to help the system raise efficiency. It is very helpful to package the business logic modules and middlewares which have not been used. On the one hand it has avoided the redundant development; on the other hand the upper applications not need to visit primitive education resources. This greatly raised application system's development efficiency.

Web Service registration level provides the service issue mechanism, realizes the sharing and cooperation of education resources.

The system establishes service registration center in and out of school, issues corresponding Web Service description interface and the disposition attribute in this unified platform. This platform verifies the function of Web Service, and classifies Web Services. No matter which school systems can select suitable Web Services though this service registration center, integrate the service in their own systems, at last realize the sharing and cooperation of resources.

The application logic level integrates Web Service application, provides the personalization service equipment visit level to realize unification visit in different ways.

Thanks to the distributed computing technology, resources for distance education teaching combine with computers (including computers group) in different places, the database, each kind of equipment and the storage device. Then a high performance computation virtual environment come ture, which can achieve the high throughput computation, the coordination project, the data access and so on. Different Branch’s designers and developers upload resources developed by themselves to the uniform interface. Other Branch’s institutions access the interface, download the files, put them on every node. it can achieve the resources synchronization and distributing. Drawing on the sharing ability, it can construct Virtual Interface support access and integration [11]. At last users and client end can easily access the distributed resources, and integrate the resources for distance education teaching.

5. CONCLUSIONS

From the application of distributed technology’ angle, the distributed technology is only a very small part of applications, it provides useful experience for the use, management and dispatches the education resources. As the distributed technology progressing, the distance education system based on the distributed technology can serve the lifelong education better.

At present, many difficulties and problems remain in how to integrate resources for distance education teaching based on distributed technology.

How to formulate plans and measures for education resources? Many of them are repeated used. We should make full use of them, then arrange or organize them according to classes, or select a part of them and join them into the education service system?

Regarding the new distance learning application procedure server and the hypothesized platform, it needs the support of fund and to consider how to install the Virtual server and interface server.

How to make full use of the distributed technology in the distance education? As the favorable mean of resources for distance education teaching, the use of the distributed technology will become more and more widespread in the field of educational institutions and education networks, although this will takes years to test in practice.

REFERENCES

[1] Xu Huaxiang, Preliminary Study on the Influence of Cloud Computing on Education[J], Computer Knowledge and Technology, N0.10, 2009, pp.2690-2692.

[2] He Xueli, Cao Miaomiao, The Research and Design of the Distance Education System Based on CSCW Technology, JOURNAL OF ANHUI RADIO & TV UNIVERSITY, No.1, 2009, pp.56~59.

[3] Yang Yang, Gird Technique and its Application to Remote Teaching, COMPUTER DEVELOPMENT & APPLICATIONS, No.8, 2007, pp.57~59.

[4] Li Mingli, Grid Technology in Distance Education, Software Guide•Educational Techology, No.8, 2008, pp.38~39.

[5] Mo Zan, Zhao Shan, Li Li, Liu Aijun, Research on the Architecture of Distant Education System Base on Grid Technology, e-EDUCATION RESEARCH, No 9, 2006, pp.23~28.

[6] Zhang Qin, Study and Implementation of Modern Long-distance Education System Model Based on Grid Computing, Computer and Automation, No.8, 2005, pp.60~63.

[7] Cheng Jiejing, Peng Feizhang, Analysis on Information Teaching Pattern Based on Grid, No.9, 2006, pp.131~133.

[8] He Yufei, Wang Haitao, The Study of Net-based Distance Education Model, JOURNAL OF CHONGQING RTV UNIVERSITY, No.5, 2006, pp.14~15.

[9] Zhang Yuanyuan, Zhao Guangyu, Yu Huahong, The integration of Cloud Computing And teaching resources, Modern Science, No 4, 2009, pp.23.

[10] Hao Zengyuan, Yan Hongyin, Research on Remote Teaching Model based on Grid, Computer Development & Applications, 2008, pp.54~58.

[11] Ma Manfu, Duan Fuhai, Huang Zhiyi, Zhang Weiguo, Grid resource selecting based on cloud theory, Journal of Computer App lications. No.4, 2009, pp.1162~1164.

Cheng Jiejing (1967-), male, Han Nationality, Qianshan Country Anqing City Anhui Province, Doctor of Education, Post-Doctoral of College of Information Management of Wuhan University, associate professor of Nanchang University, Master's degree student teacher. The research field is the development and utilization of educational technology and educational information resources. He has published two books, over fifty papers, presided over seven projects supported by provinces and ministry.

Liu Xiaoxiao (1985-), female，Han Nationality, Hengshui City Hebei Province, post-graduate student in the department of technical education in Nanchang University.

THE IMPROVEMENT OF FACILITY LAYOUT OF THE AUTOMOBILE ACCESSORY FACTORY BASED ON OPTIMIZATION ALGORITHM *

Juan Li1, Jingguo Qu2, and Haibo Yan1 and Aimin Yang2
1. Academic Affairs Department, Hebei Polytechnic University
Tangshan, Hebei Province, 063009 China
2. College of Science, Hebei Polytechnic University
Tangshan, Hebei Province, 063009 China
Email: aimin_heut@163.com

ABSTRACT

The paper mainly introduces the process of improvement for operation facility layout of workshop in manufacture enterprises, and makes analysis of material flow through the original data of enterprises and optimizes the facility layout of workshop operation of the enterprise with the application of from-to chart method and SLP method. And the Uloop arithmetic is adopted to conduct comprehensive optimization for the equipments in workshop according to mixed layout and unit layout principle, to achieve the purpose of optimization for operation facility layout of enterprise.

Keywords: Facility layout, From-to chart, System layout design, Optimization.

1. INTRODUCTION

In contemporary market environment, customers' requirements are becoming more diverse and individualized, which makes the enterprises' production modes transmit from the mass production to small batch custom-tailored production which is more flexible. Under the condition that the production requirement is more stable, the enterprise designs the production system based on actual production ability and equipment condition, and lays out facilities in the mode which is most beneficial to production [1-2]. But when the market is intensely competitive and the requirements of society are changed frequently, enterprises must take corresponding measures to improve competitiveness, such as the addition of new production equipments and the change of production system layout, and the purpose is to adapt to the internal and external environment change of enterprises, thereby, reduce production costs and increase production efficiency.

The layout of production system is an important content of production management and material foundation of spatial organization in the process of production. Factories, as the material form in production system of enterprises, are becoming more and more organized and systematic with the complexity of the enterprises' production and management activities. The system layout of operation facilities should be conducted not only for the newly built factories, but also for the running factories, and the reasonable rearrangements for the operation facilities are also necessary to improve the system's production capacity and production efficiency of the original system.

Study shows that about the 80％ of the foreign manufacture systems running have not met the design requirements completely, and 60％ of the existing problems can be ascribed to the unreasonable initial planning or misplay, which leads to 20％-50％ of the total operating cost in the system manufacture wasted in materials transportation caused by the misplay of plane design. While, good facility layout may decrease such kind of cost to 10％-30％. Moreover, good facility layout can also accelerate material transportation efficiency and reduce the residence time of products in workshops, so the production cycle can be shortened. Thus it can be seen that operation facility layout to factory is one of the most important steps in the field of production management, and it is also one of the decisive factors to affect the production efficiency of enterprise. Therefore, activity facility layout has very important significance to the successful management of enterprises, including the factors of materials flow and personnel etc.

In the production practice of an enterprise, we adopt from-to chart and SLP method to conduct optimization layout for the workshop of a certain automobile accessory factory, and apply Uloop arithmetic method to conduct optimization layout for the internal equipments in the workshop. The optimization process is from whole to part and goes into depth step by step, and an ultimate optimization layout project is gained with the method of combination of theory and practice.

2. THE CONTENT, OBJECTIVES AND PRINCIPLES OF OPERATIONS FACILITIES LAYOUT

2.1 The Content of Operation Facilities Layout

Operations Facility Layout refers to reasonable location arrangements for a number of economic activity units in a given production scope. To make decisions of operations facilities layout, the following two main aspects should be considered:

(1) overall plane layout of site according to facilities location and site terrain have been selected, makes the most reasonable arrangements for all kinds of physical facilities from the angle of the whole site;

(2) plane layout of operation units On the basis of production operation tasks that operation units should assume, make certain of the mutual position and area of organization units including each workshop section, teams and groups and workplace etc. reasonably. It is divided further into two phases of workshop overall layout and equipments (workplace) layout.

2. 2 Objectives of Operation Facility Layout

Objectives of operation facility layout is to make arrangements of all the physical facilities of enterprise for the optimum " production operation mode" or "production and operation system" and to realize the optimum combination of

* Supported by the Natural Scientific Fund Project (70471063, 70771010, 70671088) and funded by 985 second engineering project (1070082004000024).

working equipments, materials and personnel, so as to meet production and operation requirements with the most economic means and improve enterprise efficiency. Include concretely [4]:

(1) The shortest distance of removal route and the minimum number of removal;
(2) Make effective use of space(area);
(3) Effective use of the workforce;
(4) Elimination of bottlenecks;
(5) Be conducive to information communication between employees themselves, employees and managers and between staff and customers;
(6) To shorten the time of manufacturing cycle and customer service;
(7) To reduce waste or unnecessary movements;
(8) To be favorable to the entrance, exit and putting position of materials, products and personne;
(9) Integration of safety equipments and safety facilities;
(10) to improve product and service quality;
(11) To facilitate the normal maintenance activities;
(12) Be conducive to visual monitoring for operations and activities;
(13) Increase flexibility to adapt to the change.

2.3 Principle of Operation Facility Layout

Because facilities layout is of great significance to efficiency of a system, in order to realize the objective of operation facilities layout, attention should be paid to the following guiding principles in the course of the layout being considered [5]:

(1) Complies with the requirements of the production process the layout of all kinds of equipments and workshop should be in compliance with process order of production operation course of products and services, to ensure the reasonable arrangement for production operations with the purpose to adapting to the production organization forms adopted.
(2) Adapt to the requirements of productions changes because of the diversity of customers' needs, the production and management activities of enterprise which meets the customers' requirements are certain to be a dynamic process, so the layout of production system is also inevitable to be in development and changes.
(3) Facilitate overall coordination and management on one hand, the amount of physical facilities equipped with should be adapted to product output and materials circulation volume, and a reasonable ratio should be maintained mutually; on the other hand, the location of each kind of physical facilities can not be determined in isolated way, and the convenience of collaboration between these physical facilities should be considered, which is beneficial to implementation of overall management and realization of optimum integration.
(4) Optimization of transportation and removal optimization of transportation refers to minimum costs of transportation and the highest benefits of transportation. The layout of materials removal routines must be tailored to transportation characteristic of goods, which means a minimum turnover of products and avoidance of the mutual intersect for removal routes, bypassing and to and from removal.
(5) Use of space with high efficiency the production system is three-dimensional settings, so the advantages of high-rise buildings and three-dimensional layout should be fully developed and plane layout should be combined with three-dimensional layout organically, simultaneously, compacts layout as far as possible, in conclusion, all of these are the efficient and proven methods to utilize space efficiently.
(6) Comfortable environment, safe and reliable good working environment can promote the physical and mental health of employees and improve work efficiency.
(7) reasonable equipments layout, and maintain flexibility to adapt to the requirements for development and changes of enterprise the capacity and load of equipments should be reasonable, equivalent and the same with amount of operators, so as to achieve high and stable utilization rate of production equipments and personnel.

3. THE EXISTING LAYOUT OF EQUIPMENTS IN FACTORY WORKSHOP AND THE EXISTING PROBLEMS

The original facility layout drawing of a certain automobile accessory factory is shown. The main unit optimized is workshop, and the functions of each of the regions shown in figure are shown.

(1) Temporary storage area, used mainly for temporary storage of raw materials before quality testing;
(2) major production area, main machines and equipments used for processing are all in this area, and the mode function layout is adopted to conduct layout.
(3) Toolings area.
(4) Equipment Maintenance and electricity supply area;
(5) General assembly area
(6) Spaying and painting area, three sets of spaying and painting equipments and cleaning equipment are in the area.
(7) Spare parts and assembly parts storage area, spare part is in G1 and parts are in G2 area.
(8) production services area, which are mainly some small-sized drilling machines and punches.

We can see from the figure above that the original layout of workshop belongs to functional layout, that is, to arrange according process principle, and all the punches equipments, milling equipments and welding equipments are all in separate regions. The type of layout, for the machinery manufacturing enterprises whose products are complex and have many kinds, is a kind very effective layout mode of operation facility, and is right suitable for the enterprise. Nevertheless, through in-depth analysis on production site, we still find the following problems exist in the workshop:

(1) The material transportation routine is in chaos, which increases waste of transportation time and costs, and a number of wastes exist in the course of joints process between each of working procedures, resulting in the low production efficiencies of each of working procedures.
(2) Long production cycle which results in comparatively larger amount of semi-products waiting for processing, causes the waste of production time and unstable products quality and the decrease of profit of enterprise.
(3) The wider distance between manufacturing equipments leading to comparatively wider moving distance for operators, so each operator can operate only one lathe every time, which is not conducive to improvement of operators' working efficiency
(4) Difficult dispatch because the production site are in chaos, this leads to the waste of personnel and difficulty of work dispatch.

The best approach to solve the problems above is to make arrangement for the facility layout of the enterprise again. Underside, we will adopt the from-to chart and SLP method to conduct optimization layout for the factory's workshop and apply Uloop arithmetic method to conduct the t optimization layout for the internal equipments in the workshop after its optimization layout, and thus we gain the final optimization layout scheme.

4 CONDUCT IMPROVEMENT OPTIMIZATION FOR OPERATION EQUIPMENTS IN WORKSHOP

4.1 Conduct Optimization for Workshop with the Adaptation of from-to Chart

We know that the species of parts the enterprise manufactures contemporarily are up to 111 by researching. We adopt ABC classification analysis method to make analysis of the 111 species of parts based on cost, production and sales volume as well as processing technology flow, and 59 kinds of parts belongs to A category product are used for research. Work out the from-to chart based on technology routine of the 59 kinds of products and removal types and amount of products and components and parts happened between each workshop, see table 1.

Table 1. From-to chart of original part.

	Punch	Weld	Section Material	Machine Processing	Spraying-intingg	Assembly	Test	Total
Punch		33	8	3	3	0	0	47
Welding			1	3	1	0	2	9
Section Material	18	7		2	1	0	0	28
Machine Processing	1	1	1		1	0	1	5
Spraying Painting	0	0	0	0		5	0	5
Assembly	0	0	0	0	0		0	0
Test	0	0	0	0	0	0		0
Total	21	41	10	8	6	5	3	

We need to improve the from-to chart of original parts. In the improved from-to chart, we should arrange the larger numerical value of the material flow types to be close to the diagonal square. as possible in order to determine the workshop where the numerical value is in and the workshop of list where the numerical value is changed, which is the regulation of improvement of the from-to chart process. According to that regulation, we got the table 2 of the improved from-to table 2. Moreover, the original layout project of accessory factory has been improved from "punch, welding, section material, machine processing, spraying and painting, assembly and test " to " section material, punch, welding, machine processing, spraying and painting, assembly and test".

Table 2. The improved from-to chart.

from / to	Punch	Welding	Section Material	Machine Processing	Spraying Painting	Assembly	Test	Total
Punch		18	7	2	1	0	0	29
Welding	8		33	3	3	0	0	47
Section Material	1	2		3	1	0	2	9
Machine Processing	1	1	1		1	0	1	5
Spraying Painting	0	0	0	0		5	0	5
Assembly	0	0	0	0	0		0	0
Test	0	0	0	0	0	0		0
Total	10	21	41	8	6	5	3	94

Make comparison of the f1 with the f2 in the operation facility layout project according to the two from-to charts before and after they are improved. The compared objective function: under the fk of the given operation facility layout, the total distance D（fk）of the movements of all products and spares in workshop.

$D(f_k)=\sum_{j=1}^{n}\sum_{i=1}^{n}M(i(f_k),j(f_k))\cdot d(i(f_k),j(f_k))$, Thereinto, n is the total number of workshops of the factory; on condition that the *fk* is given, which is operation facility layout, $M(i(f_k),j(f_k))$ is the material flow variety and number which are form workshop i to workshop j, $d(i(f_k),j(f_k))$ is the distance from workshop i to workshop j .

We can conclude through calculation: the total distance of products movement before improvement is 154, but that after improvement is 130. By contrast, we can see that the decrease of movement distance is 154 subtracts from 130, which is equal to 24 units, that is to say, on premise that the process procedure is optimized, the transportation distance of the 59 kinds of products according to the new layout project is shorten by 24 unit distance, compared with the original workshop layout process.

4.2 The Adoption of Systematic Layout Planning (SLP) Method for the Optimization of Workshop

When conducts factory facility layout with the adoption of SLP method, the material flow condition can be studied by the dividing the mutual relation grade between operation unit pairs, and on the basis of it, the material flow related chart is introduced to show the general material flow; considering a number of qualitative factors, the mutual relation grade of non-material flow between operation units are determined with the introduction of correlation table; the comprehensive correlation tables can be acquired with integration of both. And conduct workshop layout on the basis of the areas occupied by each workshop [6].

Material flow factors are the quantitative factors to be considered in layout, and they are the total volume or the total cost of transportation between workshop pairs, as for accessory factory, the adoption of transportation cost as the quantitative factors in material flow analysis is based on the prior consideration of transportation volume and the addition of cost factors, which will make a more comprehensive analysis.

We regard the total cost of transportation for unit distance between workshops as the material flow strength between workshops, and got the material flow strength table between operation units. See table 3.

Table 3. Material flow strength tables between operation units.

	Section Material	Punch	Welding	Machine Processing	Spraying And Painting	Assembly	Test	Storage
Section Material		761	315	42	92	0	0	180
Punch			1185	75	74	0	0	250
Welding				115	5	0	155	870
Machine Processing					2	0	30	130
Spraying And Painting						29	0	20
Assembly							0	120
Test								151
Storage								

The work we need to do followed is to conduct grade division for material flow intensity with the material flow intensity analysis technology, and represent the different intensity grade with the symbols of A, E, I, O, and U respectively to draw the material flow correlation figure. See figure .

(1) Non-material Flow Analysis Between Workshops

Non-material flow analysis between workshops is the qualitative factors need to be considered in layout, and it is determined by the actual condition of factory, because the different focuses considered by each of the enterprises in actual production, the restriction of qualitative conditions are different.

According to recommendations of Richard Muther in the SLP, the number of key factors considered is from 8 to 10 for each item, and combining the actual condition of accessory factory, the following six influence factors of interrelation acute degree of workshops are determined: material flow; technological process; the use of the same equipment; the frequency degree of working interrelation; easy supervision and management; the influence of noise, vibration, dust, and the inflammable, explosive and dangerous goods [7].

After the influence factors Of interrelation acute degree of workshops are determined, then we can give the acute degree grade of each workshops, and SLP divides them into six grades of A, E, I, O, U and X. On the basis of the six influence factors chosen, and combining the actual condition of accessory factory, we can determine interrelation acute degree. We can establish non-material flow interrelation chart based on interrelation acute degree of each workshop, and then get non-material flow interrelation chart and figure, as shown in Fig. 1.

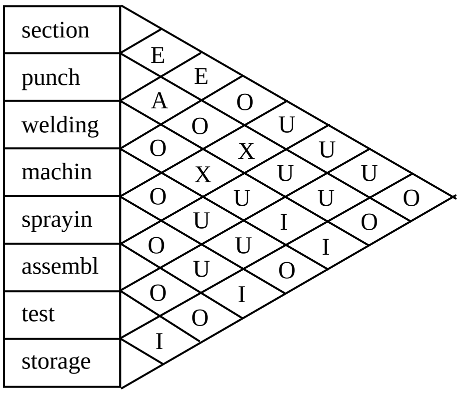


Figure1. non-material flow interrelation figure.

(2) Comprehensive Interrelation Figure

The combination of material flow interrelation figure with and non- flow interrelation figure. See Fig. 2

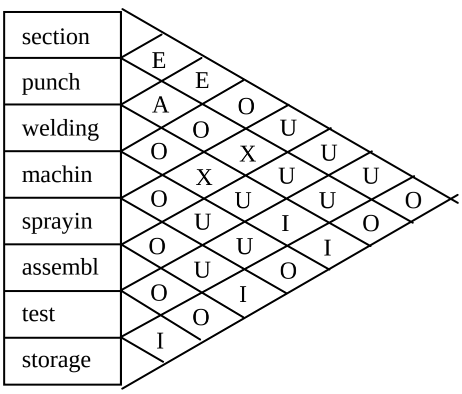


Figure2. Comprehensive density degree interrelation figure.

4.3 The Application of Principle of Mixed Layout and Unit Layout for the Comprehensive Optimization of Equipments in Workshops

We take the facility layout of punch workshop for example to conduct optimization calculation with the use of Uloop arithmetic, and with the same method for optimization, so we need not give unnecessary details

We can get the material flow relation matrix of each machine and equipments of punch workshop according to products parts technological process chart and output of products [8]

Table 4. Sparse Matrix chart.

	＃1	＃2	＃3	＃4	＃5	＃6	＃7
＃1		0	0	0	0	0	0
＃2	0		0	0	0	0	0
＃3	5	0		73	0	0	0
＃4	10	9	0		0	0	0
＃5	50	0	45	60		0	0
＃6	0	0	0	0	0		0
＃7	0	0	52	104	0	52	

Change the material flow matrix into Sparse Matrix ,see the table 4 for the result. Calculate the Sparse Matrix W’for ＃1 to ＃7, and then you can get Ri the sum of line and Ci the sum of row, and the R1, R2, R6, C5 and C7 are equal to zero, so we can dispose equipments of ＃1, ＃2 and ＃6 at the end place and ＃5 and ＃7 at the first place. The relative position between equipments of ＃1, ＃2 and ＃6, as the same as that between equipments of ＃5 and ＃7, should be determined first at random , and then, the sequence of layout is a=（＃1、＃2、＃6…＃5、＃7）;

Delete the first, second, fifth, sixth and seventh lines and rows from the Sparse Matrix W’, and then we can get a new material flow Matrix W′′, as shown in the table 5.

Table 5. material flow Matrix W′′.

	＃3	＃4	Ri
＃3	0	73	73
＃4	0	0	0
Ci	0	73	

According to I is equal to 3 and 4 in the material flow Matrix W′′, calculate Ri and Ci, and then we can get R3 and C4 are equal to zero, that is, the material flow volume from equipment 3 to equipment 4 is 73, and there is no material flow volume from equipment 4 to equipmet 3, so dispose the equipment 3 before the equipment 4, and then all the equipments have been laid out. Equipments layout sequence is: a=（＃1、＃2，＃6、＃3、＃4、＃5、＃7）; The corresponding target function value is: 15＋25+9＋5＋50＋45＋60＋52＋104＋52＝517.

According to Uloop optimization arithmetic [9], the followed thing is to optimize the sequence by exchanging the position of both the neighboring equipments respectively. That is, to find the neighboring equipments whose material flow volume Wij<Wji and exchange their position, and here W45<W54, so exchange the position of equipment4 with that of equipment 5, and then get the target function value is ：15＋25+9＋5＋50＋45＋52＋104＋52＝457; now, because of the improvement of target function value, update the layout sequence: a=（＃1、＃2，＃6、＃3、＃5、＃4、＃7）; update the sequence gradually

in the same way, and then the ultimate sequence: a=（＃1、＃2、＃7、＃6、＃5、＃3、＃4）; its target function value is 204, and other operations can not change the target function value. The ultimate optimization layout sequence is: a=（＃1、＃2、＃7、＃6、＃5、＃3、＃4）; thereinto, the position of ＃1, ＃2 and ＃7 can be exchanged mutually, so does ＃6 and ＃5, moreover, such kind of exchanges has no effect on target function value [10].

4.4 The Ultimate Layout Scheme

Combine SLP the workshop position interrelation figure from SLP with the equipments sequence and layout principle of punch workshop and welding and assembling workshop etc. from the Uloop optimization arithmetic, add the restriction of actual conditions and adopt the normative engineering symbol, we draw the ultimate layout scheme to be workshop plan layout diagram, as shown.

Change the facility position with the intention of the minimum overall material flow intensity in the process of production and process of manufacturing enterprises, and with the principle of considering the facility layout, to achieve the minimum overall material flow intensity of facility layout. By comparing the data of overall material flow intensity of facility layout before and after the improvement, we can see the sharp decrease of material flow intensity, thus, the purpose of optimizing equipment layout is achieved.

5. CONCLUSION

Through analysis the mostly factors of effecting the parallel efficiency, we give a parallel section pursuant method to solve the trinal-angles linear system equations based on the divide and rule thought. The parallel section pursuant method put forward in this text has the traits of little communication and high level parallel degree. From the theoretical analysis and the experiment, we can say that it is fit to compute under the distributed cluster environment.

REFERENCES

[1] Dennis P. Hobbs LEAN Manufacturing Implementation: A Complete Execution Manual for Any Size Manufacturer J. Ross Publish-ing 2004.
[2] Gary LeBlanc Kaizen at Hill-Rom CENTER FOR QUALITY OFMANAGEMENT JOURNAL Volumn8, Number 2 Autumn 1999page50·
[3] P.Courtois The M/G/1 finite capacity queue with delays. IEEE. Trans. Commun., 1980, COM-28: 165-171
[4] S.Fuhrmann A note on the M/G/1 queue with server vacations. Opns Res., 1984, (32): 1368-1373
[5] S.Fuhrmann and R.Cooper Stochastic decompositions in the M/G/1 queue with generalized vacations. Opns. Res., 1985, (33): 1117-1129
[6] C.Harris and W.Marchal. State dependence in M/G/1 server vacation models. Opns. Res., 1988, (36): 560-565
[7] J.Keilson, and R.Ramaswamy. The backlog and depletion-time process for M/G/1 vacation models with exhaustive service discipline. J. of Appl. Prob., 1988, (25): 404-412
[8] Y.Levy. and L.Kleinrock. A queue with starter and a queue with vacations: Delay analysis by decomposition. Opns. Res.,1986, (34): 426-436
[9] J.Teghem. Control of the service process in a queueing system. European J. of Opns. Res., 1986, (23): 141-158
[10] D.Gross. and C.M.Harris. Fundamentals of Queueing Theory. 2nd ed., JOHN WILEY&SONS, 1985

Li Juan: Lecturer, now acting as the teacher of Hebei Polytechnic University. She graduated from Yanshan University. She has already attended presided over many research projects. Her over 10 papers have been published. Her research direction is optimization algorithm and its application.

MODEL-VIEW-CONTROLLER FRAMEWORK IN PEER-TO-PEER APPLICATION DEVELOPMENT

Cheng WeiJie [1,2], Luo YunChuan [3,4], Cheng YanPing [4], Jiao YanJie [4], Zhang JinLiang [4]
[1] Wuhan University of Technology
[2] Beijing Highland Information Techonogy Co.Ltd
[3]Beijing University of Technology
[4] National Cultural Information Resources Sharing Project

ABSTRACT

The application of Model-View-Controller architecture in the development of Peer-to-Peer software is researched in this article. Basing on that a MVC framework applied to the development of the P2P software is proposed as well as a kind of thought about the usage of MVC in the field of P2P.

Keywords: MVC framework; P2P application

1. THE MVC ARCHITECTURE INTRODUCTION

MVC was just a kind of software designing pattern applied to Smalltalk-80 when it was proposed by Xerox PARC in 1980s. After decades of developing, it has been used in a wide field now. In this pattern, an application is divided into three parts: Model, View and controller. In this way the model and view is separated in codes, so that one application can have different expressions.

As the reflection of the application logic where the data is processed (handled) actually, model takes the charge of proceeding and organizing data of varies types, it encapsulates kinds of data and operates on it. The process tasks of model are more than that of the other components of MVC. MVC inform the view that which work flow has been updated beyond handle it and feedback the state of the data search. The feedback data must be neutral or free from format, so that one model can provide multiple results to several views, At the same time, the program code applied to model just need to be written one times and can be used repeatedly by multiple views, which reduced the repetitiveness of the codes.

View is the interface for the user, which connects the application program and the user. View transports the data imputed by the users to the Controller and present the feedback result to the user, meanwhile receives updated request and display the working state of the model. It is becoming more and more challenging that how to manipulate the interface of application program, one of the advantages of MVC architecture is that it can provided multiple views to your application program. In fact, the view just a pattern displaying data and permitting the user to operate on it no matter that these data are on-line stored or just an employee list, there is no actual manipulation in the view.

Controller is an assemblage between model and view including various manipulations; it takes charge of the interactive work between the view and the model, controlling the responding pattern and flow of the input of the user. There are two principle responsibilities of the controller; it distributes the user request to the relative models and reflects the changes of the model on the views promptly. The significant advantage is that by using controller connecting different models and views which are given to use repetitively to fulfill the various user requests. Controller can chose the appropriate model to manipulate the data basing on the user request, and then chose the view to display the result to the user. MVC architecture is designed to satisfy the need of the application that the same data should be displayed in different ways. The repetitiveness of the code is reduced greatly because the code applied to the model just need to be written one time and can be used by multiple views repeatedly. The efficient cooperation of the three components of MVC promotes the extendedness and maintenance of the applications significantly because it reduces the couple of them as much as possible.

Why MVC should be used? One of the important reasons is that multiple views can share one model. Now more and more visiting patterns are needed access to the application program, and MVC is a solution to the problem. No matter what interfaces are wanted by the users, one model can satisfy them. The code can be repeatedly use as much as possible because the data and business rules are separated at the representing layer. The feedback result from the model can be used by different interfaces because the data has not been formatted.

Model can change the rules of the data layer and the business layer of the application program easily because it is embodied and separated from the controller and view. Because the three components of the MVC applied to the program are independent from each other, the change of one of them will not affect the other two, the design idea are benefit to the construction of the loosely coupled units with best structures.

MVC design pattern is a best way to develop software, some of the principles it proposed are easily to be understood for example the rules of separating content and representing. However, how to split the structural components of model, view and controller need to be considered in depth, especially the framework of the application program. The soft ware will be robust if the MVC architecture is adopted the extra work and complexity brought by it can be dealed with, the repetitiveness of the cold and the structure of the program will be promoted significantly as well.

2. THE APPLICATION OF P2P AND THE CHARACTER OF THE SOFTWARE DEVELOPMENT

P2P (peer to peer) is neither a brand new idea nor innovative concept from the developing history of network. It was emerged with the original network communication. Just as the local area network connecting PCs in a building and the long distance network via Modem connecting different buildings, the communication pattern basing on TCP/IP protocol became the foundation of the internet of nowadays. From this aspect, P2P is not innovative technology but a kind of new technological application pattern.

Many network services can be included into P2P, such as immediate communication system ICQ, AOL, Instant Messenger, Yahoo Pager, MSN Messenger of Microsoft and

QQ of China, which are all the popular applications of P2P.
At present, the development pattern basing on overlay network construction is the main way to develop the application software of P2P. Among them, the overlay network just provide framework support including bottom network information transfer, router and locator, however, there are no enough consideration for the common problems of software engineering emerged during the development of P2P application which are focus in the logical business of upper middle layer. Just because of that, the same data cannot be represented in different ways, the development of P2P applications are loaded down with trivial details. Especial in the development of P2P applications with similar functions, because the programmers cannot use the developed resources repeatedly, they must redesign the program and rewrite the codes from the beginning, which reduces the working efficiency in a great degree. At the same time, because the high coupling of the different program modules, it is difficult to modify the program.

In nowadays, the P2P is so popular that it is necessary to construct a mobile and efficient development frame work appropriating to P2P application, which can manage the work flow of he development of the P2P applications software and make the programmer concentrate on the business logical of the application. That means, the whole structure of the program is provided by the developing framework which reduce the work load of every programmer respectively and strengthen the cooperation among them as well as promote the developing efficiency of the program.

3. THE INTRODUCE OF MVC ARCHITECTURE TO P2P APPLICATION SOFTWARE

Basing on the idea of MVC architecture, a framework of MVC applied to the development of P2P application is proposed. The new framework inherits the technological thought of MVC, which is the application of MVC pattern in the field of P2P. From the view of software engineering, MVC assists the designing and developing of P2P application software. That the layer thought is used in the developing procedure of P2P application software promote the extendedness and maintenance of the program.

The advantages of introducing MVC architecture to the development of P2P software

a) The different layers of the software have their own responsibilities respectively, they do not disturb each other and there is no influence to the others if any one of them is updated.

b) The MVC architecture is useful to fulfill the division of the responsibilities, so that the programmer can concentrate on the designing and developing on a specialized level which leads to the improvement of working efficiency and quality.

c) The potential of the units can be developed. Because of the specific division the units of different levels can take important responsibilities independently.

The introduction of MVC to the development of P2P application software will be discussed in details from the three components of MVC architecture.

(1)Model

Model is the core of the developing framework; usually it is compiled by JavaBean. Model receive the data request from the view, after manipulate it properly, model return the feedback to view. This course is the key of the whole framework. Taking the advantage of JavaBean, model realizes the repeated usage and extendedness of the code, and makes the maintenance of software be more convenient.

The business local units distributed on ever Peers compose principal part of the model of the new framework, these units encapsulate the visit to the data arranged at the bottom layer of the platform as well as procedure of realizing the business logical way. When the P2P application software is operating, multiple business logical units distributed on several Peers consist of the model, and these units have the common characters of satisfying some requests which are described by the view. Further more, Controller manages and cooperates these units distributed on different Peers.

During the development of the P2P application software depending on the new framework, design and developing technology basing on units will be adopted, both simple units and complex ones of high coupling are all included.

(2) View

View represents the interface to the users, which can be fulfilled via html, xml, applet of java ect.. It collects the data and manipulates it; the so called manipulation here refers to display mainly. In the new framework, view is the final result displayed before the users when the P2P application software is operating under the condition that a certain request of the users has been satisfied. The certain request of the program user is initiated by the business logical unit located on some peer at first, which is spread around the network, every peer suitable to the request character will attend the overlay network which are called collaborating group.

Request character is the core of the view of the new framework, which defines the *permitting rules* of the Peer where the P2P application software belonging to joining the application collaborating group. It accords to language standard and description of XML, and manipulated by the controller. In fact, request characters are element labels and attribute labels encapsulating the content of the users' request.

(3) Controller

Controller receive request from the user and transport it to the model instructing the model to manipulate the request with a certain module. It coordinates the collaboration between the view and the model, which is the pivot of the framework. Controller is usually compiled via Servlet. In the new framework, controller receives application requests from the users of P2P and handles them, provides feedback to the users and maintenance the overlay network the P2P belonging to.

It is fulfilled by the cooperation of three unit modules: request manipulation module, view re-construction module and overlay network module. In order to reach the target of repetitiveness of the code during the development, these 3 unit modules are designed and developed basing on the unit technology. Meanwhile, the middle technology is adopted in the integrated designing of the controller so that P2P software in different types can communicate with each other.

MVC framework used in the development of the P2P application software is shown in figure 1.

4. OPERATING FLOW OF THE P2P APPLICATION SOFTWARE UNDER THE FRAMEWORK OF MVC

After introducing the MVC architecture, the working flow of P2P application software is illustrated in figure2,which is as following.

1) A specific request needed by the users is initiated by the

business logical located on some peer when the P2P application software is operating.
2) The request of the user is transformed into the request character by the controller; the matching of which and the unit metal-data describing the business logic in the controller of this Peer is computed. If it is successful, the controller of this peer will be interconnected to itself, and anticipated result will be reported to the user via the view reconstruction module of the controller; On the contrary, if the matching is fail, the overlay network module of the controller of this peer will start, the information will be spread to the equal network the P2P application software belonging to. The overlay network module of the controller in every Peers of the equal network will distinguish and match them until the peer including the unit metal-data according to the request character is located.
3) This Peer will join the collaboration group which is defined by the request character, and then the work of the group will start. That means the controller of this peer will cooperate with initiative peer via the controller itself and screen the details of the cooperating at the bottom of the initiative peer.
4) The view reconstruction module of the controller of the initiative peer returns the result to the user

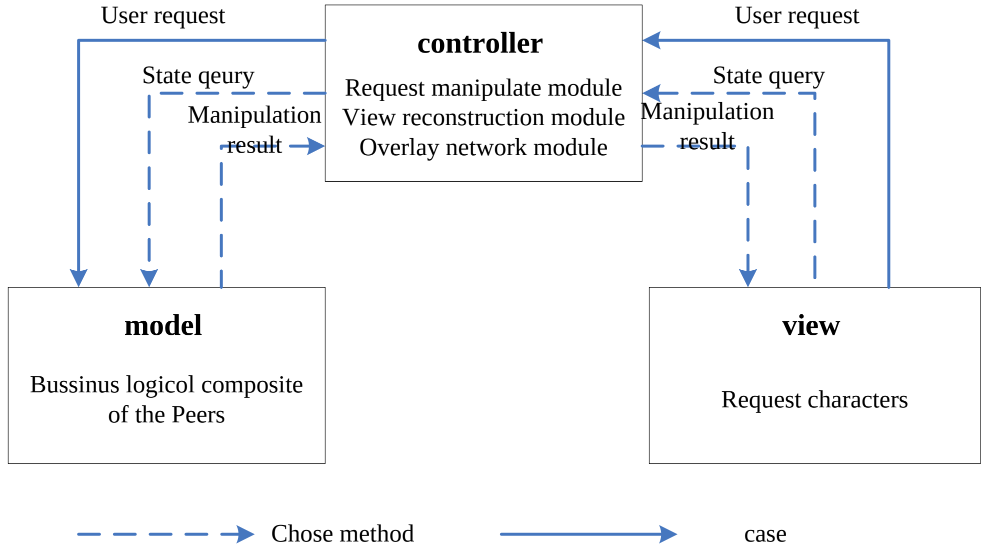


Figure 1 New framework of MVC in the development of P2P

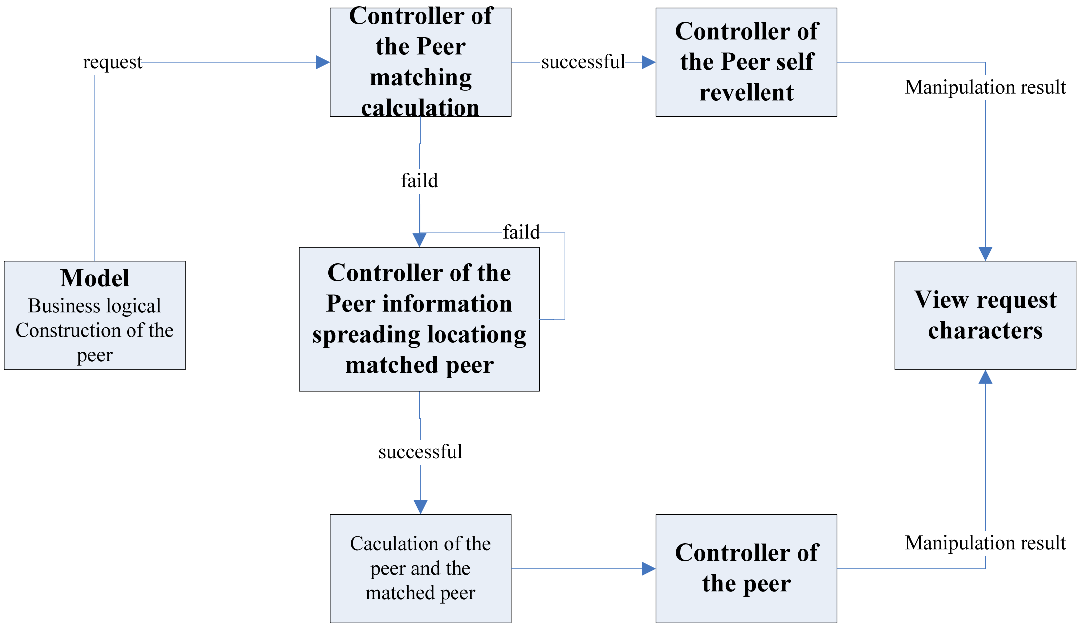


Figure 2 the MVC working flow of P2P application software

5. CONCLUSIONS

The MVC architecture is introduced at the beginning in this article as well as the P2P application and the problems in the development of the software. A new development framework introducing MVC pattern is proposed through combing MVC architecture with the development of P2P application software. According to the layer idea, the framework fulfills the designing and developing the P2P application software in layers, and reduces the high coupling of the different layers of the software. Every layers take their own responsibilities respectively and the units belonging to same layers have the common characters, which are useful to the management of the program codes in the way of engineering and programming.

REFERENCES

[1] BUSCHMANN F, MEUNIER R, ROHNERT H, et al. Pattern-Oriented Software Architecture, Volume 1, A System of Patterns[M].[S.l.]:John Wiley & Sons, 2000.
[2] BEHNEL S, BUCHMANN A. Models and Languages for Overlay Networks[J]. Lecture Notes in Computer Science, 2007, 4125(2): 211-218.
[3]BEHNEL S. SLOSL – A Modeling Language for Topologies and Routing in Overlay Networks: proceedings of the 15th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, Naples, 2007[C]. New York: IEEE, 2007.
[4]BEHNEL S, BUCHMANN A P. Overlay Networks – Implementation by Specification: proceedings of the 6th International Middleware Conference, Grenoble, 2006[C]. Germany: Springer, 2005.

FAULT DIAGNOSIS RESEARCH ON THE MONITOR SYSTEM OF FUEL CELL STACK *

Xiang Gao, Jian Deng, Wuxing Wei , ShuHai Quan
College of Automation, Wuhan University of Technology
Wuhan, 430070,China
Email: gx82119@yahoo.com.cn

ABSTRACT

According to the actual questions exiting in the monitor system of fuel cell stack, This article carries on fault diagnosis for the monitor system of fuel cell stack. The diagnosis method is founding fault diagnosis dictionary storehouse to carry on the fault diagnosis of the monitor system of fuel cell stack from top to bottom through the Fault Tree model. The key point is focused on the system signal processing unit with the BP neural network method. Signal processing unit can be effectively monitored and diagnosed through the creation and training of BP neural network. This article can achieve effectively real-time diagnosis and improve the stability and reliability to the monitor system of fuel cell stack through a combination of these two kinds of diagnostic methods.

Keywords: Voltage monitor, Fault Tree, Fault diagnosis, BP neural network.

1. INTRODUCTION

High-power fuel cell stack is usually composed of single cells in series. Because of the structure of cell in the process of fuel cell stack running, single cell unusual would influent performance and security of the whole stack. It is necessary to in real time monitor every single cell voltage and in time convenient for control system to make right decisions. So the normal performance of the monitor system is essential to monitor fuel cell stack and guarantee stack running in normal state. In order to carry out real-time monitoring of the stack, it is indispensable to ensure the normal operation of the monitor system. When the monitor system is in question, the fault sources must be able to be immediately detected. Therefore, Fault Diagnosis Research on the Monitor System of Fuel Cell Stack is particularly important.

2. MONITOR SYSTEM INTRODUCTION AND DIAGNOSTIC STRATEGIES

Fuel cell stack is generally composed of a few hundred cells in series, Although the single cell voltage is not high (0.5~1.2V), hundreds of single cells are connected in series to make cumulative electrical potential to reach hundreds of electric voltage. Routine testing chip input voltage is generally within tens of volts and difficult to sustain hundreds of fuel cell voltage. In order to solve the contradiction between single cell voltage relative to the earth at a high cumulative electric potential to chip voltage value, The system uses the distributional examination, so that all chips of the detection unit can meet the pressure requirements. In order to overcome the problem high common-mode voltage lies in the Monitor System of Fuel Cell Stack, The single cell voltage uses the grouping measuring method, every 30 cells of stack are divided into a group, Each detecting unit examines 30 road voltages. Every detection unit uses isolated DC-DC power supply, each other is not altogether with the same ground. Detecting unit is composed of several parts of signal processing units, Analog channel select switch with multi-channels, A/D converter circuit with external reference voltage, PIC18F258, man-machine interface, CAN interface circuit and so on. Detection unit structure is as shown in Fig. 1.

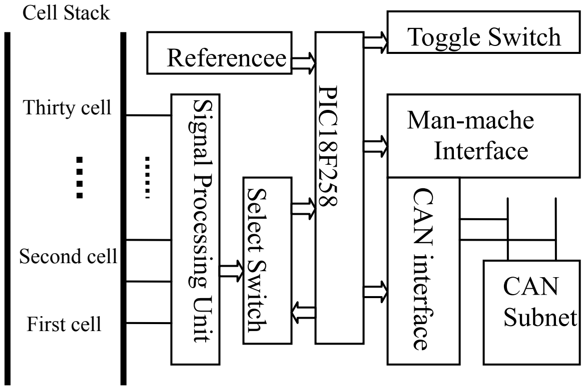


Figure 1. Detection unit structure.

Fault Tree model is a model based on the measured object. It is as the top case with the fault being most not wanted to occur. Other cases maybe cause the top case fault are as the middle and bottle cases, and express the connection between the cases with a kind of reciprocal form structure. Through calculating the probability that the top case of Fault Tree happens, probability of happening and important degree of the minimum cutset of Fault Tree, This artlcle provides quantitative analysis, thus reduces the difficulty of fault diagnosis greatly, makes the hit rate of once while searching in the trouble source greatly improve at the same time. Because Fault Three is connected by the logical relations of "and" and "or" which form all bottom cases, therefore structural function can be used as the mathematics tool, This article sets up the mathematical expression of Fault Tree, in order to calculate the trouble quantitatively as well as qualitative analysis. This article assumes that the study device and its components, etc have only two types of normal and fault status and assumes that the components, component failures are independent of each other. This article Studies a Fault Tree formed by n independent bottom cases, and sets x_i as state variables of the bottom case I, x_i only gets the two state 0 and 1. Φ expresses the state variable of a top case. Φ also only get the two states of 0 and 1. There are the following definitions:

$$x_i = \begin{cases} 1 \ \textit{Bottom case i happens} \\ 0 \ \textit{Bottom case i does not happen} \end{cases} \qquad (i=1,\ 2,\ \ldots,\ n)$$

$$\varphi = \begin{cases} 1 \ \textit{Top case happens} \\ 0 \ \textit{Top case does not happen} \end{cases}$$

The top case of Fault Three is the fault state not wanted to happen, Φ=1; The corresponding bottom case state is a trouble

* This work is funded by the Hi-tech Research and Development Program of China under contract 2006AA11A133

state of the components and parts, X_1=1. case state Φ is totally determined by the bottom case state of Fault Tree X, as $\Phi = \Phi(X)$, with X=(X_i , X_2 , X_3 …, X_n), Call $\Phi(X)$ the structural function of Fault Three. Structural function is a kind of Boolean function which expresses the equipment state. Its variable makes up the state of the unit for this device. Different Fault threes have different logic structures, thus correspond to different structural functions. Detect unit circuit board fault tree structure is shown in Figure 2:

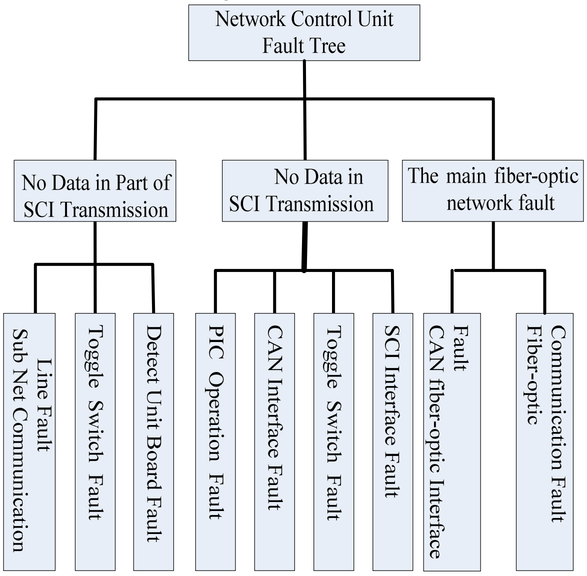


Figure 2. Detect Unit Circuit Board Fault Tree Structure.

3. SIGNAL PROCESSING UNIT FAULT DIAGNOSIS

A. Signal processing unit

Fuel cell singal cell voltage signal is a differential-mode small-signal, and contain mass common-mode parts, its value of numbers sometimes is much bigger than the differential-mode signal. Therefore, operational amplifiers of signal processing unit should have a strong ability of inhibiting common-mode signals. 3 common operational amplifiers make up Precision operational differential amplifier circuit, as shown in Figure 3:

Where $R_1 = R_3$, $R_2 = R_4$.

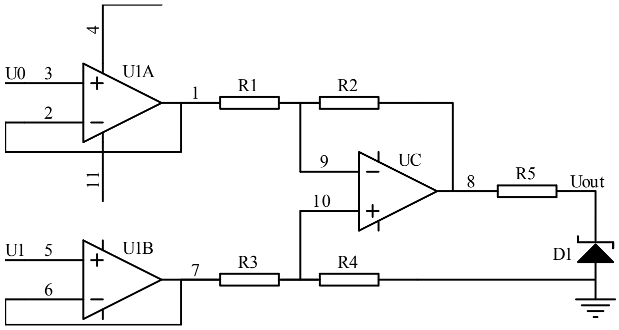


Figure 3. differential amplifier circuit.

$$U_{out} = \frac{R_2}{R_1}(U_1 - U_0) \tag{1}$$

B. BP net Introduction

Back-Propagation Network is a multilayer forward Network training Weights of Non-linear differentiable function. Fig 4 is its structure.

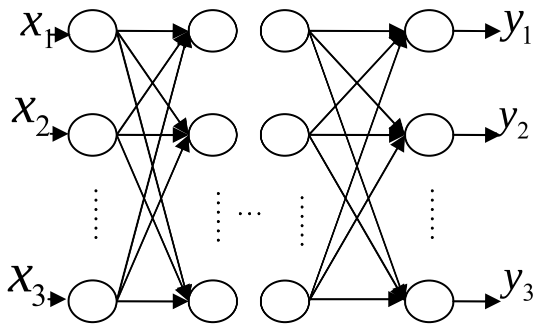


Figure 4. BP Net structure.

As shown in Fig 4，BP network is a kind of neural network with 3 layers or more than 3 layers including input layer, hidden layer and output layer. The output of every neuron node according to Connecting weights is weighted as the output of the next layer，there is no feedback between the level and the level.

C. Step of BP nerve network

Fault diagnosis steps based on the BP neural can be summarized as follows：

(1) Initialization. every connecting weight W_{ij} , V_{jt} , threshold θ_j and r_t are were given the random value between interval (-1,1).

(2) a group of input and objective samples $P_k = (a_1^k, a_{21}^k \ldots, a_n^k)$ and $T_k = (s_1^k, s_2^k \ldots, s_p^k)$ chosen by random are given to the network.

(3) every unit input S_j of middle layers is calculated by the input sample $P_k = (a_1^k, a_{21}^k \ldots, a_n^k)$, the connecting weight W_{ij} and the threshold θ_j ,then every unit output b_j is calculated with S_j through the transfer function.

$$S_j = \sum_{i=1}^{n} w_{ij} a_i - \theta_j \quad \text{j=1,2...p} \tag{2}$$

$$b_j = f(S_j) \quad \text{j=1,2…p} \tag{3}$$

(4) every unit output L_t of output layer is calculated by the input of middle layers b_j ,the connecting weight V_{jt} and the threshold r_t, then every unit response of the output layer C_t is calculated through the transfer function.

$$L_t = \sum_{j=1}^{p} V_{jt} b_j - \gamma_t \quad \text{t=1,2…q} \tag{4}$$

$$C_t = f(L_t) \quad \text{t=1,2,…q} \tag{5}$$

D. BP network design

When using BP neural network for fault diagnosis, at the equivalent of setting up fault dictionary in BP neural network, so the ideal of Fault Diagnosis Dictionary is as the following: to extract the circuit feature of the circuit under various faults at first, then to form a dictionary with one-to-one correspondence of the characteristics and faults. In the actual diagnosis, as long as obtaining real-time characteristics of the circuit, the corresponding fault can be detected from the Fault Diagnosis Dictionary. The eigenvector of the circuit is a vector responding characteristics of the circuit status, would be obtained from the voltage of test nodes with mathematical treatment. Thinking about allowance of the component, if the eigenvector obtained from test values is m-dimensional eigenvector p, numbers of circuit fault status f are n, the classifier of BP neural network is to achieve mapping from the points of m-dimensional feature space and their fields to the points of n+1 dimensional space and their fields. according to Fig 3- signal processing unit, in the process of network testing, The numbers of network output nodes are the same as the numbers of tested faults. In the design of neural network fault models have been encoded. Network output that corresponds to 1 is considered the corresponding component K as faulty. the corresponding component K has no fault when network output corresponds to 0. R_1, R_2, R_3, R_4 and R_5 are regarded as measuring notes. The component open circuit is as faulty mode. If the operation amplifier is in the normal working state, first of all, the artice must obtain the objective and input samples of the network, and obtain 6 groups of state sample datas through monitoring the signal processing units, as shown in table 1:

Table 1. Sample datas of signal processing unit.

fault category	Sample input data					Binary
Nnormal	0.73	1.08	1.09	1.08	1.08	000
R_1 Fault	0.73	0.71	1.09	1.82	1.82	001
R_2 Fault	0	22.6	14.1	-2.9	-0.7	010
R_3 Fault	-0.71	-1.07	-1.06	-1.07	-1.07	011
R_4 Fault	-0.71	-1.05	-1.06	-10.27	-4.2	100
R_5 Fault	-0.71	-1.07	-1.07	-1.08	-0.13	101

The structure of this BP network can be designed like this: Input layer has 5 neural networks. According to Kolmogorov theorem, implicit layer has 13 neural networks. the transfer function of middle layer neuron is the S-tangent function. the transfer function of ouput layer neuron is the S-logarithmic function. M means the input sample vector of the network, N means the objective sample vector of the network. So M is the sample data. N=[000, 001, 010, 011, 100, 101].

E Creation and training of BP network

The training function of the network is+. Learning function defaults to learngdm. Performance function defaults to learngdm. Network can not be directly used after created. This article set total system error as 0.02, the network was trained 200 times. When the network is completely trained, Connecting weights between the neurons obtained from learning sample set will be stored in the Knowledge Basic. After 193 times training, The network error reaches the establishing minimum. As showned in Fig 5:

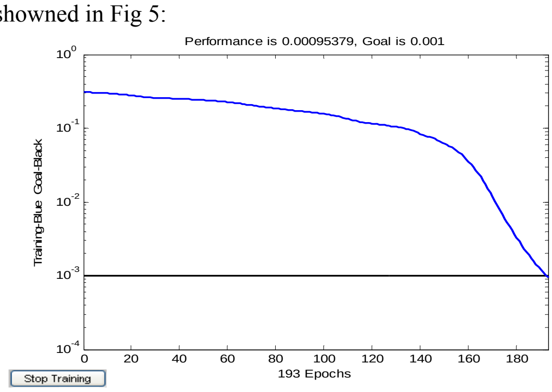


Figure 5. The diagram of Neural network simulation.

4. CONCLUSION

In this article, the faults of Monitor System of Fuel Cell Stack were diagnosed. This article proposes the diagnosis way of the whole system of the Fault Tree, expresses the connection between the incident with a kind of structure of reciprocal form. Meanwhile, and tests faults of signal processing unit connected with the stack with the BP neural network, so fully utilize circuit fault information of the test notes. This indicates that it is one kind of effectual methods through a large number of experiment tests.

REFERENCES

[1] S.Y.Yang, Shi-Bai Tong. Analog System Fault Diagnosis and Reliability Design.BeiJing: Tsinghua University Press,Feb 2001, pp,81-96

[2] Y.Q. Qiu,Y L,H Qiao Wu. Fault-Tree Diagnosis based on Optimization of Examples, the Proceedings of Ordnance Engineering College,Dec 2002,pp,60-75.

[3] F.M. Frank, "Fault Diagnosis in Dynamic System Using Analytecal and Know ledge -Based Redundancy -A Survey and Some New Results," Automatica, Vol, 26, No. 3 pp. 459-474, 1990.

[4] Youn W.C., Hammen J. M.Aiding the operator during novel fault diagnosis. IEEE Trans. On system, Man and cybernetics, Jan 1998, pp, 142-147.

[5] Kiran K.V., Joanne B. D. Automatic synthesis of fault trees for computer-based systems. IEEE Transactions on reliability, April 1999, pp, 394~401.

Jian Deng, (1964-), male, Associate Professor, Master's director. Automation department, Wuhan University of technology. Main research area: Intelligent examination and electric instrument, Electric automobile electron and control system. E-mail: djwhut@sina.com .

Xiang Gao, (1982-), male, undergraduate student, Automation department, Wuhan University of technology. Main research area: Control theory and Control engineering. E-mail: gx82119@yahoo.com.cn.

WuXing Wei, (1984-), male, undergraduate student, Automation department, Wuhan University of technology. Main research area: Control theory and Control engineering. E-mail:weiwuxing84@163.com

ShuHai Quan, (1953-), male, Professor, Vice-president. utomation department, Wuhan University of technology. Main research area: Intelligent Information Processing and Intelligent Control, Automotive electronics and electric vehicle control, Pattern Recognition and Image Processing.

NONLINEAR CONTROL ON THE SPEED OF A DUAL-FUEL ENGINE

YuhuiLi ,YanchaoGuo
School of Energy and Power Engineering, Wuhan University of Technology
Wuhan 430063, China
Email: krislee@whut.edu.cn

ABSTRACT

Dual-fuel engine as a kind of alternative fuel engine has both advantages in emission and disadvantages in multivariate coupling and nonlinear dynamic character. For the purpose of a Diesel/CNG dual-fuel engine speed stablization, a non-linear model has been established to simulating the speed regulation process in this paper. On the base of feedback linearization method, a Fuzzy-sliding mode controller has been designed to control the engine speed. By means of the xPC Target platform, a control system has been designed to perform the testing bench signal control. The result of the simulation proved that the controller can meet the demands of the engine in all situations. From the engine performance on the testing bench, the responding time of the speed control system is less than 4s and its transient regulation rate is 7%.

Keywords: Nonlinear Control ,Fuzzy-Sliding Mode Control ,Dual-Fuel Engine.

1. INTRODUCTION

The alternative fuel technology is a focus in the field of energy research. Diesel/ CNG engine is one of the most practical low emission engines at present. The performance of emission and power of this kind of engine is affected by the ratio of diesel and CNG. Bench testing is the only way to get the optimized ratio in the phase of calibration. In order to stabilize the engine speed during the diesel/ CNG ratio changes, a good performance of the engine governor is demanded.

The typical design mode of the governor that is used in mechanical governor is to liberalize the dynamic characters of engine, fuel pump and everything in the system. Ignoring the nonlinear factors in the system makes the design easy and the performance of speed stabilization pale[1]. This design mode is on the base of that the nonlinear dynamic characters of the system is not strong. However, in dual fuel engine, the controlled factors are more than one and the fuel ratio effect the performance of the engine as well as the fuel quantity. The nonlinear character of the system is so strong that it can not be ignored. That means a new method should be adopted to design the governor to improve the performance and meet the demand of the engine speed stabilization in bench test during the phase of calibration.

Recently, more and more research on the engine focus on the nonlinear character in especial on speed stabilization. Along with the development of nonlinear control technology, many method such as feedback linearization, Variable structure control(VSC), piecewise linear PID control, adapt control, Fuzzy logic control, artificial neural network control and so on, have been used to control engine[2]. Among these measures, the nonlinear feedback linearization is outstanding because of its unique property in mature theoretical system [3], so does VSC because of its robustness.
Variable structure control(VSC) is a synthesis method which belongs to the nonlinear control theory on the basis of phase-space. It usually used to designed nonlinear controller combined with feed-back linearization technique which is based on differentiable manifolds theory.VSC basis principle is that system state variables reach some switching surface firstly and then slide to the origin step by step on this surface. When the sliding movement has good quality, the purpose of control is accomplished[4]. The main characteristic of this kind of control is extremely strong robustness. That is, it is insensitive to the model error, system parameter variation and external disturbance [5]. It is very suitable for the engine governor design.

In this paper, a nonlinear model is designed to describe the performance of 2135 diesel/ CNG dual fuel engine. By means of feed back linearization method, the switch surface of the nonlinear system is constructed. Sliding mode controller is designed on the base of switch surface along with Fuzzy logic technique to avoid the switching tremble which is the typical phenomenon in the classical VSC controller. In order to inspect the controller quality, the dual fuel engine with the nonlinear controller is tested on test bench. The result of the experiments shows that the performance of the controller can meet the demand of speed stabilization.

2. DESIGN OF THE NONLINEAR ENGINE MODEL

The process of the engine speed regulation is usually seen as a quasi-static process. The change of engine speed is determined by the difference between the output torque of engine M and the load torque M_p. The engine speed acceleration is defined by $J_s \times \frac{d\omega}{dt} = M - M_p$ Here, J_s is the total equivalent inertia of all the rotated parts in the system including engine and load mechanism.

The output torque of engine M is determined by the quantity of diesel q, the amount of CNG c and the speed of engine n as the following equation $M = f_M(q, c, n)$.

The following figure (Figure1) shows an example relationship which is among the engine power (I.M.E.P) , the quantity of diesel and engine speed in the condition of CNG valve opening at 5 steps which is driven by stepping motor.

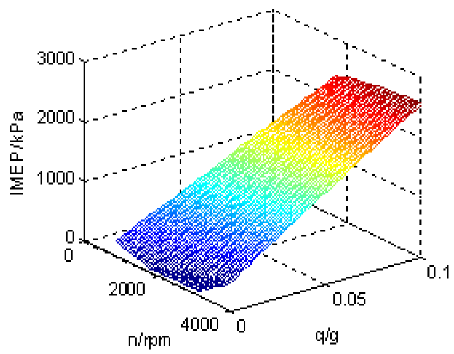


Figure 1. The Character of Engine Power
A hydraulic dynamometer is deployed in the test bench. Its

dynamic character can be described as $M_p = a + bn^2$, Here, parameter *a* is a constant and parameter *b* can be seen as a turbulent variable during working condition changes. The quantity of diesel *q* is determined by the position of pump rack *d* and engine speed *n* $q = f_q(d, n)$.

The function can be get by the data fitting according the result of experiment (Figure2).

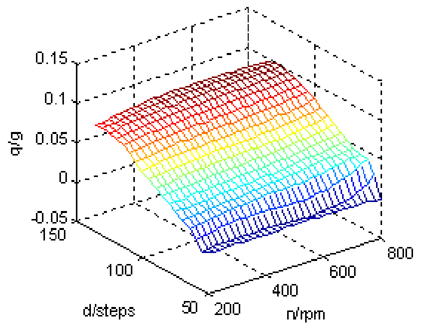


Figure 2. The Character of Diesel Pump

The quantity of CNG *c* is determined by the opening of CNG valve *v* and engine speed *n* (shown in Figure 3) $c = f_c(v, n)$

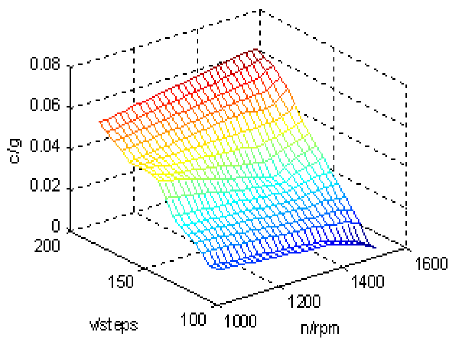


Figure 3. The Experiment Result of CNG Valve Flow

In order to design the engine model, the diesel pump rack position and the opening of CNG valve are taken as the control input variables and the quantity of diesel, the amount of CNG, engine torque and engine speed are taken as the state variables. Load can be seen as a turbulent variable. Using the relationships among those variables, which are discussed above, the model can be described as the following equations:

$$\dot{q} = f_1(q,n) + g_1(q,n) \bullet d$$
$$\dot{c} = f_2(c,n) + g_2(c,n) \bullet v$$
$$\dot{n} = f_3(q,c,n)$$
$$y = h(n)$$

All the equation in the model is get by means of polyfitting according the result of experiments. The fitting errors is shown in the Table1. The results in the chat reflect that the data fitting error is low. That is, the accuracy of the model is available.

Table1. Data Fitting Error

	Diesel pump	CNG valve	Engine torque
Data fitting error (Max)	3.1%	5.4%	3.5%

The model structure is described as following figure (Figure4).

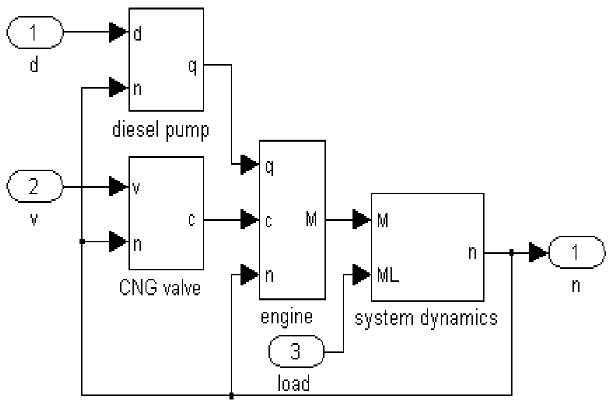


Figure 4. The Structure of Engine Model

3. VARIABLE STRUCTURE CONTROLLER DESIGN

Although the accuracy of the model is available, the error does exist. The cylinder combustion cycle fluctuation makes the engine has some tiny different performance between the engine cycles. Both these factors determines that the robustness should be taken into account in the control design. The characteristic of variable structure control is to make the system respond fast and to be robust extremely, i.e., to be not sensitive extremely to mode error, change of target parameters and perturbations outside [6]. Because VSC and feed back linearization have some unique advantages in model uncertainty, VSC and feed back linearization methods are used in this controller design to control the quantity of diesel and CNG.

The desired state X_d be considered, the new state variables *X* can be used to create a set of new equations to describe the model in which the state variables $X=(X_1, X_2, X_3)^T$ are the difference between the desired values and real values. That is

$$X_1 = q_{desired} - q$$
$$X_2 = c_{desired} - c$$
$$X_3 = n_{desired} - n$$
$$U_1 = d$$
$$U_2 = v$$

The system can be described as the following equation:

$$\begin{cases} \dot{X} = f(X) + g(X) \cdot U \\ Y = h(X) \end{cases}$$

The system model defined by the above equations is a kind of affine nonlinear system. According to the VSC design method, the sliding mode switch surface can be set as $s = C * X$

The equivalent control variable is

$$u_{eq} = -[C * g(X)]^{-1} C * f(X)$$

The switch surface equations is

$$\begin{cases} \dot{X} = [I_n - g(X)(C * g(X))^{-1} C] f(X) \\ C * X = 0 \end{cases}$$

By means of the differential geometry feed back linearization measure, the pole assignment can be proceeded to get the switch surface which has quality dynamic characters.

To design the control law of the system, the function switch control method is adopted. That is the controller is composed of two parts: $u = u_{eq} + u_{vss}$. Here, the switch control part u_{vss} should be connected to the desired state vector. It has the following form:

$$u_{vss} = g(X)^{-1}(ds/dt - C*(X_d, \dot{X}_d)^T - k*\mathrm{sgn}(s))$$

In the switch controller described above, the factor *k*sgn(s)* is the key of switching tremble. In order to improve the performance of switching, a Fuzzy logic is adopted to approach this factor.

In the Fuzzy logic algorithm of approaching k*sgn(s), s and ds/dt are defined as the input variables, and m as output variable: $m = Fuzzy(s, ds/dt)$ The Fuzzy logic set is defined as{ PB (positive big), PM (positive medium), PS (positive small), ZE (zero), NS (negative small), NM (negative medium), NB (negative big) } Setting s, ds/dt and m evaluate these 7 Fuzzy subsets. Their domains are {-3, -2, -1, 0, 1, 2, 3}.

From experience, the control law is designed as the following chart (Table2). The defuzzification of the Fuzzy logic system is centroid method.

Table2. Fuzzy Logic Control Law

		s						
		NB	NM	NS	ZE	PS	PM	PB
ds/dt	NB	NB	NB	NB	NB	NM	NS	ZE
	NM	NB	NB	NM	NM	NS	ZE	PS
	NS	NM	NM	NM	NS	ZE	PS	PS
	ZE	NM	NM	NS	ZE	PS	PS	PM
	PS	NS	NS	ZE	PS	PM	PM	PM
	PM	NS	ZE	PS	PM	PM	PB	PB
	PB	ZE	PS	PM	PB	PB	PB	PB

Synthesized the results discussed above, the SVC controller is

$$u = g(X)^{-1}(\dot{s} - C*D - m) - [C*g(X)]^{-1}C*f(X)$$

4. REALIZING THE CONTROLLER AND CONTROL SYSTEM BENCH TEST

In the environment of MATLAB/ SIMULINK, many kinds of mathematic blocks are used to express the control algorithm. Combined with the algorithm blocks, some logic blocks are used to handle the other functions which are necessary for engine operation. The following figure shows the real controller structure (Figure5) .

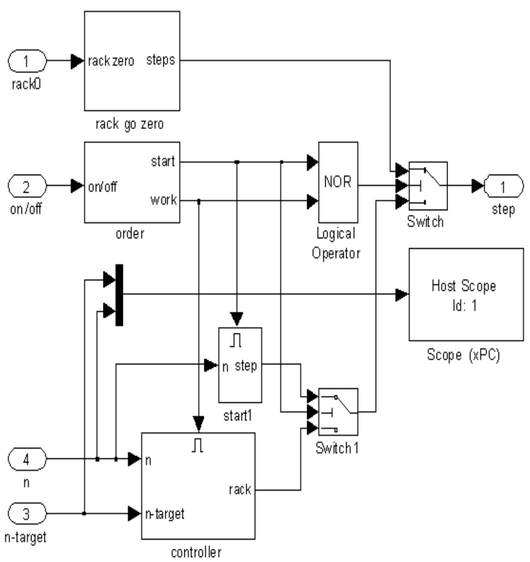


Figure 5. Controller Structure in MATLAB / SIMULINK

For the purpose of convenience in realizing and modifing algorithm, MATLAB/ xPC Target is employed to control the signal of engine. This platform is a subsystem of MATLAB/ SIMULINK. It is an environment that uses a target PC, separate from a host PC, for running real-time applications. The xPC Target software environment includes many features to prototype and deploy a real-time system.

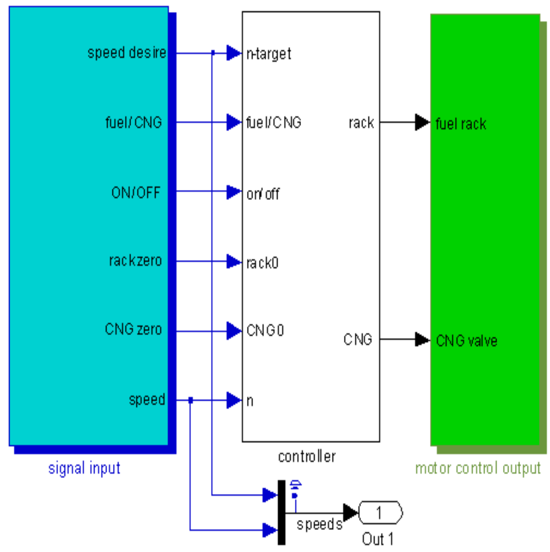


Figure 6. Controller Structure in Xpc Target

The input and output function are realized by Advantech adapters PCI-1713, PCI-1760 and PCI-1780 and the drivers for these adapters are developed. Some target signal feedback measures are used in the system to monitor and control the engine situation. The structure of controller in xPC Target is shown as the Figure 6.

Some engine experiments in test bench can inspect the performance of the controller in the situation of some sudden changes especially when the engine speed changes from 800rpm to 1000rpm then loads suddenly, the process of speed and fuel are shown in Figure 7.

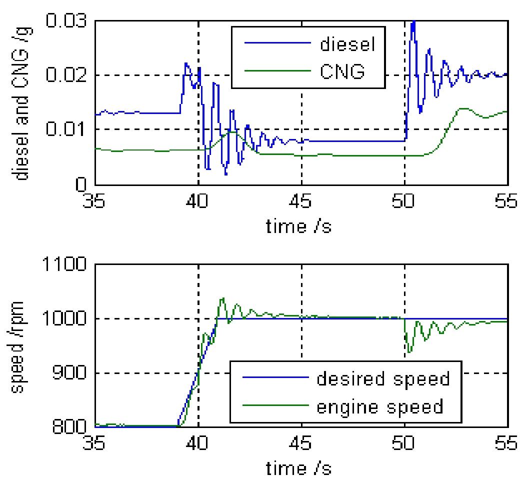


Figure 7. Speed Changes from 800rpm to 1000rpm

When the engine speed changes from 1000rpm to 1200rpm then unloads suddenly, the process of speed and fuel are shown in Figure8.

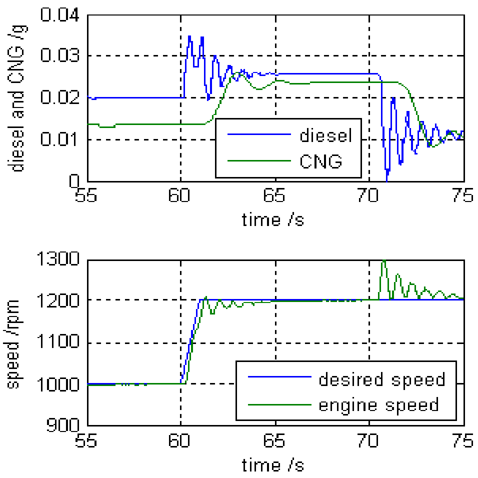


Figure 8. Speed Changes from 1000rpm to 1200rpm

When the engine speed changes from 1000rpm to 1400rpm and then loads suddenly, the process of speed and fuel are shown in Figure9.

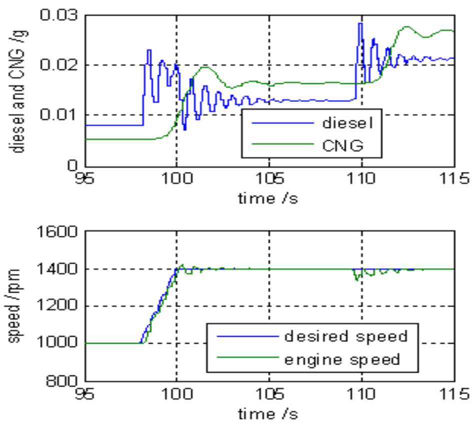


Figure 9. Speed Changes from 1000rpm to 1400rpm

The transient speed regulation rate of engine is about 7%. The response time is less than 4 seconds (Δ=95%). The results of experiments reflect that the system has a quality dynamic performance.

5. CONCLUSIONS

A non-linear model has been established to simulating a dual-fuel engine speed regulation process in this paper. Combined with feedback linearization method, a Fuzzy-Variable Structure Control method has been applied to design the engine governor. By means of the xPC Target platform, a control system is realized and meet the demands of the engine speed regulation in bench test.

The engine performance on the testing bench reflect that the VSC method, combined with feedback linearization method and Fuzzy logic algorithm, is a useful tool for engine governor design. The engine with a VSC controller has a smooth process in speed regulation and a quality stabilization.

REFERENCES

[1] D. E. Winterbone S. Jai-In, The application of modern control theory to a turbocharged diesel engine powerplants, Proc Institute Mechanical Engines,1991, (Part I): (67~83)

[2] Li Guoyong, Ma Zhifang, Intelligent control system based on Fuzzy-reasoning. 6th International Symposium on Test and Measurement, (Vol 2), China, June 1-4, 2005,2029-2032.

[3] Li Yuhui,Feedback Linearization Method in Speed Control on Electronically Controlled Diesel Engine, Journal of Wuhan University of Technology (transportation Science and Engineering),2003. 6:901~904

[4] V. Utkin, Variable structure systems with sliding modes, IEEE Trans. Automat. Contr., vol. AC-22. pp. 212-222.1977.

[5] S. Sarpturk, S.Y. Istefanopulos, and O. Kaynak, On the stability of discrete-time sliding mode control systems, IEEE Trans. Automat. Contr., vol.32.pp. 930-932, 1987.

[6] Bouyoucef K., and Khorasani K., Variable structure-based nonlinear congestion control for networked controlled systems, IEEE Physics and control. 2005: 583~586.

NATURAL GAS PIPELINE LEAKAGE DETECTION BASED ON OPTIC FIBER INTERFEROMETER SENSOR IN LABORATORY-SCALE EXPERIMENT*

Wangjia Zhang [1], Qiang Wang [2]
[1] College of Modern Science and Technology
[2] Department of Quality and Safety Engineering
China Jiliang University, Hangzhou 310018, Zhejiang, China
Email: qiangwang@cjlu.edu.cn

ABSTRACT

A fully distributed optical fiber sensor of gas pipeline leak detection system is demonstrated based on spectrum analysis of interferometer optical time domain reflectometer (OTDR) system. To effectively detect pipeline leaky acoustic signals, Sagnac interferometer method was performed during source localization process, combined with wavelet transform. It improved the spectrum null-frequency localization precision. The feasibility of separately locating multiple leaky source was theoretically analyzed. The laboratory-scale experiment is conducted. It was demonstrated that the optic fiber sensor can measure a single source leak in real time, with a spatial positional resolution better than 4 m in a 2 km spacing of optical fiber system.

Keywords: Gas Pipeline, Leak Detection, Optical Fiber Sensor, Interferometer.

1. INTRODUCTION

In China, natural gas main pipelines include western to eastern line, San-Jing line and Zhong-Wu line *et al*. In total, the large scale oil/gas pipelines are successfully laid in different geological conditions. But the pipeline safety operation monitoring and leakage detecting technologies lag behind, especially the gas pipeline leak testing technology. To 2020, natural gas consumption will reach 2×10^{11} m^3, 2.5×10^4 km new gas pipeline will be constructed and underground gas tank dischargeable capacity will be 95×10^8 m^3 in China [1].

However, leaks, which are the main faults of gas pipelines, can cause serious problems related not only to the environment but also to economy. Therefore, many methods and techniques for leak detection with various applicability and restrictions have been proposed to prevent further loss and danger. The primary leak detection methods include acoustic monitoring, optical monitoring, gas sampling, soil monitoring, flow monitoring, magnetic flux leakage, and dynamic model-based methods [2-3].

Optical fiber leak detection system has many advantages, such as being flexible, insulating, corrosion resistant, etc. Thus, fiber optic technology has received a great deal of attention and some pipeline leak detection systems based on fiber optic sensing technology have been developed [4-5].

Optical monitoring methods can be classified as either passive or active. Active methods involve the illumination of the area above the pipeline with a radiation source, usually a laser or a broad band source. Then the absorption or scattering caused by gas molecules above the surface is monitored using an array of sensors at specific wavelength [6]. In contrast to active methods, passive methods do not require a source. They detect the radiation emitted by the natural gas, or the background radiation serves directly.

Distributed fiber-optic sensing is to detect and locate non-concurrent external signals along the whole length of a fiber. The most distinguished techniques for distributed sensing involve the use of the optical time domain reflectometer (OTDR) and the optical frequency domain reflectometer (OFDR) [7-9]. They sense signals by measuring the optical characteristics of various reflective lightwaves coming from spatially-distributed backscattering such as Rayleigh, Brillouin or Raman. However, there are some limitations on these traditional methods. For example, the scattering light is usually so weak that the performance requirement of detection techniques is high but the sensitivity is low. Also, the injected pulsed light is not suitable for detecting time-varying signals. In these cases, more economic and effective methods are in great demand in the distributed sensing areas [10].

In this paper, based on hybrid interferometer (Sagnac and Mach-Zehnder interferometer) proposed by Huang et al 2007[11], we investigated the multi-leak localization method of natural gas pipeline, and multi-resolution method wavelet transform is implemented to improve signal SNR and capture null-frequency position. The sensor is able to locate and recover the leak signal simultaneously in near three kilometer spacing. The operation performance of the proposed device is theoretically analyzed and experimentally demonstrated.

2. PRINCIPLE OF FIBER INTERFEREMETER

The hybrid configuration of Mach-Zehnder and Sagnac interferometers scheme had proposed by Huang *et al* in 2007. We conducted this scheme in laboratory scale. The Figure 1 presents the hybrid configuration of Mach-Zehnder and Sagnac interferometers with light source as the broadband source. The interferometer is composed of two couplers. The two terminals of the first coupler are connected to a source and a detector respectively. One terminal of the second coupler is connected to a sensing fiber before Faraday rotator mirror (FRM). There are two fiber arms between these two couplers where the longer arm acts as a delay line fiber.

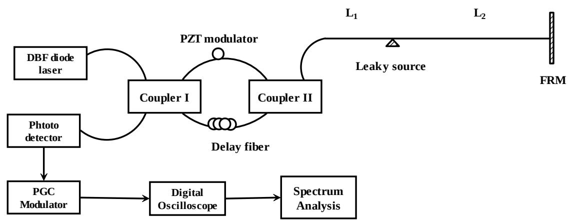


Figure 1. The Hybrid Configuration of Mach-Zehnder and Sagnac Interferometer (Huang *et al* 2007)

3. LEAK SOURCE LOCATION PRINCIPLE

Achieve single leakage accurate positioning is one of the keys to the entire fiber-optic detection system. From Fig. 2, we can see that paragraph in the leak detection to Sagnac interference

* the project supported by the Scientific Research Fund of Zhejiang Provincial Education Department (No. 20070644) and Natural Science Foundation of Zhejiang Province of China

effect of two lights, are able to receive a high signal to noise ratio (SNR).

Path I

$$E_1 = E_{10}\exp\{j\omega_c t + j\Delta\phi(\omega_s)[\sin\omega_s(t-\tau_1) + \sin\omega_s(t-\tau_2)] + j\Delta\phi_m[\sin\omega_m(t-\tau_{m1})] + j\psi_1\} \tag{1}$$

where E_{10} is the amplitude of the electric field of the light beam propagates through Path I to the detector D, ω_c the angular frequency of the optical carrier, ψ_1 the arbitrary optical carrier phase angle, ω_s the angular frequency of the leakage induced acoustic phase signal, $\varphi(\omega_s)$ the amplitude of the leakage induced acoustic phase signal, ω_m the angular frequency of the modulation carrier phase signal, φ_m the amplitude the modulation carrier phase signal, c the velocity of light, n the refractive index of the fiber core, $\tau_1 = n(B + L_1)/c$ and $\tau_2 = n(B + L_1 + 2L_2)/c$ the time duration of the first and second time pass through leaky position in the Path I, $\tau_{m1} = nB/c$ is the duration which the light beam pass through the modulator in the Path I.

Path II

$$E_2 = E_{20}\exp\{j\omega_c t + j\Delta\phi(\omega_s)[\sin\omega_s(t-\tau_3) + \sin\omega_s(t-\tau_4)] + j\Delta\phi_m[\sin\omega_m(t-\tau_{m2})] + j\psi_2\} \tag{2}$$

where E_{20} is the amplitude of the electric field of the light beam propagates through Path II to the detector D, ψ_2 the arbitrary optical carrier phase angle, $\tau_3 = n(D+ L_1)/c$ and $\tau_4 = n(D+ L_1 +2L_2)/c$ the duration of the first and second time in the Path II to pass through leaky position and $\tau_{m2} = n(D+2L_1 +2L_2)/c$ is the duration to pass through the modulator in the Path II.

(1) When $\sin\omega_s\ (\tau_d/2)=0$, the zero frequency has happened, τ_d with fiber-optic delay-related, and
$\sin\omega_s(\tau_d/2)= N\pi, f_s = N \bullet C / (4n \bullet (D \bullet B))$

where D is the delay fiber length, D-B equal the length of fiber delay. In order to reduce the impact of it, As far as possible make it in a broadband range of larger values can be maintained in the optical path design, can make use of Matlab software simulation, in the range of 0-60kHz to find the best value, so that the highest signal to noise ratio demodulation.

(2) When $\cos(\omega_s\tau_a) = 0$, there will be zero frequency.
$f_s = (2N+1) \bullet c / (4n \bullet L_2)$

L_2 for the leakage points to the distance between the Faraday rotation mirror. And zero-point frequency and leakage sensitivity, can be expressed as $df_s / dL_2 = c / 4nL_2^2$, therefore the sensitivity of the system as leakage points, or interfere with the location of the different change, is not a fixed. Simplification of the interference term can be expressed as:

$$I_{\text{int}} = \cos[4\Delta\varphi\cos\omega_s(t-\tau_T/2)\sin\omega_s(\tau_d/2)\cos(\omega_s\tau_x)].$$

Therefore the frequency and length of optical fiber delay will affect the interference light intensity, ω_s is the angular frequency of leakage signals, τ is the time of light after delay fiber, τ for the time of fiber-optic leak spread to Faraday rotation mirror. If the τ_x is constant, there exists one or more zero frequency points. The first zero frequency point in total spectrum, is the corresponding leaky point. [Huang 2007].

4. DEMODULATION OF ACOUSTIC SIGNAL

When a high-pressure pipe leaks, the broadband acoustic signal it generates can be up to 60 kHz [12]. If the demodulating bandwidth of the PGC demodulation is too narrow, it will cause serious signal distortion. Hence, if we expect to demodulate this broadband signal completely, we have to use higher carrier frequency for the PGC demodulation.

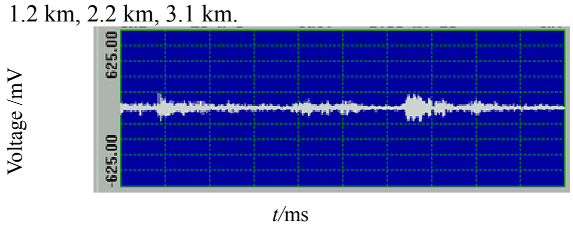
By Frederick Dimmer optical interference items to be necessary demodulation, is about to leak source / sources of interference caused by phase changes in a complete extraction of interference signals. Preliminary studies show that the gas pipeline leakage signals are generally concentrated in the following 50-60 kHz broadband signal, Taking into account the high sensitivity and large dynamic range, the phase demodulation circuit will have a common carrier phase (PGC) demodulation circuit modified with a wide frequency response (bandwidth design around 100 kHz), and to adjust for the carrier frequency 100-400 kHz, which is conducive to the best phase information[13-14].

5. LABORATORY-SCALE LEAK SIMULATION TEST

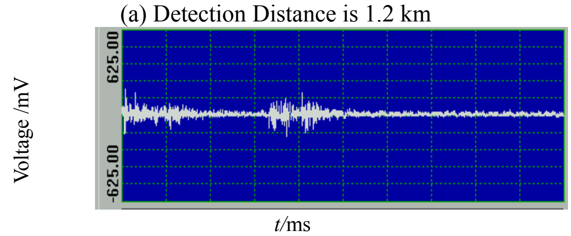
5.1. Experimental Data Analysis

In order to measure the leaky spectrum of sensing fiber, we devise a suit of experiment. The experimental configuration is shown in Fig. 1. The set-up includes NI multi-channel data acquisition system, FFT spectrum analyzer, signal generator, photoelectric detector, Tektronix TDS 3200 oscilloscope such as the air compressor used in laboratory. Optical devices and laboratory equipment: Faraday rotation mirror, optical isolator.

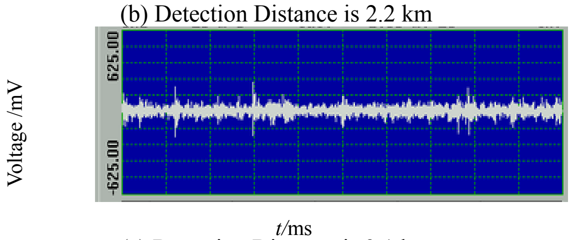
In experiments, found that as the pressure increases, the intensity of the leakage signal will be enhanced, the majority of signal strength is concentrated in the 50 kHz [12]. below the corresponding sensor can be measured fiber length is about 1 km to infinity, the more the better low-frequency, if this band as a target signal leakage measurement region can receive more good results. Therefore, the optic fiber leakage detection system owns higher sensitivity and larger dynamic range. Partial results of the gas pipeline leak detection experiment were shown in Figure 2: abscissa units are kHz, ordinate units are normalized spectrum intensity, Figure 2 (a, b, c) are acoustic signals of three conditions: leaky source distances are 1.2 km, 2.2 km, 3.1 km.



(a) Detection Distance is 1.2 km



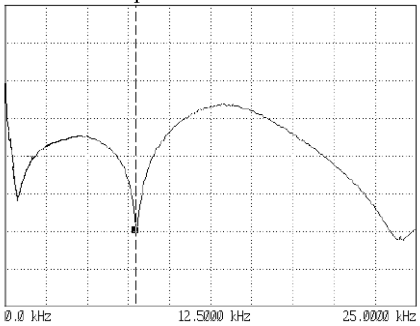
(b) Detection Distance is 2.2 km



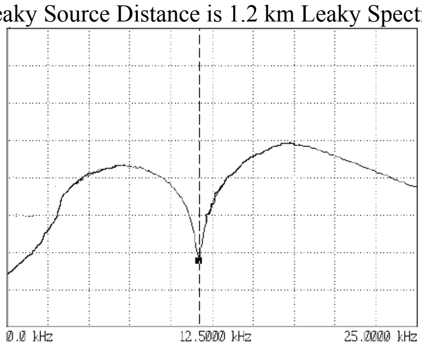
(c) Detection Distance is 3.1 km

Figure 2. Different Leaky Source Localization (1.2 km, 2.2 km, 3.1 km) Signal Gathered by Optic Fiber Detection System under Same Pressure (0.5 MPa)

In Figure 3, based on the Sagnac interferometer localization theory, we search the null-frequency in spectrum chart. Through the null-frequency, the leaky source position is gained. In the dotted line corresponds to the location of zero frequency.



(a)Leaky Source Distance is 1.2 km Leaky Spectrum



(b)Source Distance 2.2 km Spectrum

Figure 3. Leaky Spectrums of Null-frequency Point Capture Leaky Source of Three Distances

6. CONCLUSION

Based on hybrid configuration of interferometer as sensing frame to detect the leakage position of the gas pipeline. The experimental set-up is conducted. We investigated the leak localization method of natural gas pipeline, and Sagnac method is implemented to improve signal SNR and capture null-frequency position. Then the laboratory-scale experiment proved leaky position can be acquired from null frequency of the output spectrum.

We have experimentally demonstrated that the sensor can measure both the frequency and position of a single vibration in real time, with a spatial positional resolution better than 4 m in about 2.2km prototype system. Moreover, we have successfully verified the separately locating scheme by performing the experiment of two vibrations with different frequencies acting at the same position. With further improvement, multiple leaky sources with different frequencies acting at different positions are on the way to be experimentally investigated.

REFERENCES

[1] Yang Xiaohen. Oil and Gas pipeline Safety engineering. Beijing: China Petrochemical Press. 2005 (in Chinese)

[2] Ming Liu, Shu Zang, DongHua Zhou, “Fast Leak Detection and Location of Gas Pipelines based on an Adaptive particle Filter”, International Journal of Applied Mathematics and Computer Science, Vol. 15, No. 4, 2005, pp.541~550.

[3] Benkherouf A. and Allidina A.Y. , “Leak detection and location in gas pipelines”, IEE Proc., Vol. 135, Pt. D, No. 2, 1988, pp. 142~148.

[4] Lijun Hang, Cunfu He, Bin Wu, “Novel distributed optical fiber acoustic sensor array for leak detection”, Optical Engineering, Vol.47, No.5, 2008, pp. 054401-1~054401-6.

[5] A. Carrillo, E. Gonzalez, A. Rosas, A. Marquez, “New distributed optical sensor for detection and localization of liquid leaks Part I. Experimental studies”, Sensors and actuators. A, Physical , Vol.99 , No. 3, 2002, pp. 229~235.

[6] T. Iseki, H.Tai and Kimura K., “A portable remote methane sensor using a tunable diode laser”, Measurement Science and Technology, Vol. 11, No. 6, 2000, pp. 594~602.

[7] J. P. Kurmer, S. A. Kingsley, J. S. Laud, “Distributed fiber optic acoustic sensor for leak detection”, Proceeding. SPIE, Vol. 1586, 1991, pp. 117~127.

[8] A. A. Chtcherbakov, P. L. Swart, S. J. Spammer, B. M. Lacquet, “Modified sagnac/mach- zehnder interferometer for distributed disturbance sensing”, Microwave and Optical Technology Letters, Vol. 20, No. 1, 1999, pp. 34~36.

[9] W. W. Lin , “Novel distributed fiber optic leak detection system”, Optical Engineering, Vol. 43, , No. 2, 2004, pp. 278~279.

[10] Q. Z, Sun et.al, “Distributed fiber-optic vibration sensor using a ring Mach-Zehnder interferometer”, Optics Communications, , Vol. 281, No. 6, 2008, pp.1538~1544

[11] S. C. Huang, W. W. Lin, M. T. Tsai. , “Fiber optic in-line distributed sensor for detection and localization of the pipeline leaks”, Sensors and Actuators A: Physical, Vol.135, No. 2, 2007, pp.570-579.

[12] W. A. Wassef, M. N.Bassim, “Acoustic emission spectra due to leaks from circular holes and rectangular slits”, Journal of Acoustical Society of America, Vol.77, No. 3, 1985, pp. 916~923.

[13] P. R. Hoffman, M. G. Kuzyk, “Position determination of an acoustic burst along a Sagnac interferometer”, Journal of Lightwave Technology, Vol.22, No.2, 2004, pp. 494~498

[14] Z. Y. Zhang and X.Y. Bao, “Distributed optical fiber vibration sensor based on spectrum analysis of Polarization-OTDR system”, Optics Express, Vol. 16, No. 14, July 2008, pp.10240~10247.

Zhang Wangjia was born in 1989. She is an undergraduate student in College of Modern Science and Technology, China Jiliang University. Her major is safety engineering. She focuses her attention on long-distance gas pipeline leak detection system and safey assessment, and was familiar with C language and Matlab software programming. She has published one Journal paper. Her research interests are in safety inspection technology.

Wang Qiang is an associate Professor in Department of Quality & Safety Engineering, China Jiliang University. He graduated from Zhejiang University in 2005; Ph. D major in control theory and engineering. Post doctoral in electronic engineering of Zhejiang University from 2005 to 2007. He is currently a Senior Research Associate with intelligent sensor system, in China Jiliang University. He has published over 10 Journal papers. His research interests are in parallel signal processing, optic fibre sensor and non-destructive testing.

APPLICATION OF BP NETWORK IN EVALUATION OF GROUNDWATER QUALITY

Yanbin Ning, Shuiming He
School of Mathematics and Physics, China University of Geosciences
Wuhan, Hubei 430074, China
Email: iamanxin1314@163.com

ABSTRACT

The aim of this paper is to evaluate the condition of the groundwater quality in lower reaches of YiTong river, which is based on the principle of the golden section to search the perfect number of hidden layer nodes, and then to establish the BP neural network. Furthermore, by comparing with fuzzy comprehensive evaluation method, gray clustering method and Nemero index method[2] respectively, it turns out BP neural network got a well performance.

Keywords: BP Neural Network, The theory of golden section, Hidden layer nodes

1. INTRODUCTION

The BP neural network is an artificial intelligence system, which simulates the nerve network of biological and also has some good qualities, such as robustness, fault tolerance, self-organization, self-learning ability and fuzzy characteristics, etc. In the mean while, the BP network has good high-dimensional nonlinear mapping ability and generalization ability, which can be able to approximate any continuous function and the square integral function at any precision, especially in the evaluation of groundwater quality, it can distinguish the ambiguity, uncertainty and non-linearity of groundwater environment efficiently.

The number of hidden layer nodes, which determines the performance, is the most crucial part of the BP neural network. Since there is no theory which can prove itself is the best one to search the perfect hidden layer nodes and most of the methods are still based on the empirical formula or priori information. This paper is mainly based on the principle of golden section [1], and make full use of the priori information to get the interval where the perfect hidden layer nodes frequently appears, finally, an MATLAB program was designed to search optimization directly rather than to test all the nodes one by one to meet a certain accuracy, the simulating experimental results show that this method can avoid blind search for perfect number of hidden nodes, so it's time-saving.

2. INTRODUCTION OF BP NETWORK AND THE PRINCIPLE OF GOLDEN SECTION

2.1 Brief Introduction of the BP Network

The three-layer BP Network is able to approximate any continuous function at any precision, so the BP Network with only one hidden layer can establish the relationship of non-linear mapping between two space, and if the required precision can not meet, one more hidden layer is needed. However, in practical applications, it is necessary to ensure that the trained BP Network has good generalization ability for the testing samples, that is to say, the network should approximate the inherent law of the samples, rather than just pursue the ability to fit the training samples. The Structure of BP Network is shown in Fig.1:

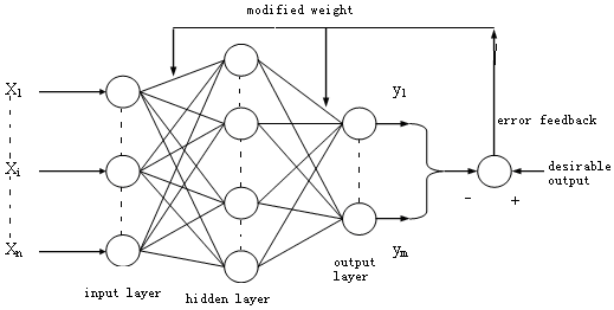


Figure1. The Structure of BP Network

2.2 Data Preprocessing

For input layer, the number of nodes is up to the characteristic parameters of training sample. For the output layer, they are determined by the characteristic parameters of the desired output. Generally, the training sample and the object vector are often not on the same dimension, so certain dimension conversions are required before the training of the network. There are lots of conversions. In accordance with the practical condition the maximum normalization was introduced to this paper.

$$x_{ij}^{'} = \frac{x_{ij}}{\max\limits_{1\le j\le n}(x_{ij})} \quad (i=1,2,\cdots,m;j=1,2,\cdots,n) \tag{2.1}$$

2.3 The Principle of Golden Section

Generally, the number of hidden layer nodes can be determined by the following empirical formula, for the hidden layer, the input layer as well as the output layer, the number of nodes is represented by m, n and l respectively:

$$m = \sqrt{n+l} + \alpha \text{ , } 0 < \alpha < 10 \text{ ;} \tag{2.2}$$

$$m = \log_2 n \tag{2.3}$$

$$m = \sqrt{nl} \text{ ;} \tag{2.4}$$

In this paper the empirical formula is base on the principle of golden section, the formula is as follows:

$$a = \frac{n+l}{2} \le m \le (n+l) + 10 = b \tag{2.5}$$

The interval [a, b] is the scope of the number of hidden layer nodes, which can be obtained through the empirical formula (2.5). All the points of the interval would be tested one by one until reach the required precision, if the trial and error method is chosen. However, if the principle of the golden section is used, blind search for the perfect number of hidden layer nodes is avoided.

The origin of extreme and mean ratio:

$$r_1 = (\sqrt{5}-1)/2 = 0.618; \quad r_2 = r_1^2 = 0.382; \tag{2.6}$$

The golden mean optimization procedure is as follows [1]:

(1) Obtain the interval [a, b] of the approximate number of hidden nodes by the empirical formula (2.5).

(2) Find the fist point x_1 in the interval [a, b], which is the extreme and mean ratio. It is worked out by the formula $x_1 = 0.618\times(b-a)+a$, then record the test result.

(3) Find the second point x_2 in [a, b], which can be obtained by the formula $x_2 = 0.382\times(b-a)+a$, then record the test result.

(4) Comparing the two test results, if the result of x_1 is better than the x_2, then reserve the $[x_2,b]$ and remove the $[a,x_2]$; if not, reserve the $[a,x_1]$ and remove $[x_1,b]$; but if the results of them are equal, remove both the interval $[a,x_2)$ and $(x_1,b]$, just reserve the $[x_2,x_1]$.

(5) Repeat the above-mentioned process in the remaining interval until meet the required precision.

(6) If the perfect number of hidden layer nodes, which can meet the required precision during the interval of $[a,b]$, can not be found, that is to say, the interval should be extended. The endpoint of b will be served as the new extreme and mean ratio, so another endpoint c can be got by the formula: $b = 0.618\times(c-a)+a$ and then repeat the above-mentioned steps in $[b,c]$ until it reaches the error precision.

3. APPLICATION OF BP NETWORK IN EVALUATION OF THE WATER QUALITY

The evaluation criteria is referenced to the GB/T14848-93 of The standard of the ground water, the grade of classification will be served as the training sample. The function of UNFRAND in the toolbox of MATLAB can be used to generate 10 random numbers for every grade as the training sample in order to enhance the generalization capability. Because the grade of IV and V have the same boundary line, in order to distinguish them, the Compromise Algorithm is required, that is to say, the standard I and V are kept , but the standard II, III and IV will be the average value of I and II, II and III, III and IV respectively. The learning samples are shown in the table 2:

Table 1 The Evaluation Standard (1)

The Evaluation standard of groundwater (unit:mg/L)						
	K^+, Na^+	Ca^+	Cl^-	SO_4^{2-}	HCO_3^-	TDS
I	50	50	50	50	150	300
II	100	100	150	150	300	500
III	150	150	250	250	450	1000
IV	200	200	350	350	550	2000
V	>200	>200	>350	>350	>550	>2000

Table 2 The Evaluation Standard (2)

The Evaluation standard of groundwater(2) (unit：mg/L)						
	K^+, Na^+	Ca^+	Cl^-	SO_4^{2-}	HCO_3^-	TDS
I	50	50	50	50	150	300
II	75	75	100	100	225	400
III	125.5	125.5	200	200	375	750
IV	175	175	300	300	500	1500
V	>200	>200	>350	>350	>550	>2000

Table 3 The Date of Groundwater

The date of groundwater(unit：mg/L)						
	K^+, Na^+	Ca^+	Cl^-	SO_4^{2-}	HCO_3^-	TDS
1	16.40	157.39	174.85	236.10	541.29	967.68
2	66.60	142.92	134.17	211.75	477.07	868.70
3	58.20	117.84	126.98	206.10	411.25	782.28
4	23.90	82.65	101.09	191.75	353.84	640.24
5	34.20	63.13	88.35	170.00	280.08	561.81

Table 4 The Desired Response

The desired response					
I	1	0	0	0	0
II	0	1	0	0	0
III	0	0	1	0	0
IV	0	0	0	1	0
V	0	0	0	0	1

Since the sample data is not on the same dimension, so the training samples and testing samples should be preprocessed before training. Experiments show that the method of maximum normalization can get the better outcome, because the different amount of different elements has different effects on human health, sometimes with high ratio of some elements, the water will be of high quality, sure the opposite thing happens as well.

(1) For the Characteristic parameters which show that the more the content of it, the better the water quality is, using the following approach [3]:

$$x_{ij}^{'} = \begin{cases} 0 & x_{ij} > s_{i\min} \\ 1 - x_{ij}/s_{i1} & 0 \le x_{ij} \le s_{i\min} \end{cases} \tag{2.7}$$

The $s_{i\min}$ represents the first level (best) of the *ith* grade.

(2) For the Characteristic parameters which show that the more the content of it, the worse the water quality is, using the following approach:

$$x_{ij}^{'} = \begin{cases} 1 & x_{ij} > s_{i\max} \\ x_{ij}/s_{i\max} & 0 \le x_{ij} \le s_{i\max} \end{cases} \tag{2.8}$$

The $s_{i\max}$ represents the worst level in the *ith* grade, in addition, $s_{i\max} = s_{i5}$ in this paper.

The use of the above-mentioned methods can effectively avoid the Uniform-processing of all the data, reflecting the different content of the elements will have different impact on the human body.

To establish the BP neural network with the Preprocessing data, the Trainlm algorithm is served as the training function, the training error is 1×10^{-5}. The search of the perfect number of hidden nodes is based on the principle of Golden Section, and finally, after the training time reached 237, the network met the required precision and the number of hidden nodes is 13. Fuzzy comprehensive evaluation[4] as well as the gray clustering method [5] and so on was used to evaluate the water quality of the YiTong River as well, the results are as follows:

(1) Evaluation results of BP Network

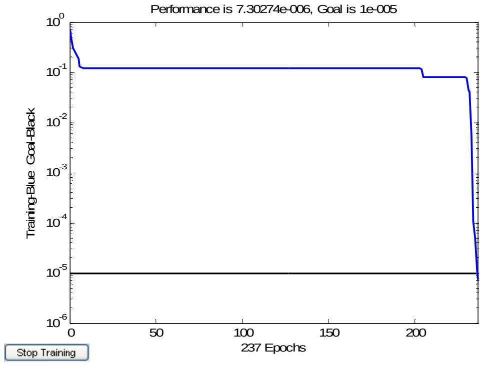


Figure2. The Training Processing of BP Network

Table 5 The Evaluation Results of BP Network

grade / sample	The evaluation results of BP Network				
	I	II	III	IV	V
1	0.2665	0.3954	**0.8883**	-0.6974	0.3560
2	0.2558	0.0594	**0.9292**	-0.4853	0.2324
3	0.1207	0.0976	**0.9489**	-0.5837	0.4217
4	-0.0559	0.5595	**0.8803**	-0.4952	0.5449
5	-0.1255	**0.8456**	0.5826	-0.2121	0.4598

(2)Evaluation results of fuzzy comprehensive evaluation method

Table 6 The Results of The Fuzzy Comprehensive

grade / sample	The evaluation results of the fuzzy comprehensive (fuzzy matrix)				
	I	II	III	IV	V
1	0.0221	0.1644	**0.3750**	0.2289	0.2095
2	0.0319	0.1418	**0.5148**	0.3115	0
3	0.0625	0.1517	**0.7010**	0.0849	0
4	0.0484	0.3761	**0.5755**	0	0
5	0.1922	**0.4734**	0.3344	0	0

(3) Evaluation results of gray clustering method

Table 7 The Evaluation Results of Gray Clustering

grade / sample	The evaluation results of gray clustering (Grey matrix)				
	I	II	III	IV	V
1	0.2051	0.1004	0.4136	**0.6413**	0.3831
2	0.0689	0.2902	0.6772	**0.7206**	0.2411
3	0.1378	0.3300	**0.8071**	0.6829	0.0656
4	0.2051	0.4779	**0.8200**	0.4357	0
5	0.3300	**0.7089**	0.6959	0.2510	0

(4) Evaluation results of Nemero index method

Table 8 The Results of Nemero Index Method

grade / sample	Evaluation results of Nemero index method (Evaluation index PI)	
1	1.2534	IV
2	1.1319	IV
3	0.9890	III
4	0.8332	III
5	0.7332	II

Note: if PI<0.72,it is grade I,if 0.73<PI<0.76,it is grade II,if 0.77<PI<1.00,it is grade III,if 1.01<PI<7.29,it is grade IV,if PI>7.29,it is grade V.

(5) Evaluation results of the index mean method

Table 9 The Results of the Index Mean Method

grade / sample	Evaluation results of the index mean method (Evaluation index PI)	
1	1.0289	IV
2	0.9716	III
3	0.8680	III
4	0.6851	III
5	0.5939	III

Note: if PI<0.20, it is grade I, if 0.21<PI<0.37, it is grade II, if 0.38<PI<1.00, it is grade III, if 1.01<PI<2.54, it is grade IV, if PI>2.54, it is grade V.

(6) Evaluation results of the Index superposition method

Table 10 The Results of the Index Superposition Method

grade / sample	Evaluation results of the Index superposition method (Evaluation index PI)	
1	6.1732	II
2	5.8295	II
3	5.2078	II
4	4.1104	II
5	3.5632	I

Note: if PI<4.04, it is grade I, if 4.04<PI<7.38, it is grade II, if 7.39<PI<20.00, it is grade III, if 20.01<PI<50.85, it is grade IV, if PI>50.85, it is grade V.

(7) Comprehensive comparing the evaluation results of some of the above-mentioned methods

Table 11 Comprehensive Comparing of the Results

sample	BP Neural network	Fuzzy Comprehensive Evaluation method	Gray Clustering method	Nemero Index method
Sample1	III	III	IV	IV
Sample2	III	III	IV	IV
Sample3	III	III	III	III
Sample4	III	III	III	III
Sample5	II	II	II	II

It can be seen from the evaluation results, the BP network has the same evaluation results as the fuzzy comprehensive evaluation method. However, the BP network need not to set up function as the membership function of the fuzzy comprehensive evaluation method and the whitened function of gray clustering method, and it also avoid adding too much subjective factors.

4. CONCLUSIONS

As the result of self-learning, self-organization and adaptive capacity, compared with other models, the BP network is of strong general-purpose and objectivity in the assessment of groundwater quality. The key to the Neural network is that often does not converge with large-scale and can not satisfy the precision with the small-scale. The serial method of training can be used by adding the network gradually to meet the training accuracy. The serial network can be changed to parallel network in the forecast and Series-parallel hybrid can make the network self-adaptive adjustment in accordance with specific issues. It's capable of selecting the learning parameters and training samples flexibly to modeling, therefore, it has a strong flexibility and adaptability.

BP Neural network is a large-scale parallel computing network, she is to imitate the human brain, but in fact, when we use BP neural network, it's only a single BP neural network for a particular aspect (such as in this article). Large-scale neural networks (in brain) there is too much BP neural network computing and applying at the same time, so the BP Neural network can identify an object from all aspects (side).

In this paper, the BP Neural network used in the assessment of groundwater quality, which is based on the principle of Golden Section to search the perfect number of hidden layer nodes, finally achieved a good assessment, what's more, the evaluation of water quality based on the BP network computes faster, more accurate so that it has a wide range of applications prospect, which has opened up a new way to evaluate the water quality.

REFERENCES

[1] Kewen Xia, Changbiao Li, Junyi Shen. Algorithm on the number of hidden layer nodes in Feed-forward Neural Network [J].Compute science, 2005 Vol.32 No.10.

[2] WeiZhang, Xuejun Yan. Evaluation of groundwater quality in Tianjin City [J].The protection of the water resource, 2005, 21(2):31-35.

[3] Jianfeng Cao, Jianhua Ping. Application of improved BP neural network to groundwater quality evaluation [J].Advances in Science and Technology of water Resources, 2006, 26(3):21-23.

[4] Chunxiang Duan, Quanrong Wang. Application of Fuzzy Comprehensive Assessment to Groundwater Quality Assessment and its Features Analysis[J].Journal of Huaihua University, Aug,2008,Vol.27.No.8.

[5] Beifang He, Xiaoyu Wang, Xiaoju He, Xixia Ma, Yifeng Liu. Water quality assessment method based on Gray-clustering Decision [J].Journal of Zhengzhou University (Engineering Science), 2002, Vol.23 No.1.

APPLICATION OF VIRTUAL INSTRUMENT BASED ON ARM PROCESSOR *

Jianjun Ding
Physics and information school, Jianghan University
Wuhan, Hubei Province, China 430056
Email: ddjjd@126.com

ABSTRACT

This paper introduces a kind of virtual instrument designed with ARM, which software and hardware are given. A method of task dispatch and management is given which is based on Linux embedded operating system. Finally, the example for explanation is given.

Keywords: Virtual Instrument, ARM, Linux, Real Time
.

1. INTRODUCTION

With the deep development of computer, network technology and micro-electronics, especially, the design of some kinds of fine performance system on chip (soc), embedded operating system is widely used in industry control, mobile communication, electric appliance and consumer electronic products etc.

2. BASIC WORK PRINCIPLE OF SYSTEM

This paper introduces a specific application on embedded operating system and data acquisition (DAQ). Structural and modularize design is used in software and hardware design and debug. A general test platform is developed based on embedded control、high speed data acquisition、handle and friendly Graphic User Interface.

3. DESIGN OF HARDWARE

The core chip of the system is S3C2410X and ADS7864, which is a 6-channel, 12-bit, 500 kHz SAR-type A/D converter. It contains 6 sample/hold amplifiers. Its parallel interface can be connected with six FIFO registers in order to acquire data fast. The width of every channel is 16-bit.The range of input signal is 0～5V. Ambipolar signals can be transformed to 0～5V through circuit of signal debug. The signal from A/D converter must be converted due to voltage of I / O of S3C2410X is 3.3V. So 74LVTH16245A is used.

After pretreatment of multiplexer analog signal, it is input A/D converter to take data acquisition by ADS7864 via multiple analog sample/hold switches. After A/D convert, digital signal is read by S3C2410X which has a core of ARM920T. It is displayed on LCD after handling, we can choose or control by keys. Figure 1 is the hardware system.

There is no ADC I/O input on board QT2410, we cannot use it directly. So we designed a general data acquisition circuit by QT2410, which contains voltage converting, logical coding unit, 8-channel analog input and 16-channel digital I/O. So that, some kinds of analog signal and digital signal can be put in /out with adjust signal (such as breadth limited, filter, convert and insulate etc).

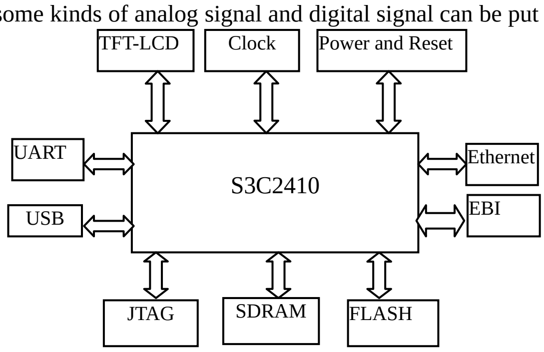


Figure 1. Hardware System

4. SOFTWARE DESIGN

Software design on data acquisition circuit contains: initialization program of the start-up S3C2410X, A/D data acquisition and adjust program, I/O control program, waveform display on LCD program.

4.1 Transplant of Linux

Most of source code of Linux was written in language C. But the only language read/write on processor and register is assemble language, so we still used assemble language to complete the code related to processor. Transplant was considered when Linux was designed, so the transplant was focused on the part of processor, including code related to processor and setting of the applications related to files. ADS1.2 was used to be compiler to accomplish the work above. Then Linux can be run on the ARM chip.

4.2 Development of Application Process

All the peripheral equipment was regarded a kind of special file named "device file" in Linux. Drive process is the interface of the core and external device in operation system. It controls the operation system and hardware. Users can control the physical devices by using normal application process which came from driving process. Device drive process shielded specific details of hardware, so application process can control hardware device like other normal files.

Drivers of GPIO, keys, LCD, DA, and applications process were developed based on Linux. This text introduces the development of DA circuit. With the interface of ADS7864 and S3C2410X.The mode of continuous and interrupt is used to convert by using ADS7864.It works on highest frequency. Read result from ISR.

* Sponsored by Hubei province natural science foundation (2007AB A302), Wuhan science & technology bureau (20085179952 4-06), Wuhan educational bureau (2008K057).

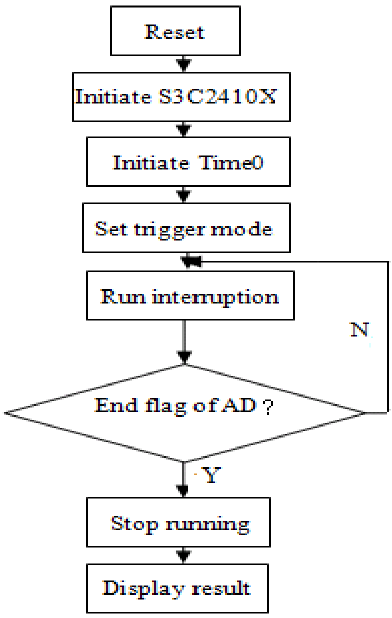


Figure 2. Program Flow Chart on ADS7864

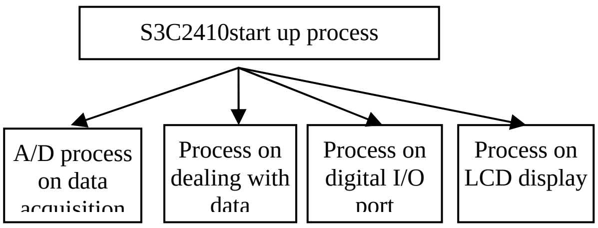


Figure 3. Start up Process

Register device first while start ADS7864.
#define ADC_MAJOR 245//main device number
#define DEVICE_NAME"ADS7864"//name of device
ret=register_chrdev(ADC_MAJOR,DEVICE_NAME,&ADS7864_fops);
setting trigger mode
set_gpio_ctrl(GPIO_F1|GPIO_MODE_EINT|INT_EINT1);
set_external_irq(IRQ_EINT1,EXT_FALLING_EDGE,GPIO_PULLUP_DIS);
start operation of device ,register intrupt
request_irq(ADC_MAJOR,ADS7864_irq,SA_INTERRUPT,DEVICE_NAME,NULL);
MOD_INC_USE_COUNT;
accordingly, realease interupt while shut down device
free_irq(ADC_MAJOR,NULL);
MOD_DEC_USE_COUNT;
Start ADS7864 to convert send data from ISR to user
static DECLARE_WAIT_QUEUE_HEAD(ADS7864_wait);//define a waiting queue
(void*)(adc_cs)=ioremap(0x08008000,0x2);//mapping 16-bit address
(volatile unsigned short)(adc_cs)=0;//start convert while writing
interruptible_sleep_on(&ADS7864_wait);//process block
copy_to_user(buffer,&adc_value,sizeof(adc_value));
in 7864_irq(ISR)，read high-12-bit
(void*)(adc_convst)=ioremap(0x08008800,0x2);
adc_value=*(volatile unsigned short*)(adc_convst)>>4;
wake_up_interruptible(&ADS7864_wait);//process awaken
Converted data can be read from ADS7864 while run read in process application. In order to making ADS7864 working under certain circumstances, driver of timing device need to be ran. Control of timing is the basic principle, include setting of time, writing ISR of timing, apply and release of interrupt.

5. EXAMPLE

Taking Spring testing system as an example, main purpose is testing 3-channel digital I/O and 1-channel analog I/O so as to form a system with these functions.

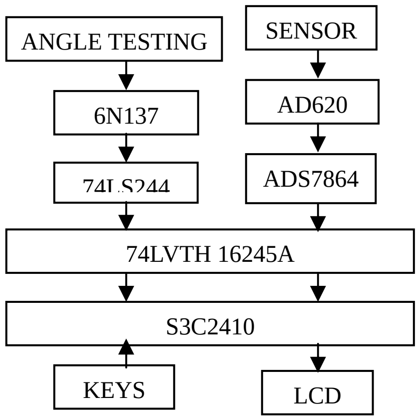


Figure 4. Spring Testing System

Figure 4 is about the whole system. Data which come from sensor was put in virtual instrument after handling. Then it is put in ADS7864 to take sample/hold and acquisition. Digital signal is sent to S3C2410 which has core ARM920T after converting, then, the data displayed on LCD. Keys can be used to select and control. At the same time, task of attemper, manages and communication can help the system accomplish measure and controlling. Transmission of teledata has been added to Linux so as to satisfy the need of DAQ from long-distance. The point is GSM MODEM communicates with RS232, transfers the data which come from GSM MODEM to station of DAQ.

The way of structural and modularize is the key point of the design of application of virtual instrument mentioned above. The system is provided with high value and bright future in industry measurement and control thanks to embedded system, high speed DAQ, real time handle and friendly.
Embedded data acquisition and processing system based on ARM has well-defined architecture, good universality and extensibility. It could provide an integrated hardware and software solution for various embedded applications. Also, it has a wide application prospect in the field of industrial measurement and control.

REFERENCES

[1] Balfazs Vince Nagy, Gyorgy Abraham. Spectral test instrument for colorvision measurement. Journal of Bionics Engineering, 2005, 2(2), pp.075-079
[2] JIN Yongjun, JIANG Youyi. Gradient descent approach for minimizing dissimilarity measure in log-polar imagery. Optoelectronics lettres, 2005,2(4), pp. 0302－0304
[3] Donald.l.Birx. Chaotic oscillator and CMFFNS for signal detection in noise environments IEEE Internation Joint Conference on Neural Networks, 1992.Vol.2, pp.821-888
[4] Shang ping Ren, Gul A. Agha. A modular approach for programming embedded systems. Lecture Notes in Computer Science, 2006.4.11, pp. 170-207

FAST IMPLEMENTATION OF RC6 USING INTEL'S SSE2 INSTRUCTIONS

Qing Hu[1], Xiaodong Xu[2]
[1] College of Computer Science and Technology, Wuhan University of Science and Technology
Wuhan 430081, China
[2] School of Computer Sciences and Technology, Beijing University of Posts and Telecommunications
Beijing 100876, China

ABSTRACT

RC6 is a symmetric block cipher, designed by RSA laboratory to meet the requirements of the AES competition. As one of the five AES finalists, RC6 achieves good performance with a high level of security, and especially fit for parallel processing. SSE2 is a set of Intel's instruction extensions in the IA-32's SIMD programming model. It provides the ability to perform SIMD operations on 128-bit packed integers which greatly enhances throughput and is a natural fit for RC6. In this paper we presents implementation of RC6 to fully exploit the paralleled processing potential of SSE2 instructions. With specific optimization techniques, our implementation can encrypt two blocks parallelly and greatly enhances the overall performance.

Keywords: RC6, SSE2, Parallel Processing, SIMD, Instruction Level Parallelism.

1. INTRODUCTION

With more and more people doing their work through on-line applications, such as E-mail, Google Docs and E-commerce, information security becomes first rate importance. Cryptographic algorithms are designed to meet the goal that ensures confidentiality. Block ciphers, as one type of symmetric-key cipher, are used to encrypt bulks of data. Since the data transfer rates directly rely on the throughput of the algorithm, efficient implementations with high throughput are becoming increasingly important. Besides security, speed is the second main goal in designing cryptographic primitives.

RC6 is a block cipher developed to meet the requirements of Advanced Encryption Standard. As one of the five finalists of the AES candidates, RC6 has not only good security properties but also has an exceptional performance results on PC platforms. Additionally, the features of RC6 benefit significantly from Intel's instruction set extensions SSE2, which is first introduced in Pentium4, to accelerate multimedia

2. AN OUTLINE FOR RC6 AND SSE2

2.1. The RC6 Block Cipher

The RC6 block cipher consists of three components: key expansion algorithm, encryption algorithm and decryption algorithm. We only describe RC6 Encryption algorithm here, for a detailed descriptions of RC6, please refer to [1].

Notations and primitive operations in RC6:

lg(x): base-two logarithm of x
+: Two's complement addition of words
xor: Bit-wise exclusive-OR of words
<<<: Left-rotation of words, the cyclic rotation of words x left by y is denoted x<<<y.

All the above operations are modulo word size.

The RC6 Encryption Algorithm:
Input: Plaintext stored in four w-bit input registers A,B,C,D; Number of r of rounds; w-bit round keys S[0,...,2r+3]
Output: Ciphertext stored in A,B,C,D
Procedure:

```
B = B + S[0]    step 1
D = D + S[1]    step 2
for i = 1 to r do
{
    t = (B*(2B+1))<<<lgw              step 3
    u = (D*(2D+1))<<<lgw              step 4
    A = ((A XOR t)<<<u) + S[2i]       step 5
    C = ((C XOR u)<<<t) + S[2i+1]     step 6
    (A,B,C,D) = (B,C,D,A)             step 7
}
A = A + S[2r+2]    step 8
C = C + S[2r+3]    step 9
```

In our implementation, the word size w = 32, and the iteration round r = 20.

2.2. Intel's SSE2 Extensions Overview

SSE2 was first introduced in Pentium4 and Xeon processors to enhance the performance of multimedia applications, such as 3-D graphics and video decoding. SSE2 extends the SIMD model with support for packed double-precision floating-point values and for 128-bit packed integers. We only care about integer instructions since operations of RC6 are all integer arithmetic.

Take two instructions for example:

paddd: Adds the 4 signed or unsigned 32-bit integers in XMM2 to the 4 signed or unsigned 32-bit integers in XMM1, **Figure1**.

eg: paddd XMM2, XMM1

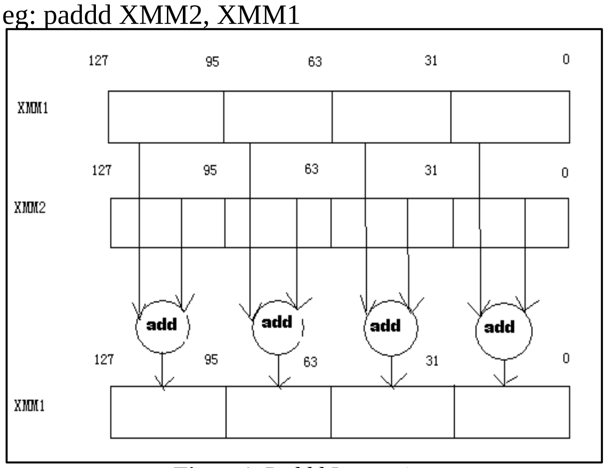


Figure1. Paddd Instruction

pshufd: Shuffles the 4 signed or unsigned 32-bit integers in XMM1 as specified by imm8, **Figure2**.

eg: pshufd XMM2, XMM1, ox1b

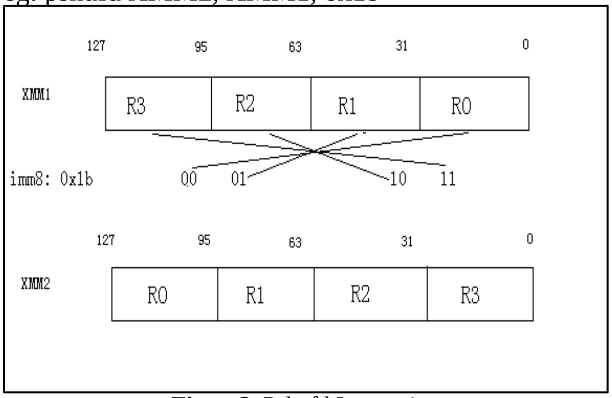


Figure2. Pshufd Instruction

3. OUR IMPLEMENTATION

3.1. Block Level Parallelism

From section 2.1, we observe that the incoming 128-bit plaintext block(Seen as four sequential word A,B,C,D) is not fully processed in parallel. Instead, A and C are processed parallelly, so are B and D. So packing the four plaintext words into one 128-bit XMM register can not benefit from SSE2 instructions. But packing two words into one XMM register do not fully exploit the advantage of SSE2 since SSE2 enables four word integers processed parallelly. So why not process two blocks at a time! Given two blocks $A_1B_1C_1D_1$ and $A_2B_2C_2D_2$, through a series of pack and unpack operations, we get $A_1C_1A_2C_2$ and $B_1D_1B_2D_2$, which can be store in two XMM registers and processed efficiently. After the whole encryption process, we pack them back to $A_1B_1C_1D_1$ and $A_2B_2C_2D_2$ and output the two encrypted blocks .This trick makes our implementation speed almost double.

The following shows the packing process.
Input: $A_1B_1C_1D_1$, $A_2B_2C_2D_2$
Output: $A_1C_1A_2C_2$, $B_1D_1B_2D_2$
Implement:
movdqa xmm0, plaintBlock1 ;xmm0: $A_1B_1C_1D_1$
movdqa xmm1, plaintBlock2 ;xmm1: $A_2B_2C_2D_2$
movdqa xmm2, xmm0
punpckldq xmm0, xmm1 ; xmm0: $A_1A_2B_1B_2$
punpckhdq xmm2, xmm1 ; xmm2: $C_1C_2D_1D_2$
movdqa xmm1, xmm0
punpckldq xmm0, xmm2 ; xmm0: $A_1C_1A_2C_2$
punpckldq xmm1, xmm2 ; xmm1: $B_1D_1B_2D_2$
After the whole encryption is finished, we should pack $A_1C_1A_2C_2$ and $B_1D_1B_2D_2$ back to $A_1B_1C_1D_1$ and $A_2B_2C_2D_2$, thus we get two encrypted blocks.

3.2. Operation Level Parallelism

After two plaintext blocks are packed and reshuffled, the following task is to translate basic arithmetic to parallel SIMD code. As RC6 possesses a good parallel structure, we can properly arrange the steps so that some of them can be processed parallelly. According to section2.1, it is easy to notice that step1 and step2, step3 and step4, step 5 and step 6, step 8 and step 9 can be executed parallelly. Our implementation is described in the following in detail. For simplicity and straightforwardness, Intel' Intrinsics for SSE2 are used instead of raw assembly code. For a detailed description of Intrinsics, please reference to [6].

Our Implementation is as follows:

Notations and definitions:
iOnes: 128-bit constant integer, (1,1,1,1)
imusk: 128-bit constant integer, (word(-1),0,word(-1),0)

step1 and step2:
pshufd(S[0], S[1], 0x44) —> $S_0S_1S_0S_1$: pack the first two words of round key array into 128-bit XMM register;

paddd($B_1D_1B_2D_2$, $S_0S_1S_0S_1$): add packed four words parallelly.

step3 and step4:
pslld($B_1D_1B_2D_2$, 1) —> temp128: logically shift left by one bit, so each four words multiply;

paddd(temp128, iOnes) —>temp128: add 1 for each four words paralleled;

Since there's no packed multiply, we have to multiply respectively.

pand(pmuludq($B_1D_1B_2D_2$, temp128) , imusk) —> low128: calculate D_1*(D_1+1) and D_2*(D_2+1), and we mask the higher part of the two 64-bit integer results.

pand(pmuludq(psrldq($B_1D_1B_2D_2$,4),psrldq(temp128,4)),imusk) —> high128: calculate B_1*(B_1+1) and B_2*(B_2+1), and we mask the higher part of the two 64-bit integer results.

por(low, pslldq(high,4))—>temp128: shift high by 4 bytes and xor with low bit by bit, then we get the value of f(x) = x*(2x+1) for B_1,D_1,B_2,D_2 respectively.

por(pslld(temp128, 5), pslld(temp128, 32-5)) —> temp128: rotate by lg32.

step5 and step6:
pxor($A_1C_1A_2C_2$, temp128)—>$A_1C_1A_2C_2$: bit by bit xor
movdqa(word temp1[4], temp128), movdqa(word temp2[4], $A_1C_1A_2C_2$): store the two 128-bit values

since there's no packed rotate, we have to do it manually.
rotl(temp2[0], temp1[1]);
rotl(temp2[1], temp1[0]);
rotl(temp2[2], temp1[3]);
rotl(temp2[3], temp1[2]).

movdqa($A_1C_1A_2C_2$, temp2): save back to $A_1C_1A_2C_2$

pshufd(S[2i] S[2i+1], 0x44) —— $S_iS_{ii}S_iS_{ii}$: pack current round key S[2i] and S[2i+1] into XMM register;

paddd($A_1C_1A_2C_2$, SiSiiSiSii): add packed four 32-bit words parallelly.

step7:
Step 7 is a permutation (A,B,C,D) = (B,C,D,A), for our two blocks, that is:
$A_1C_1A_2C_2$ = $B_1D_1B_2D_2$, $B_1D_1B_2D_2$ = $C_1A_1C_2A_2$

movdqa(temp3, $B_1D_1B_2D_2$)
$B_1D_1B_2D_2$= pshufd($A_1C_1A_2C_2$, 0xb1)
$A_1C_1A_2C_2$= temp3

step8 and step9:
pshufd(S[2r+2], S[2r+3], 0x44) —> $S_{2r+2}S_{2r+3}S_{2r+2}S_{2r+3}$: pack the last two words of round key array S[2r+2] and S[2r+3] into XMM registers;
paddd($B_1D_1B_2D_2$, $S_{2r+2}S_{2r+3}S_{2r+2}S_{2r+3}$): add packed four 32-bit integer.

3.3. Instruction Level Parallelism

Though we process the encryption arithmetic as parallelly as possible, there's still space to improve. The Pentium processor allows to execute two instructions in parallel through two five-stage pipelines, called U pipeline and V pipeline. This instruction pairing is carried out automatically and independently for programmers. However, that does not mean CPU will do it all. Instead , to pair two instructions, there's several constraints are to be satisfied. So it's up to programmers to organize the instructions elegantly, so that a better performance due to the superscalar structure is achieved. The following is pairing rules for Intel's SIMD instructions:

- The second instruction does not read or write a register which the first instruction writes to.
- MMX/SSE shift, pack or unpack instructions can execute in either pipe but cannot pair with other MMX/SSE shift, pack or unpack instructions.
- MMX/SSE multiply instructions can execute in either pipe but cannot pair with other MMX/SSE multiply instructions. They take 3 clock cycles and the last 2 clock cycles can overlap with subsequent instructions.
- An MMX/SSE instruction which accesses memory or integer registers can execute only in the U-pipe and cannot pair with a non-MMX/SSE instruction.

Table 1. Instruction Level Optimization.

Instructions(S)	C	Instructions(R)	C
1 movdqa xmm6, xmm1	1	**1** movdqa xmm6, xmm1	1
2 pslld xmm6, 1	2	**7** movdqa xmm2, xmm1	P
3 paddd xmm6, iOnes	1	**2** paddd xmm6, xmm1	1
4 movdqa xmm7, xmm6	1	**8** psrldq xmm2, 4	2P
5 pmuludq xmm6, xmm1	3	**3** paddd xmm6, iOnes	1
6 pand xmm6, imusk	1	**4** movdqa xmm7, xmm6	1
7 movdqa xmm2, xmm1	1	**5** pmuludq xmm6, xmm1	3O
8 psrldq xmm2, 4	2	**9** psrldq xmm7, 4	2
9 psrldq xmm7, 4	2	**6** pand xmm6, imusk	1
Total**: 14 Cycles**		Total**: 9 Cycles**	

C: cycle, P: paired, O: overlapped with subsequent instruction
S: straightforward implementation, R: re-arranged implementation

We take a code segment in assembly to illustrate how to optimize the code. In **Table 1**, the left two columns are straightforward implementation, while the right two columns are optimized implementation after rearranging. The instructions are labeled by the number before them.

Let's take a closer scrutiny to the code.

Instruction 2 uses pslld to multiply four integer words. It can be replaced by paddd to achieve the same results. Since the latter consumes only 1 cycle but the former consumes 2, we save 1 cycle by doing so.

Multiply instructions are quite time-consuming. According to rule 3, after issuing a multiply instruction, wait until three clocks later before using the result. So the overall throughput can be optimized to 1. In this guideline, we place instruction 9 after 5. Since 9 does not use the register which 5 writes to, its two execution cycles can overlap exactly with the last two cycles of 5. This saves two cycles.

Besides, it's easy to notice that instructions 1-4 have high data dependency. We move 7 to line 2, 8 to line 4 to remove data dependency, so that 1 and 7, 2 and 8 can be paired and this saves two cycles in overall.

After the above optimization, we save 5 cycles out of 14 and achieve up to 56% enhancements in performance.

4. OUR RESULTS

We test the cycles and throughput on Intel Core2 Duo™ 3.0G PC, as **Table 2** illustrated.

Table 2. Throughput Comparison

RC6-32/20/26	Intel C++ Intrinsic	SSE2-Assembly	Optimized SSE2-Assembly
Cycles	618/block	309/block	274/block
Throughput	621.4M bits/s	1242.7M bits/s	1401.4M bits/s

5. CONCLUSIONS

The goal of our work is to show how to speed up the RC6 block cipher utilizing Intel's SSE2 128-bit SIMD Integer Instructions. The case study of RC6 provides a way to other block ciphers. By way of three levels of parallelism(block level parallelism, operation level parallelism and instruction level parallelism), this goal has been successfully met.

6. FUTURE WORK

The optimization of RC6 that we have followed can be taken even further.

(1) The data movement between CPU and memory can be speeded up by using data prefetching. In this way the data required next can be brought into the cache in advance.

(2) The data can be aligned to 16 bytes prior to loading it into the processor. This increases the memory performance enhancement leading to a further improvement in speed.

(3) GPU offers a tremendous amount of computational bandwidth that was until now largely unusable for cryptographic computations due to a lack of integer arithmetic and user-friendly APIs. The latest generation of GPUs, which introduced integer/binary arithmetic, has been leveraged to create several implement. Two mainstream GPU camps, NVIDIA and ATI, both introduced their general-purpose GPU computing model, CUDA and ATI Stream respectively. They are designed to handle processing massive amounts of data efficiently and hold the potential for significant speedup for data parallel problems, basically for 5-10 times faster. Exploiting GPU parallel computing technology to cryptographic applications may be rewarding.

REFERENCES

[1] Ronald.L. Rivest, M. J. B. Rabshaw, R. Sidney, and Y. L. Yin, The RC6 Block Cipher, in First Advanced Encryption Standard (AES) Conference, Available at ftp://ftp.rsasecurity.com/pub/rsalabs/rc6/rc6v11.pdf

[2] A. Bosselaers, R. Govaerts, J. Vandewalle, Fast Hashing on the Pentium, Advances in Cryptology, Proceedings Crypto' 96, Lecture Notes in Computer Science 1109, N. Koblitz, Ed.,Springer-Verlag, 1996, pp. 298-312.

[3] Helger Lipmaa, IDEA: A Cipher for Multimedia Architectures? Selected Areas in Cryptography, Proceedings SAC' 98, Lecture Notes in Computer Science 1556, Stafford E. Tavares, Henk Meijer, Ed.,Springer-Verlag, 1998, pp. 248-263.

[4] Stefan Tillich, Johann Großschadl, and Alexander Szekely, An Instruction Set Extension for Fast and Memory Efficient AES Implementation, Communications and Multimedia Security, Lecture Notes in Computer Science 3677, Springer-Verlag, 2005, pp. 11-21.

[5] S. Bartolini, I. Branovic, R. Giorgi, E. Martinelli, Effects of Instruction-set Extensions on an Embedded Processor: a Case Study on Elliptic Curve Cryptography over GF(2/sup m/), IEEE Trans. Computers, ISSN:0018-9340, vol. 57, no. 5, M 2008, pp. 672-685.

[6] Craig S.K. Clapp, Instruction-level Parallelism in AES Candidates, in Second Advanced Encryption Standard (AES) Conference. Available at http://csrc.nist.gov/archive/aes/round1/conf2/Clapp.pdf

[7] http://www.agner.org/optimize/

[8] Jason C. Yang, Symmetric Key Cryptography on Modern Graphics Hardware, Available at http://ati.amd.com/technology/streamcomputing/.

[9] Intel(R) C++ Compiler User and Reference Guide.

[10] Intel® 64 and IA-32 Architectures Software Developer’s Manual.

Hu Qing is a postgraduate of ontoweb laboratory in College of Computer Science and Technology of Wuhan University of Science and Technology. She obtained her bachelor's degree in Computer Science from Wuhan University of Science and Technology in 2007.Her research interests are mainly on parallel computing and semantic web.

THE RESEARCH OF TORSIONAL SPRING MEASURING ON CONTROL SYSTEM *

Jianjun Ding
Physics and information school, Jianghan University
Wuhan, Hubei Province, China 430056
Email: hbddjjd@163.com

ABSTRACT

The article briefly describes MCU control examination system for one kind of new torsional spring testing machine: its development of software and hardware. this detector set uses the serial port for communication and programming control. Sensor signal pre-processing and connection with main chip are described in detail, and the software design process is also intruced.

Keywords: Torsional Spring, Measuring, MCU, Control

1. INTRODUCTION

Technics of manufacture equipment and material of spring in these years have a rapid development in spring industry. At the same time, quality of spring is demanded. Its parameters is demanded to be examined such as torsional moment, torsional angle and so on. Torque is the most common detection index, which shows that in the circumstances imposed by force, the angle which is produced of the spring, in order to control the strength which the automatic device needs, and does not have the oversized load to damage the torsional spring at the same time. So, to realize intellectualization of the spring measuring is becoming the spring industrial development need and the inevitable trend. Spring measuring system's design is relates to mechanical, electronic, pneumatic and other technical fields.

2. SYSTEM OVERVIEW

Torsional spring testing system realizes autofilter by measuring the torque and angle of the torsion spring, and then comparing it with pre-established indicators. Because of small size, poor rigidity, easy deformation, its parameters is measured difficulty. The traditional method of measuring can not meet the requirements, designing and developing a new type of test system have important significance to reduce the test error, ensure the accuracy and reliability of test.

In this paper, the design of test methods is based on the existing mechanical systems and related peripheral equipment. That single chip is the core device with auxiliary of corresponding electronic circuit realizes torsional spring automatic measurement. C8051F020 chip which have 12-bit ADC that satisfies the accuracy requirement serve as microprocessor-controlled. System hardware also includes signal conditioning circuits, liquid crystal display, and keys circuit and so on. The system has realized intellectualization and automation. The system obtains torsional spring's parameter of torque and angle by processing the sensor output signal.

* Sponsored by Hubei province natural science foundation (2007AB A302), Wuhan science & technology bureau (20085179952 4-06), Wuhan educational bureau (2008K057).

3. OVERALL SKETCH

Measuring the torque and angle of torsional spring, it is installed a spring measuring machine with two disk. Its one end is fixed an immovable disk that connects with the torque sensor, the other is fixed pivoted disk that connects with the electro-optical encoder. Electro-optical encoder outputs counter impulse that is related to the angle of the pivoted disk. The torque and the angle have a corresponding relation for a kind of spring. Outputted signal by the torque sensor and encoder is processed by MCU (Micro Controller Unit) processing sensor. The corresponding data and the final results is displayed on the LCD screen.

4. HARDWARE DESIGN

C8051F020 which is compatible with the MCS-51 is served as the core unit of hardware. The hardware also includes host computer minimum system unit, human-computer interaction unit, data and parameter storage unit, torque and angle measurement unit.

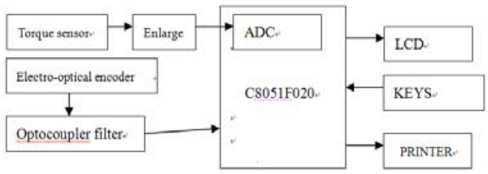


Figure 1. System Hardware Structure

4.1 C8051F020 Module

The C8051F020 devices are fully integrated mixed-signal System-on-a-Chip MCU with 64 digital I/O pins. It is high-Speed pipelined 8051-compatible CIP-51 microcontroller core (up to 25 MIPS), specific product feature selection and true 12-bit 100 ksps 8-channel ADC with PGA and analog multiplexer.

C8051F020 which is used in this design not only has a high speed, but also has a 12-bit 8-channel ADC. It is suitable to this system to realize the system function. Integration of module can make the hardware design easier. It makes the system more stable and reliable. (Fig.2)

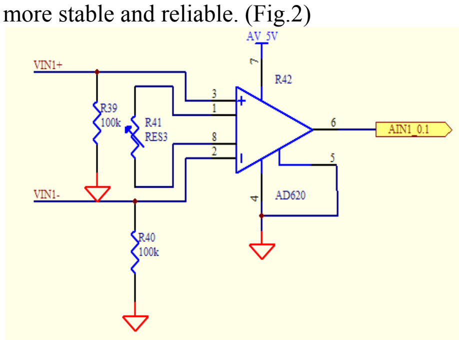


Figure 2. Hardware Connection

C8051F020 is connected to AD620 as a amplified chip, 6N138 as a photoelectric encoder, LCD and key. The weak signal come from sensor must be amplified in order to enhance its distinguish ability. The range of the voltage should be amplified to the ADC max measuring range for the sake of system precision.

AD620 is used as an amplified chip. The AD620 is a low cost, high accuracy instrumentation amplifier that requires only one external resistor to set gains of 1 to 1000.

The AD620 is a monolithic instrumentation amplifier based on a modification of the classic three op amp approach. Absolute value trimming allows the user to program gain accurately (to 0.15% at G = 100) with only one resistor.

Pins are seen in Fig.3. The one resistor connect to pin 1and 8 to adjust times of amplify. Pin 4 and 7 should connect positive and negative working voltage. Voltage input pin is 2 and 3, output pin is 7. Pin 5 is reference voltage. Formula 1 and 2 shows gains function of the chip. G is the gain multiple and RG is the resistance value.

$$G=49.4K\Omega/RG+1 \quad (1)$$

$$RG=49.4K\Omega/(G-1) \quad (2)$$

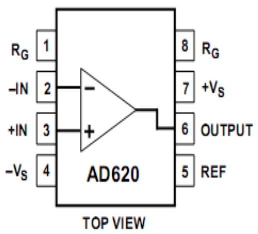


Figure 3. AD620 Pins

The AD620, with its high accuracy of 40 ppm maximum nonlinearity, low offset voltage of 50 µV max and offset drift of 0.6 µV/°C max, is ideal for use in precision data acquisition systems. We get the mV voltage from torque sensor in this system. if G=50, then RG=1KΩ.

4.2 Signal Isolation

Output of photoelectric encoder is pulse signal. Signal from photoelectric encoder is unilateralism. Input is separated from output in order to make the entire signal safe.

The 6N138/9 photoelectric encoders consist of an AlGaAs LED optically coupled to a high gain split darlington photodetector. The split darlington configuration separating the input photodiode and the first stage gain from the output transistor permits lower output saturation voltage and higher speed operation than possible with conventional darlington phototransistor photoelectric encoder. In the dual channel devices, an integrated emitter - base resistor provides superior stability over temperature. The combination of a very low input current of 0.5 mA and a high current transfer ratio of 2000% makes this family particularly useful for input interface to MOS,CMOS, LSTTL and EIA RS232C, while output compatibility is ensured to CMOS as ell as high fan-out TTL requirements.

4.3 Torque Sensor

It is very important to select a torque sensor that owns appropriate precision and measure range according to range, max, precision of spring torque, installation and others parameters. The role of static torque sensor transforms strain into electrical signal. The electrical signal is linear with torque. AKC-11which owned large range, high precision, stabilization, good manufacture technics is chosen.

4.4 Angle Sensor

Photoelectric encoder is an angle sensor for measuring the revolution angle of the spring. Photoelectric encoder is widely used to be the angle sensor. Its output is pulse. ZLG-2500-G-05E is chosen as a angle sensor.

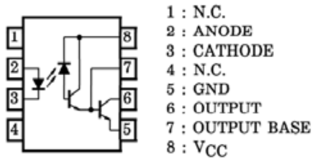


Figure 4. 6N138 Pins

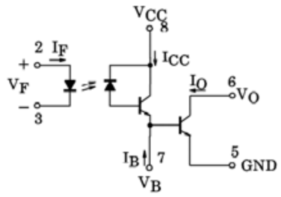


Figure 5. 6N138 Inside Structure

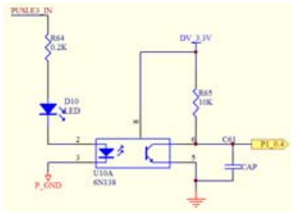


Figure 6. 6N138 Connect

4.5 Input/Output System

According to the characteristic of the system, output contains keys and LCD.

5. SOFTWARE DESIGN

The software design is programmed in C language in order to modify and maintain it easily. Structural and modular software design is easy to read, debug and function expansion. It includes main programmer, initialization programmer, keyboard scan programmer, LCD display programmer, signal acquisition programmer, signal store programmer. The main software chart and the measuring chart are below (Fig.7, Fig.8).

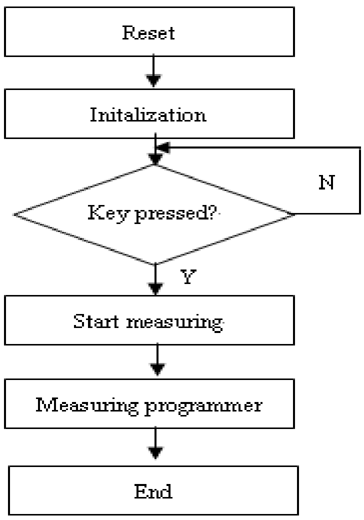


Figure 7. Main Programmer Chart

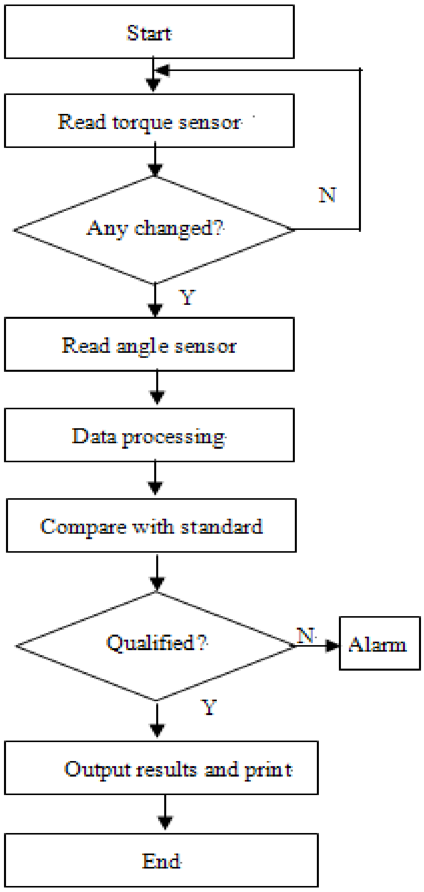


Figure 8. Measuring Programmer Chart

The aim of the system is to obtain the relation between the rotation angle and the torque. The angle of rotation can be obtained by output pulse of photoelectric encoder, and the torque can be obtained by output analog signal of the torque sensor. Different kind of have different relation.

6. CONCLUSIONS

This paper has presented a scheme for measuring torsion spring by the relationship between rotation angle and torque. The hardware and software have been designed. And its function is introduced. The results indicated that the measuring can effectively measure the rotation angle and torque, and then select qualified product.

REFERENCES

[1] M. I. Gurevich, O. V. Tel'kovskaya and D. A. Shkarovskii. Graphical Representation of the Initial Data and Computational Results for the MCU Computer Code. Atomic Energy, Volume 99, Number 3 2005,9 pp.597-601

[2] N. N. Ponomarev-Stepnoi, V. I. Bryzgalov, E. S. Glushkov, E. A. Gomin, M. I. Gurevich, V. E. Demin, G. V. Kompaniets, V. A. Lobyntsev, V. I. Nosov, D. N. Polyakov, O. N. Smirnovand O. V. Telkovskaya Using the MCU computer program to analyze the results of critical experiments with HTGR fuel pellets on the astra testing stand. Atomic Energy, Volume 97, Number 4 / 2004,10, pp.669-677

[3] Ying Zhang. Krishnendu Chakrabarty Adaptive Check pointing with Dynamic Voltage Scaling in Embedded Real-Time Systems. Embedded Software for SoC, 2004, pp.449-463

[4] Qiang Huang, Xiaofeng Liang and Yingying Zhu. Evaluation of Dependable Embedded System. Communications in Computer and Information Science, 2007, pp.614-622

[5] Susan Lysecky, Frank Vahid. Automated Application Specific Tuning of Parameterized Sensor-Based Embedded System Building Blocks. Lecture Notes in Computer Science, Volume 4206/2006 , pp.507-524

BASED ON LEAST SQUARES SUPPORT VECTOR MACHINE CLASSIFICATION OF THE CRANKSHAFT POSITION ALIGNMENT RESEARCH

Rui Zhang[1], Rongqiang Hu[1], Xiuzhi Yang[2]
[1.] School of Automation, Wuhan University of Technology, Wuhan
[2.] Materials Science and Engineering , Huazhong University of Science and Technology
Hubei, Wuhan, China
Email: whlgdxzr@126.com

ABSTRACT

It’s introduced to strengthen the crankshaft rolling and straightening technology, and studied the least squares support vector machine classification principle and multi-classification method using least squares support vector machine to establish the location of the crankshaft alignment classifier. And the establishment of the classification system is used in the crankshaft fillet rolling machine to strengthen alignment, and achieved good results.

Keywords: Crankshaft, Least Squares Support Vector Machines, Rolling, Alignment, Categories

1. INTRODUCTION

Nowadays crankshaft straightening process rolling is the crankshaft to improve fatigue strength of the technology of the most effective methods, the axis usually can have the distortion after the trundle.，and if the nonstraightness crank is big, it needs align. Overseas uses the trundle alignment craft generally from 1980s, but its core technologies still was at the security condition to our country, our country is occupying to this aspect research starts the stage. The author participation the topic-based group grinds the domestic first crank fillet trundle alignment engine bed, already succeeded in the production practice the application.

The process of rolling a specific alignment is as follows: to strengthen the front of the rolling process, the pay-off will make the crank axle waiting for machining ready, as shown in Figure 1, Four trundle pliers divide to four cylinder crank 17 positions three times carry on the high pressure trundle strengthening, the crank various fillet place plastic deformation produces the remaining compressed stress level, strengthened the fatigue strength. Rolling makes the crankshaft bending deformation. as a result of material inhomogeneity, the size and direction of bending is a random variable. After rolling crank axle comes into the aligning process, the aligning process diagram as shown in Figure 2, it needs measurements of non-straightness, and the expert system determines whether qualified Crankshaft, and finds out the school, the alignment strength size, or the position which the unqualified crank appears. Finally we carry on the trundle alignment regarding the unqualified crank. Alignment process of the three key issues are as follows: (1) the need to accurately measure the non-straightness of the crankshaft, usually before the use of poor methods of testing; (2) the establishment of rolling pressure and the inverse model of deformation, deformation vector input, output roller Roll pressure and position; (3) As a result of the influence of cold hardening, it may produce with the original distortion direction opposite distortion, and need to use greater pressure roller alignment, so the expert system output is the incremental pressure roller. To ensure a higher efficiency and smaller axial elongation, the number of alignment is generally 0-2 times. In this article, we mainly use the least squares support vector institutions to built expert reasoning supporting alignment of the crankshaft position classifier.

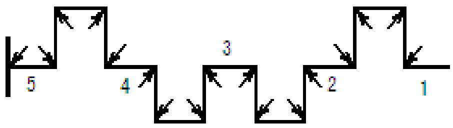


Figure 1. Schematic Diagram
of Four-cylinder Crankshaft Fillet Rolling

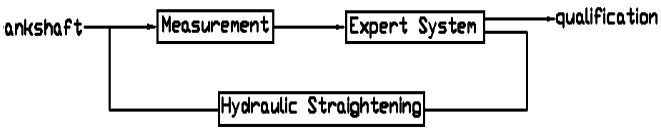


Figure 2. Crankshaft Straightening Rolling Flow Chart

2. LEAST SQUARES SUPPORT VECTOR MACHINE PRINCIPLE

Support Vector Machine (support vector machine, SVM), a kind of machine based on statistical study theory new machine learning method, is handed in by Vapnik [1].Principle is that the original sample space through the kernel function is mapped to the high-dimensional feature space, and that in high-dimensional space based on structural risk minimization ideas solve convex quadratic optimization problem, in order to get a better result in the limited training samples for the training study. As statistical study theory in structure risk minimum criterion concrete realization, the SVM has some advantages, such as the simple structure, the better overall situation, widely used, and so on. So it has been widely researched.

The least squares support vector machines (LS-SVM) proposed by Suykens [2,3] is an improvement of the SVM, which is the inequality in the traditional SVM replaced by equality constraints binding. And it will be the error sum of squares loss function as a training experience of loss set, so that solution of quadratic programming problems put into a system of linear equations to solve problems, to solve the problem to improve the speed and convergence precision, this method is applied to pattern recognition and nonlinear function estimation, achieved good results.

The key of the SVM’s use is the choice of $k(x_i, x_j)$ kernel. The nuclear function is the characteristic vector inner product, which is to meet the follow symmetric function: $K(x_i, x_j) = \langle \varphi(x_i \cdot x_j) \rangle = \langle \varphi(x_j \cdot x_i) \rangle = K(x_j, x_i)$

Commonly used kernel function that is the following four kinds: (1) linear nuclear $K(x,x_i)=x_i^T x$; (2) d-order polynomial kernel function $K(x,x_i)=(x_i^T x+1)^d$;(3) Radial Basis nuclear (RBF) $K(x,x_i)=\exp(-\frac{\|x-x_i\|^2}{2\sigma^2})$, etc. In this paper, to select the kernel function RBF. Supposes the training sample collection meet the follow function:

$\{x_k, y_k\}, k=1,2,...,N, x_k \in R^n, y_k \in R$

Which x_k is for the input, y_k is for the output, so the fitting question transforms as the minimum problem, such as formula (1) mentions:

$$\min_{w,b,e} J(w,e)=\frac{1}{2}w^T w+\frac{1}{2}\gamma\sum_{k=1}^{N}e_k^2, \gamma>0 \tag{1}$$

Which γ is for the penalty coefficient, taking the greater value generally; The constraints is $y_k = w^T\varphi(x_k)+b+e_k$, which $\varphi(x)$ for the Kernel space mapping function. B is for the deviation, e_k is for Fitting error. Definition minimum problem Lagrange function is for Type (2):

$$L(w,b,e,\alpha)=J(w,e)-\sum_{k=1}^{N}\alpha_k\{w^T\varphi(x_k)+b+e_k-y_k\} \tag{2}$$

α_k for the Lagrange multipliers

For L on w,b,e,α and the partial derivatives equal to zero, Derivation step as follow:

$$\frac{\partial L}{\partial w}=0 \Rightarrow w=\sum_{k=1}^{N}\alpha_k\varphi(x_k)$$

$$\frac{\partial L}{\partial b}=0 \Rightarrow \sum_{k=1}^{N}\alpha_k=0$$

$$\frac{\partial L}{\partial e_k}=0 \Rightarrow \alpha_k=\gamma e_k$$

$$\frac{\partial L}{\partial \alpha_k}=0 \Rightarrow w^T\varphi(x_k)+b+e_k-y_k=0$$

Eliminate variable w,e, get parameters b,α the analytical solution for Type (3):

$$\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 & l^T \\ l & \Omega+\gamma^{-1}I \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{3}$$

Which $y=[y_1,y_2,\cdots,y_N]^T$, $l=[1,1,\cdots,1]^T$

$\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_N]^T$ Ω is a symmetric matrix.Which the first i elements in row j column for $\Omega_{ji}=K(x_i,x_j)=\Omega_{ij}$

So the least squares support vector machines achieves the classification function as type (4).

$$\hat{y}=\sum_{k=1}^{N}\alpha_k K(x,x_k)+b \tag{4}$$

Seeing from the model, LSSVM classified geometry significance is uses two kind of samples in the characteristic space two parallel planoids to be smallest first two rides approaches, but causes these two planoids the average gap to be biggest.

3. SUPPORT VECTOR MACHINES ALGORITHM FOR MULTI-CATEGORY

In the view of the functional point on algorithm, the SVM is essentially a two types of classifier, and can not be directly used to solve the multiclass classification problems. Therefore, how to develop the basic types of support vector machine (Binary-classSVM, BSVM) extended into the multi-class support vector machine (Multi-class SVM, MSVM) becomes a serious problem. At present, there are three common methods ,that is "a class of congruence class", "into categories" and directed acyclic graph support vector machine [5]. In this paper, we use a class of congruence class (One Against Rest, 1-ar) methods constitute a crankshaft position alignment of multiple classifiers.

Categories for a sample of M data, 1-ar classification is to structure SVM as much as M. No. i in the structure of SVM sub-classifier, it will be i belonging to the first class of the sample data are marked as category, the other types of sample data for the negative category tags, the use of BSVM method in the first i-type and the remaining M-1 category construction of M between a separating hyperplane, then any sample of x for classification of the discriminant method is to sample x, respectively, for classification into the above-mentioned decision-making function of M, with the largest function value of that type, that is, for classification of samples x categories. This method is simple, classification shorter. There are functional (5) exist.

$$\max \hat{y}_i=\sum_{k=1}^{N}\alpha_k K(x,x_k)+b, i=1,2,...,M \tag{5}$$

RBF (σ) lected kernel function equals1 and a larger penalty coefficient γ equals 10000.Calculated parameters of the analytical solution b,α . Obtain the decision-making function classification 11, as shown in table 2 of the test samples will be output to structure a good classification of 11 decision-making function, get the maximum value, maximum value is the category where the category of the sample. Comparison of Tables 1 and 2 shows least squares support vector machine classifier with good generalization performance.

4. CONSTRUCTION OF THE CRANKSHAFT POSITION ALIGNMENT CLASSIFIER

Four-cylinder crankshaft alignment may be the main axis of the neck and five Crankpin 4 Department for the over-round rolling alignment, alignment of the crankshaft before the bending deformation of the shape of common S-type and arch type. 5 spindle necks to the vertical displacement sensor 5 is installed, measurement has been poor before, before the amplitude and phase difference as classifier input.

Before poor - classification of training samples alignment example as shown in table 1, I will be the overall effective sample data 33 is divided into 11 categories. Table 1 lists the S-type and a class of arch-type of the typical data, which is the arch-type samples 1,2,3,4,5,6 is the S-type.

5. CONCLUSIONS

Least squares support vector machines with the greatest access to the average interval separating hyperplane, the method and combination of multiple classifiers can get global optimal solution of the classification decision-making function. I use this method to build the crankshaft position classifier alignment and strengthen the crankshaft fillet rolling machine successfully applied in practice.

Table 1. Fore Poor-classification of Training Samples Alignment Example

the number of Samples		Worse before					Category
		1 Spindle neck	2 Spindle neck	3 Spindle neck	4 Spindle neck	5 Spindle neck	
1	Amplitude	0.1	0.2	0.26	0.2	0.1	1
	Angle	0	0	0	0	0	
2	Amplitude	0.1	0.2	0.28	0.2	0.1	1
	Angle	0	0	0	0	0	
3	Amplitude	0.1	0.2	0.3	0.2	0.1	1
	Angle	0	0	0	0	0	
4	Amplitude	0.1	0.2	0.1	0.26	0.1	11
	Angle	0.5	0	0.5	0.5	0.5	
5	Amplitude	0.1	0.2	0.1	0.28	0.1	11
	Angle	0.5	0	0.5	0.5	0.5	
6	Amplitude	0.1	0.2	0.1	0.3	0.1	11
	Angle	0.5	0	0.5	0.5	0.5	

Table 2. Before Bad-example of Alignment of the Test Samples

the number of Sample		Worse before					Decision-making function of the maximum	Category
		1 Spindle neck	2 Spindle neck	3 Spindle neck	4 Spindle neck	5 Spindle neck		
1	Amplitude	0.1	0.2	0.25	0.2	0.1	0.8268	1
	Angle	0	0	0	0	0		
2	Amplitude	0.1	0.2	0.27	0.2	0.1	0.9496	1
	Angle	0	0	0	0	0		
3	Amplitude	0.1	0.2	0.1	0.27	0.1	1.0001	11
	Angle	0.5	0	0.5	0.5	0.5		
4	Amplitude	0.1	0.2	0.1	0.29	0.1	1.0001	11
	Angle	0.5	0	0.5	0.5	0.5		

REFERENCES

[1] VapnikV.Statistical Learning Theory [M]. New York: John Wiley, 1998.

[2] SuykensJAK, Vandewalle J. Least squares support vector machine classifiers[J].Neural Processing Let-ters,1999,9(3):293-300.

[3] Suykens J A K, Lukas L, Vandewalle J. Sparse approximation using least squares support vector machine [A].IEEE Int Symposium on Circuits and Systems[C]. Geneva, 2000.(Ⅱ):757-760.

[4] Du Zhe, Liu Sanyang. Least squares support vector machine algorithm variants [J]. Xidian University (Natural Science Edition), 2009 (36) 2:331-337.

[5] Fang. Support Vector Classification Algorithm for Multi-[J], Fujian PC, 2009 (4) 84-86.

DESIGN AND IMPLEMENTATION OF SYNCHRONIZED COLLABORATIVE SYSTEM UPON HETEROGENEOUS CAD SYSTEMS *

Xiaoyong Zhang, Wanfeng Dou, Xiaodong Song
School of Computer Science and Technology,
Jiangsu Research Center of Information Security & Confidential Engineering,
Key Lab of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University
Nanjing, Jiangsu 20097, China
Email: douwanfeng@njnu.edu.cn

ABSTRACT

The synchronized collaborative design within heterogeneous CAD systems becomes a significant challenge because of the great differences in the modeling operations and underground data structures between different CAD systems. This paper first proposes a collaborative architecture with heterogeneous CAD systems, which is composed of CAD Adapter component and CoCAD Middleware component. The CAD Adapter is responsible of capturing of local operation commands and replaying of remote operations, and the CoCAD Middleware takes charge of translation of operation commands between different CAD systems based on a language of neutral modeling command (NMC). The system design and prototype of synchronized collaborative design is discussed based on two common CAD systems Pro/E and UG with their APIs and VC++.

Keywords: Neutral modeling command, Heterogeneous CAD systems, CSCW, Collaborative mddleware.

1. INTRODUCTION

In recent years, more and more complex design need to be collaboratively developed by multiple departments or groups geographically dispersed. This new development paradigm requires new approaches and tools that effectively support collaborative design activities [1].

The current approaches of synchronized collaborative design are divided into types[1-2]: (1) Visualization design systems, which support the function of viewing, annotating, and inspecting design models in a web or a CAD environment, and (2) Co-design systems, which provide users the function of modeling and modifying models interactively and collaboratively online. The visualization-based CAD system needs to download the product data using files and a variety of 3D streaming-based communication methods for collaborative design [3]. Their work aims at supporting visualization of multiple CAD models in a distributed CAD environment.[4], authors developed a mechanism of updating facet models, where a change of a model at an interval is captured, encoded in an incremental editing manner, transmitted and embedded into the associated faceted models at remote sites. The co-design system usually can effectively support collaborative modeling and modifying design models among designers. Such systems can be divided into two types according to CAD systems used at different sites: homogeneous and heterogeneous. The homogeneous systems require to use the same CAD system at any collaborative sites, it means that users have to move from their accustomed design systems into the new systems, and some additional costs for this new system is also applied to enterprise, and even more users need to take great efforts to learn this new system.

Integrated with the Computer Aided Design (CAD) technology and Computer Supported Cooperative Work (CSCW) technology, the Collaborative CAD (CoCAD) supports designers from different sites to finish a product design task at the same time by means of networks. The CoCAD makes best use of remote resources and reduces the design cost, has made a rapid and further progress in the past decade. However, comparing with homogeneous systems, collaborative design based on heterogeneous CAD systems is much more welcome to most users because different designers in different companies or departments are often accustomed to distinct CAD systems, and they prefer to work on their familiar CAD systems. The synchronized collaborative design within heterogeneous CAD systems becomes a significant challenge because of the great difference between the modeling operations and underground data structures of different systems. The challenging problem with synchronous collaborative design based on heterogeneous CAD systems is how to effectively exchange CAD models or modeling operation between CAD systems in real-time regardless of the great difference between the data structures and operation commands of different CAD systems [5].

A feasible method to resolve the above problem is to directly transmit operation commands to reduce the transmitting data comparing the method of exchange complete solid models among collaborative design systems. However, different CAD systems have different operation command set and direct exchange is obvious impossible because it is required to build the mapping of operations between every two CAD systems. To solve these problems, it is required a neutral modeling command set to aid to realize the exchange of information of operations between different CAD systems. Chun et al. introduced the data mapping between different CAD systems and proposed the description, definition and recognition technology of operation semantics and mapping model based on ontology [6-7]. Since this method is based on Standard for the Exchange of Command in computer aided design (STEC), it doesn't implement real-time interaction between different CAD systems. Guk-Heon Choi et al. proposed a macro-parametric approach to exchanging CAD model between different CAD systems [6]. By analyzing the general commands of several Cad systems, they build a series of neutral commands to exchange the information of operations between different CAD systems based on the macro command files. This method focuses on off-line exchanging of the whole CAD model, not considering real-time exchanging of single operation command required by synchronized collaborative design. Considering the heterogeneity of CAD systems in distributed environments and macro-semantics feature, Yang J. et al proposed an approach to exchange operation commands based on macro semantics commands [8]. By analyzing the modeling commands and macro files of several commercial

* Supported by the Natural science Foundation for Higher School of Jiangsu Province of China under Grant Nos.07KJD520112.

CAD systems, the definitions of Macro Command Group (MCG) and Macro Semantic Command (MSC) are proposed, and a set of MSCs is set up. In this approach, the MCGs corresponding to the modeling operations of each CAD systems are exchanged by MSC in the form of XML. However, this method requires that users must abstract the recent operation command from macro files by interactive way. Min L. et al. proposed an approach to set up a number of neutral commands corresponding to the basic modeling operations of CAD systems [9-10]. Based on the neutral modeling command, data exchange between heterogeneous CAD systems is achieved.

2. COLLABORATION ARCHITECTURE

Heterogeneous collaborative CAD system supports a group of users dispersed from geographical position under the constraints of network environment and computation environment resources to simultaneously conduct design tasks. A replicated structure of collaborative CAD system is given in Figure 1. Each client side is running a specific single-user CAD system. Therefore, not only this structure may apply in the same type CAD application, also may apply heterogeneous CAD application. For instance, in collaborative CAD system each client site may be possible to use the different CAD application. Each collaborative site is deployed the corresponding CAD middleware. The role of the CoCAD middleware is responsible for the transfer of operation information, the mapping of operations from different CAD systems, concurrent control, and session management. Thus this structure may realize transparent access to the design view from the multiple CAD system at different sites. On the other hand, once a CAD middleware at a site breaks down, this cannot interfere with the entire collaborative task, hence can raise the whole system's robustness.

In order to process heterogeneous CAD application deployed at each client site to support the shared awareness during collaborative design, the client end must run a module, called CAD Adapter. The main role of the CAD Adapter is first to capture different kinds of events or operations from specific CAD system deployed at each client site. Then the CAD Adapter carries on information filtration and accumulation according to operation sequence and distinguishes these operation semantics by means of the API functions provided by the specified CAD system. At the same time, the CAD Adapter also must transfer the local operation to the following CAD middleware, and receive and replay the operation of other sites from the corresponding CAD middleware deployed at those remote sites. Obviously, the CAD Adapter is close correlative to the concrete CAD application, namely it needs to understand the application characteristics or semantics.

The structure of CoCAD middleware is shown in Figure 1. Its functions include the following: (1) Operation transformation: it first receives a local operation from the upper CAD Adapter, then transforms it into the neutral operation according to the CAD application interoperability standard, again transfers it to the other cooperative sites. Operation transformation is the core part of this function of the CoCAD middleware. It implements two aspects of function. On the one hand it will transform the local operation into the neutral operation; on the other hand, the remote operation from other site is transformed into the specific CAD operation so as to be executed by local CAD system. It is obvious that there is a need of neutral operation set to finish the operation transformation. The neutral operation set will be discussed in next section. (2)Concurrent control: it check the relationship between multi-user concurrent operations from different sites according to the history operation buffer shown in Figure 1 and resolves the conflict of operations according to concurrent control strategies so as to preserve consistency of design views at different CAD systems.

The part of operation transformation is responsible for the exchange of operation information between different CAD systems. It includes two modules: NMC-to-SMC and SMC-to-NMC. The task of the NMC-to-SMC component is to map neutral modeling command (NMC) into system modeling command (SMC). The task of the SMC-to-NMC component is to map the SMC into NMC. As shown in Figure 1, to make NMC-to-SMC component and SMC-to-NMC component work properly, neutral modeling command set should be required to accomplish the transformation between CAD system commands. This problem will be discussed in the following section.

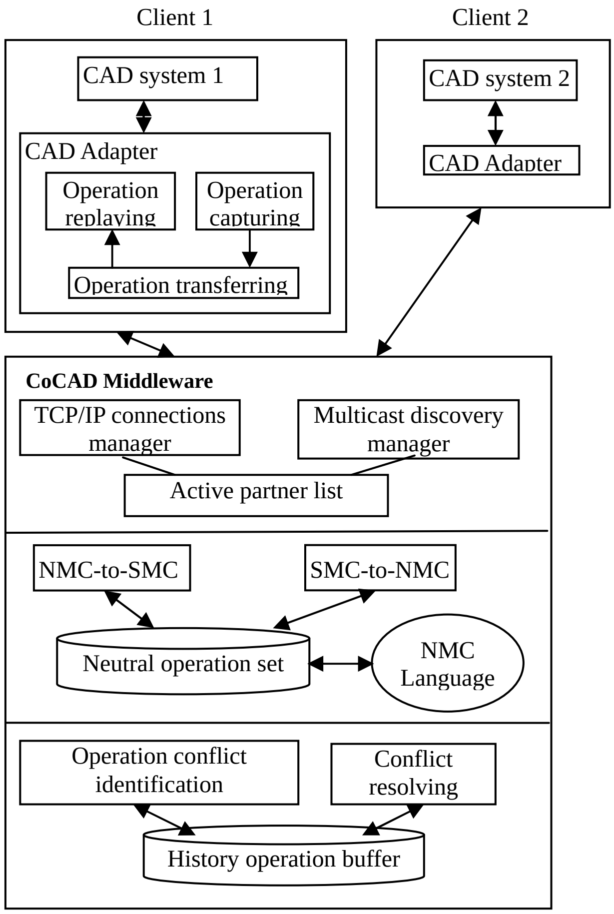


Figure 1. The Structure of Collaborative CAD Application with Heterogeneous CAD Systems

The data exchange upon heterogeneous CAD systems includes two main components, NMC-to-SMC and SMC-to-NMC. The tasks of NMC-to-SMC component is responsible for mapping of operation command from neutral modeling command (NMC) to system modeling command (SMC), and the task of SMC-to-NMC component is to transform a specified CAD system command into neutral modeling command according to transformation rules or vice versa. As shown in Figure 1, to make NMC-to-SMC component and SMC-to-NMC component work properly, a neutral modeling command set must be required to accomplish the transformation between CAD system commands. Obviously, a description language of neutral modeling command is aid to fulfill the transformation based on the neutral modeling command set. Their roles are to bridge between different CAD systems and to support the mapping of operation command.

NMC set plays an important role in achieving real-time operation command exchange within heterogeneous CAD systems to support collaboration among designers. To guarantee the rationality and validity of the constructed NMC set, parametric feature modeling can effectively support product modeling with parametric features, variable design and intelligent design. Hence, parametric feature modeling operations is the base to construct NMC.

We construct NMC set following the principles of parametric feature modeling. The NMC set proposed in this paper includes nineteen basic features as basic operation command set according to basic geometry feature, aided geometry feature and reference feature. We observe that the essential modeling operations provided by all commercial CAD systems are similar though their parameters of the corresponding operations may be different. To ensure that every SMC can be translated into a NMC easily or vice versa, it is desirable to make NMC correspond to the intersection of parametric feature modeling operations.

To effectively support the realization of two conversions, SMC-to-NMC and NMC-to-SMC, it is necessary that a suitable representation language of neutral modeling command be built so that the conversion is easily completed. The NMC set built by parametric modeling feature is a minimum command set. For each operation command in NMC set, its parameters take the union of parameters of operation command from all CAD systems. Taking extrusion as an example, there is a bi-extrusion attribute in SolidWorks and UG while MDT and Invertor only support single-extrusion, so the parameters of NMC Extrusion will include bi-extrusion parameter.

3. SYSTEM DESIGN

To effectively support synchronized collaboration between different CAD systems, we adopt component techniques and design patterns to fulfill the design of CADAdapter and CAD middleware respectively.

3.1 The Class Structure of CAD Adapter

For the sake of support the interoperation of different CAD systems, an interface ICADAdapter is first defined, it provides two interfaces: captureOpr() and replayOper(). The role of the captureOper() is to capture the commands from a specific CAD system and to assemble them into a CAD operation. The role of the replayOper() is to replay or execute an operation from remote sites on the local CAD system. Figure 2 shows the structure of CAD Adapter by means of UML. The CADAdpter class is abstract class in terms of the ICADAdapter interface, and it is associated with Operation class and OperTransfer class to implement the transferring of operations. Three subclasses,UGAdapter, Pro/EAdapter, and SolidWorksAdapter inherit the CADAdapter to implement the specific tasks.

3.2 The Class Structure of CAD Middleware

The CAD Middleware is relatively independent component, and may be deployed flexibly and easily. It tasks includes operation transferring, operation transformation, group management, and concurrency control. The class structure of CAD Middleware is shown in Figure 3.

The CADMiddleware requires NMC set to aid to accomplish the translation of operation commands. This is done by CommandTranslator. The CommandTranslator provides three methods: SmcToNmc(), NmcToSmc(), and checkIntegrality(). The SmcToNmc() implements the translation of a system operation command to a neutral modeling command, and the NmcToSmc() is a neutral modeling command to a system operation command, and the checkIntegrality() checks the integrality of the translation. Here, we define a Rule interface. It is used to resolve the differences of translation of different features for different CAD systems.

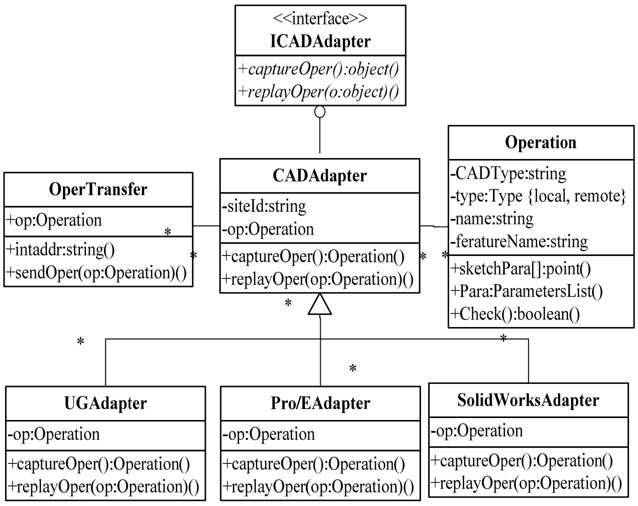


Figure 2. The Class Structure of CADAdapter

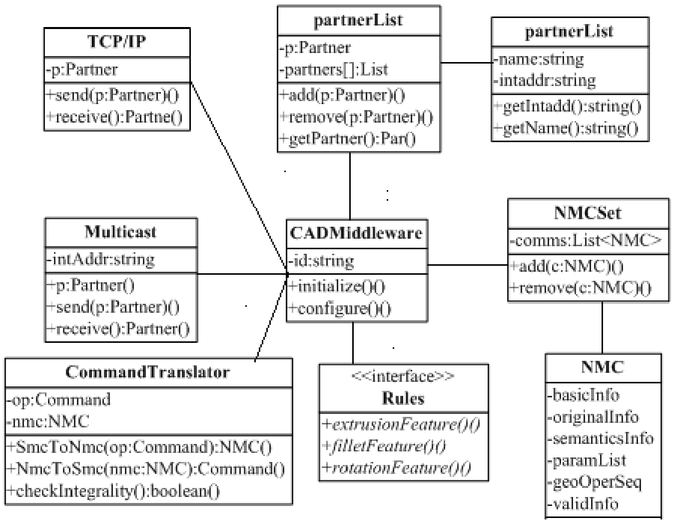


Figure 3. The Class Structure of CADMiddleware

For the translation of a SMC into a NMC, namely translator of SMC-to-NMC, there are three situations: direct match and indirect match. Direct match denotes that a SMC has the same semantics with a NMC in NMC set even more the same name of operation command. Such translation is one-to-one, and all parameters of the NMC may directly be gained from the SMC. Indirect match denotes that a SMC has no same semantics but the same feature modeling functionality with any NMCs in NMC set. Such translation is completed by translating the SMC into a corresponding NMC with the same feature modeling function, but the original information of the SMC including command name and CAD type should be saved in the originalInfo part in NMC representing language so that it is used during the translation of this NMC into a SMC. And all parameters are gained by direct parameters and indirect parameters. Here direct parameters refer to those that are also the parameters of the SMC and hence can be obtained from the

SMC directly. In contrary to the direct parameters, the indirect parameters are calculated based on the feature mode.

For the translation of a NMC into a SMC, namely the translator of NMC-to-SMC, we can find a match relationship between a NMC and the SMC. Since the NMC set in our paper takes the intersection of SMCs of all CAD systems, a NMC must find a corresponding SMC or several SMCs. In order to avoid information losing during the transformation, and to improve the efficiency of transformation, we need to find the original semantics in originalInfo part of the NMC so as to complete direct transformation. For example, when the UpToSurface command in SolidWorks is executed at local site, then it must be translated into a NMC so as to send it to other remote sites. Since the UpToSurface command doesn't correspond to any NMCs in the NMC set directly, it must be translated into a NMC with the same function, say Extent, meanwhile UpToSurface and extrusion to a surface are saved respectively in originalInfo and semanticsInfo of the NMC. Before this NMC is executed in Inventor at remote site, it must be translated into a SMC like DistanceExtent of Inventor system, but from the viewpoint of semantics it is more suitable to translate it into ToFaceExtent because there is directly corresponding relationship between them.

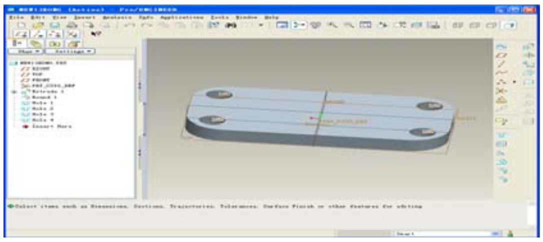
4. IMPLEMENTATION

We have developed a preliminary system of synchronized collaborative design based on UG and Pro/E. The prototype is implemented with Visual C++ 6.0 and APIs of the CAD system. The CAD Adaptor is compiled into the plug-in of native CAD system, running as background process after system starts to work. The CoCAD middleware includes both SMO-to-NMC and NMC-to-SMC translators implemented with Visual C++ 6.0. The CoCAD middleware is independent of any CAD systems, hence is implemented as a sole process. The communication between the CoCAD middleware and the CAD Adaptor is achieved by named pipes, and the communication between the CoCAD middleware deployed in different sites is achieved by Transmission Control Protocol (TCP).

In the developed prototype system, two geographically dispersed users, using Pro/E and UG respectively, completed the part design shown in Figure 4. One user uses Pro/E and the other uses UG. The process of the collaborative design is as follows.

When the user using Pro/E creates one base extrusion feature in his or her site, of the moment, the operation capturing of the local CADAdaptor resided in Pro/E is notified by the modelChanged event fired by Pro/E API. And then it gets the new created feature and captures the geometric information of the sketch and sketch segments, the values of dimensions and parametric information of the extrusion feature. Finally, the operation transferring of the local CADAdaptor sends them to the local CoCAD middleware. After the operation transferring of the local CoCAD middleware received the feature command and related parametric information, the operation transformation of the local CoCAD middleware is responsible for the translation of the local operation into the neutral modeling command. This task is completed by SMC-to-NMC module in operation transformation. Then the neutral modeling command transformed is sent to the CoCAD Middleware running at the remote site.

Further, the user using the CoCAD Middleware in UG at another site receives the neutral modeling operation, then is translated into system modeling command suitable to UG, and sent to the CADAdaptor resided in UG. Finally, this operation command is delivered to UG system to be executed. Similarly, one fillet and four through holes are completed at the two sites respectively, and the views of the two systems are shown as Fig4



a) Pro/E view

b) UG view

Figure 4. The Views Of The Two Systems After One Base Extrusion, One Fillet, and Four Through Holes Are Created

5. CONCLUSION AND FUTURE WORK

Synchronized collaborative design based on heterogeneous CAD systems may effectively supports design tasks among designers at different sites by means of Internet, and has been paid more and more attentions in recent years. But due to the differences in both data structure and modeling commands between different CAD systems, the real-time exchange of CAD operation commands or models is very difficult.

In this paper, a replicated architecture on synchronized collaborative design is presented and the components composed of it are analyzed in detail. The class structure of system design is described in UML language. A prototype of collaborative architecture with heterogeneous CAD systems is discussed.

Future work will focus on the mechanism of mapping of operations between different CAD systems, and on concurrency controlling and conflict resolving mechanism while multiple operations produced by different sites are executed concurrently.

REFERENCES

[1] Li,M., Gao, S.M. and Wong, C.L., Real-time collaborative design with heterogeneous CAD systems based on neutral modeling commands. International Journal of Computing and Information Science in Engineering, 2007, vol.7, pp.113-125

[2] Li,W.D., Lu, W.F., Fuh, J.Y.H. and Wong, Y.S., Collaborative computer-aided design—research and development status. International Journal of Computer Aided Design, 2005, vol.37, no.9, pp.931-940

[3] Qiu, Z.M., Wong, Y.S. and et al. Geometric model simplification for distributed CAD, International Journal of Computer Aided Design, 2004, vol.36, no.9,

pp.809-819

[4] Wu, D. and Sarma, R. The incremental editing of faceted models in an integrated design environment, International Journal of Computer Aided Design, 2004, vol.36, no.9, pp.821-833

[5] Li M, Gao S M, Li J. A preliminary study on synchronized collaborative design within heterogeneous CAD systems. In Proceedings of the 8th International Conference on Computer-Supported Collaborative Work in Design, Coventry, UK,2004, pp.1-6

[6] Chen C, Zhang S S, Chi Z, Wang C., An operation-based exchange mechanism in heterogeneous CAD collaboration. In Proceedings of the 9th International Conference on Computer-Supported Collaborative Work in Design, Coventry, UK,2005, pp.553-557

[7] Chen C, Zhang S,Li L, Zhang J Y. Real-time collaborative design based on operation semantic in distributed CAD environment. Chinese Journal of SouthEast University, 2004, 20(4), pp.442-447.

[8] Yang J, Dou W F. Synchronized collaborative design with heterogeneous CAD systems based on macro semantics commands. In Proceedings of the 12th International Conference on Computer-Supported Collaborative Work in Design, Xi'an China,2008, pp.183-188

[9] Li M, Gao S M, Li J, Yang Y D. An approach to supporting synchronized collaborative design within heterogeneous CAD systems. In Proceedings of ASME 2004 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Salt Lake, Utah USA,2004,pp.511-519

[10] Li M, Gao S M, Wang C L. Real-time collaborative design within heterogeneous CAD systems based on neutral modeling commands. Journal of Computing and Information Science in Engineering, 2007, 7(2),pp.113-125.

TAX REVENUE FORECASTING BASED ON BP NEURAL NETWORK

Cungen Shen [1,2] Kaijun Zhou[2] Wenzhen Zhang [2]
[1] School of Business, Hohai University,Nanjing 210098,China
[2] Party School of State Administration of Taxation, China
Email: yzscg@sohu.com

ABSTRACT

It is difficult to accurately analyze forecasting of tax revenue. Based on BP neural network, this thesis establishes a tax forecasting model which is characterized by its convergence speed, accuracy and generalization. The thesis proves that it is practical to analyze the highly nonlinear tax revenue forecasting through the neural network. It is of promotion value.

Keywords: tax revenue, forecasting, nonlinear, neural network.

1. INSTRUCTION

Growth and development of tax revenue can be influenced by various factors. On one hand, relationship between tax revenue and policies is very close, on the other hand, tax revenue grows with economic growth which includes GDP, fiscal expenditures, investment in fixed assets, sales of social products, volume of export and import, and total amount of wages, etc. It also subjects to some uncertain factors.

Technologies of statistics are available to forecast tax revenue, including the load model consisting of time series model and regression model. Time series model can not take full use of factors like GDP, leading inaccuracy and instability; regression model, where the above mentioned shortcoming is overcome, requires functional relationship between those factors and tax revenue itself, this is often quite difficult. It is known that MFNN (Multi-layer Feedforward Neural Network) is able to approach to discretionary nonlinear input-output relationship, so it is innovative to establish a nonlinear tax revenue forecasting model. This paper introduces such a model based on BP neural network and proves that this model should be applied widely thanks to its convergence speed, accuracy and generalization.

2. THE PRINCIPLES AND ALGORITHMS OF BP NEURAL NETWORK

2.1 The Principles of BP Neural Network

Artificial neural network is a nonlinear system to assimilate human information process and the interaction between real biological neural system and objects. It enables neural network to memorize and define through training before processing various kinds of information. It can learn, adapt, memorize, proceed in parallel and transfer in a nonlinear way, avoiding complicated math reasoning and ensuring stable output even samples are lacking and parameters are unstable. It is suitable to analyze and process issues containing very complicated information, inadequate background and unclear rule of reasoning.

As one of the artificial neural networks, Error Back Propagation (BP) is a Multi-layer Feedforward Neural Network (MFNN), and its neuronal transfer function is S type, output is the continuous quantity ranging from 0 to 1, so it is able to realize discretionary nonlinear mapping from input to output. Namely,

$f: R^n \rightarrow R^m$, $f(X)=Y$

Kolmogorov theory: a given continuous function

$f: U^n \rightarrow R^m, f(X)=Y$, U^n is unit solid body , f could be accurately realized by a three-layer MFNN, and this neural network has n processing units in its first layer, 2n+1 processing units in its middle layer and m processing units in its output layer.

2.2 The Algorithms of Three-Layer BP Neural Network

A typical BP network is a model consisting of three neuron layers, namely, input layer, output layer and middle hidden layer. Neurons at these three levels are in a full mesh but there's no link among neurons at a same layer. The input signals will be propagated to the neurons at hidden layer for function transforms before being propagated to the neurons at the output layer. Calculation of three-layer FMNN model is as follows [1]:

1) Initialization. Apply random value to all the joining weights and set an initial value for the threshold value.

2)Set the first learning data, namely, a pre-processed training sample set $\{x_{ml}\}$ and corresponding expected output set $\{y_{ml}\}$, m and l respectively means amount of samples and input vector.

3)Calculate O_{pi} of each layer.

For input layer, input is the same as the output: $O_{pi}=x_{pi}$, x_{pi} means the i^{th} value of sample P;

For hidden layer and output layer, neuron's output is characterized by:

$O_{pj}=f(\sum(w_{ji}O_{pi}-\theta_j))=f(net_{pj})$

O_p is current input of neuron j, and O_{pj} is its output, f(x) is a nonlinear and non-increasing sub-differentiable function, S type function and its derivatives are normally applied, namely, $f(x)=\frac{1}{1+e^{-bx}}$, b is a constant.

4)Calculate errors at each layer.

Output layer: $\delta_{pi}=(y_{pj}-O_{pj})(1-O_{pj})O_{pj}$

Hidden layer and input layer:

$\delta'_{pl}=O_{pl}(1-O_{pl})\sum_k \delta_{pk}w_{pk}$

5)Back Propagation, corrected weight and threshold value.

Output layer: $w_{ij}(t+1)=w_{ij}(t)+\delta_{pj}O_{pj}$, $\theta_{pj}(k+1)=\theta_{pj}(k)+\eta\delta_{pj}$

Hidden layer: $v_{lj}(t+1)=v_{lj}(t)+\delta'_{pl}O_{pl}$,

$\theta_{pl}(k+1)=\theta_{pl}(k)+\eta\delta'_{pl}$

η means learning speed.

6) Update learning data and reselect a model couple randomly before going back to step 2) until error function E of the whole network is lower than the preset error of fitting before finishing network training.

7) Return to 2) if learning data ends.

8) Update learning frequency for a new cycle of learning.

9) Return to 2) if learning frequency is not met.

3. TAX REVENUE FORECASTING MODELING BASED ON BP NEURAL NETWORK

3.1 Access to Data

Based on the influences imposed by relating factors and refer to reference [2], comparability of data as well as requests of the forecasting model, this thesis selects 8 indicators as the input variables of the neural network, namely, added value of primary industry, added value of secondary industry, added value of tertiary industry, total investment in fixed assets, total volume of import and export, total volume of fiscal expenditure, resident consumption level and total amount of money issued.

All the data, ranging from year 1994 to 2006, were from China Statistical yearbook 2007 and measured by billion yuan. Refer to Table 1.

3.2 Establishment of Neural Network Model

Fig.1 shows the tax revenue forecasting model based on BP neural network. Its input variables are the 8 indicators and the output layer neuron transfer function is S logarithmic function. Tax revenue of selected years is the target of network training. Under MATLAB environment, instruction for training the established neural network by using neural network toolkit function is[3]:

```
net=newff(minmax(P),[17,1],{'tansig','logsig'},'trainlm');
net=train(net,P,T);
```

{P,T} is the trained sample, trainlm is the selected Levenberg_Marquardt BP training function.

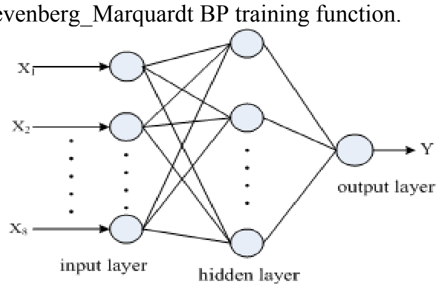


Figure 1. Neural Network Model

3.3 Pre-processing of Data

Data before 1994 is beyond consideration due to the major tax reform in 1994 which imposed great influences on tax revenue, so data of 1994-2006 is the sample. To test generalization of the established model, data of year 1999 and 2006 is selected in this paper to test generalization, leaving others as the training sample. Meanwhile, normalization is required and $\hat{x} = \dfrac{x - x_{\min}}{x_{\max} - x_{\min}}$ is adopted in this paper.

3.4 Training of the Network

Maximum training frequency in this paper is set as 2000 and expected error 1e-12, learning rate 0.1. Training accuracy would be realized and training ends within 1000 steps based on MATLAB programming. Amount of hidden layer neurons is crucial for convergence speed of network training, but experience is the best to rely on when selecting optimal value. This paper selects 13, 15,19, 25, 32 neurons respectively to train the network, proving that convergence speed is the fastest and error is minimum (lower than 5%) when amount of the hidden layer neuron is 15. For example, targeted data of tested sample is 12581.5 when tax revenue of year 2000 is the tested sample, and the actual value of the trained neuron is 12580.25, leaving a difference of 0.01%; and targeted data of tested sample is 34809.7when tax revenue of year 2000 is the tested sample with the actual value of the trained neuron being 33752.99, leaving a difference of 3%. This result proves that this network model is worth of applying to the tax revenue forecasting.

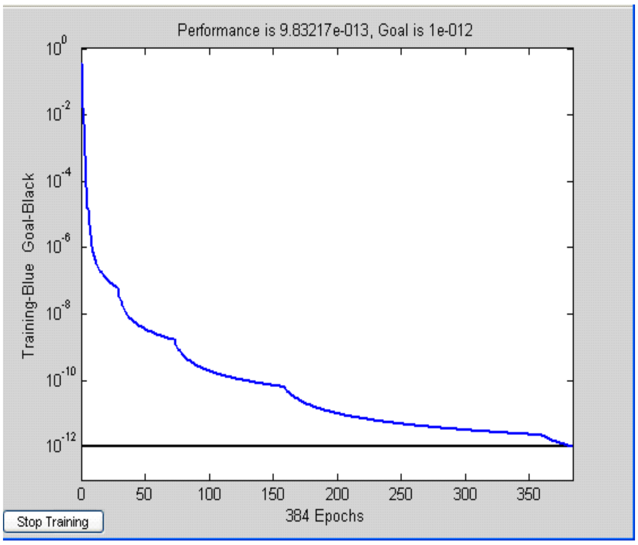


Figure 2. Convergence during training

4. CONCLUSIONS

A highly accurate and non-linear tax revenue forecasting model is established based on BP neural network in this paper. It has been proved that it is practical to apply neural network computing intelligence technique, a widely used one in physical science, to tax revenue forecasting modeling; amount of neurons at the hidden layer in the tax revenue forecasting model is crucial for convergence speed and should be optimized; it has been proved that interpolate value can be better tested compared with extrapolated value; to prevent over- adaption of network training where training result model fits the sample while other data witnesses more errors, it is practical to monitor network training by testing sample to avoid a situation in which difference between output error and targeted error is getting smaller when training sample is applied while getting bigger when test sample is applied. In fact, in-depth study is required on how to improve learning speed of network.

It should be made clear that the latest and adequate set of sample should be selected based on tax policy for the sake of important impacts imposed by tax policy and characteristics of neural network training.

REFERENCES

[1] Wang hui Assessment on Ability of Innovation of Hi-tech Enterprises Based on BP Neural Network. Science and Technology Management Research. Volume 11, 2007.

[2] Chang Qing. Multi-lateral Analysis on Tax revenue Based on Grey Correlation Analysis, Shandong Social Science. Volume 7, 2006.

[3] Feisi R&D Center of Technological Products, Neural Network Theory and MATLAB 7 Realization, China Publishing House of Electronics Industry.

[4] Xu Li, Intelligence Control and Intelligence System, China Machine Press.

Table 1. Sample Data of Tax Revenue Model

Year	Tax revenue	Added value of primary industry	Added value of secondary industry	Added value of tertiary industry	Total investment in fixed assets	Total volume of import and export	Total volume of fiscal expenditure	Resident consumption level	Total amount of money issued
1994	5126.9	9572.7	22445.4	16179.8	17042.9	20381.9	5792.6	1833	46923.5
1995	6038.0	12135.8	28679.5	19978.5	20019.3	23499.9	6823.7	2355	60750.5
1996	6909.8	14015.4	33835.0	23326.2	22974.0	24133.8	7937.6	2789	76094.9
1997	8234.0	14441.9	37543.0	26988.1	24941.1	26967.2	9233.6	3002	90995.3
1998	9262.8	14817.6	39004.2	30580.5	28406.2	26849.7	10798.2	3159	104498.5
1999	10682.6	14770.0	41033.6	33873.4	29854.7	29896.2	13187.7	3346	119897.9
2000	12581.5	14944.7	45555.9	38714.0	32917.7	39273.2	15886.5	3632	134610.4
2001	15301.4	15781.3	49512.3	44361.6	37213.5	42183.6	18902.6	3869	158301.9
2002	17636.5	16537.0	53896.8	49898.9	43499.9	51378.2	22053.2	4106	185007.0
2003	20017.3	17381.7	62436.3	56004.7	55566.6	70483.5	24650.0	4411	221222.8
2004	24165.7	21412.7	73904.3	64561.3	70477.4	95539.1	28486.9	4925	254107.0
2005	28778.5	23070.4	87364.6	73432.9	88773.6	116921.8	33930.3	5463	298755.7
2006	34809.7	24737.0	103162.0	82972.0	109998.2	140971.4	40422.7	6111	345603.6

THE METHOD OF TAX COLLECTION AND TECHNICAL REALIZATION BASED ON THE THIRD-PARTY ONLINE PAYMENT MODE

Qifeng Yang, Jun Wang, Ping Song, Xuan Ma
School of Economics, Wuhan University of Technology
Wuhan, China, 430070
Email: yangqifengwhut@163.com; wangjunwhut@163.com; songpingwhut@163.com

ABSTRACT

The third-party online payment has become the main payment modes of C2C/B2C e-commerce by virtue of its obvious advantages in the field of C2C/B2C, but the resulting tax issues also become an important issue faced by online payment. Combining with the realization of micro-payment mechanism of the online payment, the paper analyzes the reasons of the tax evasion of online payment. According to the characteristics of the third-party online payment mode, it proposes two methods of tax collection; designs the liquidation process, transaction process and database table of the tax collection that the Inland Revenue Department commissions the third-party online payment platform to do; and provides an available solution for the tax collection of the third-party online payment mode.

Keywords: The Third-Party Online Payment, Management of Tax Collection, Technical Realization

1. INTRODUCTION

The development of e-commerce makes tax evasion become easier [1]. (1)The characteristics of e-commerce make it hard to be monitored. The transnational nature of e-commerce makes the concept of national boundaries become increasingly blurred. Customers can stay at home to buy foreign products, e-commerce transactions do not need the fixed surface, paper-based contract and so on in traditional activities, and the products can be non-material category, such as music, books, software and other digital products. This has led to the loss of tariff and value-added tax [2]. (2)The law about e-commerce tax is imperfect; all countries don't have a common view about the tax problem of e-commerce. Because e-commerce in developed country is perfect, they advise some of e-commerce (C2C) not to be taxed [3]. But in our country the business of e-commerce should be taxed. (3)The administration of the tax of e-commerce is chaotic. Many dealers even don't register in the business sector and don't have business license either. This increases the difficulty of the work of anti-tax evasion. Some tax evaders preferred online payment, because it reduces their costs and risks of tax evasion[4]. (4)The methods of tax collection and regulatory technical fall behind and the tax management system fail to achieve network connection with e-commerce transactions and online payment system. A variety of e-commerce transaction data and bills are in electronic form, and the quantity is large. At present the tax department's capability of electronic data collection, processing and audit is weak, which leads to the loss of a large number of tax sources. What's more, in the present mode, the network among tax system, industry and commerce, public security, banking, customs and other departments is separate. It can not achieve seamless connection, and also can't control the sources of tax revenue and can't deal the revenue business processes fully automated effectively.

Realizing the actual funds transfer from the buyer to seller through the payment system in the process of e-commerce is an important symbol of whether the e-commerce occurs or not, and it is also the fundamental basis of tax collection. Online payment is the latest form of e-commerce financial payments, the research of the tax collection and management of online payment has important theoretical and practical significance.

Many scholars have already researched the e-commerce tax issues in our country. Zhou Feng, Wang Xiaolin, Zhou Tsang (2003), etc. proposed a method of tax collection which is based CA certification system and SET protocol[5].Gang Zaobin, Wang Yafang, Wei Dengwen, etc. (2004) built the model of process of tax in the environment of e-commerce[6]. According to the characteristics and problems of e-commerce, Cheng Dan(2003) designed a tax collection and management system model adapt to the requirements of e-commerce[7].These bring a lot of help to solve the tax problems in e-commerce. But the research of tax collection based on the third-party online payment mode is rare.

2. THE THIRD-PARTY ONLINE PAYMENT MODE

2.1 Topology of the Third-Party Online Payment Mode

The third-party online payment mode is the most active online payment mode at present. It is an honest broker for both transaction sides, its main purpose is to guarantee the credit of both sellers and buyers so as to reduce the uncertainty of transaction risk on internet, and increase the possibility of transaction.

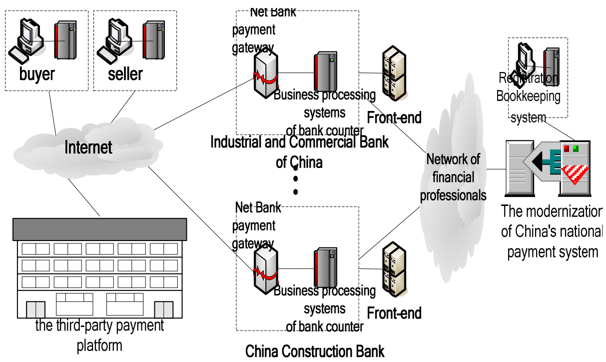


Figure2-1 Topology of the Third-Party Online Payment Mode

In Figure2-1 both transaction sides connect with the third-party payment platform through internet, then through commercial banking system connect to the finance network

(Industrial and Commercial Bank of China and China Construction Bank, are examples in Figure2-1),at last they complete online payment liquidation process in the modernization payment systems.

2.2 Process of the Third-Party Online Payment Mode

In this mode，payer and payee must apply for an account on the third-party payment platform firstly, at the same time they should provide their bank account number to the payment platform, and the payer must transfer the funds in real account to the account in payment platform. The specific process is in Figure 2-2:

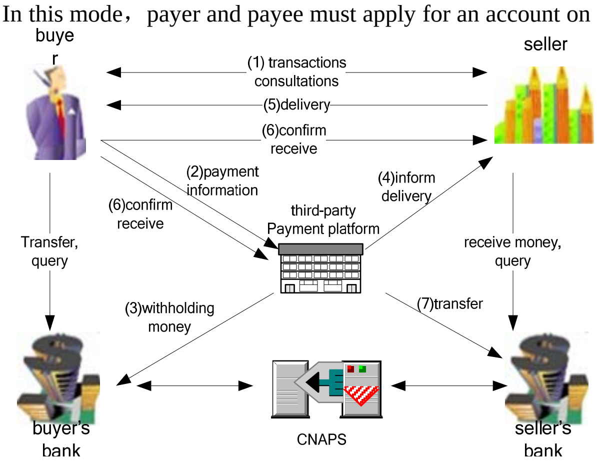


Figure 2-2 Process of the Third-Party Online Payment Mode

(1)The buyers search the website of sellers' and visit it, then select products and discuss price with the sellers. At the same time the buyers tell the goods information they want to buy to the sellers through their account number and get an order number.

(2)The buyers send their accounts and orders information to the third-party payment platform and link to the secure payment server directly, and then apply for the payment through different payment tools.

(3)The third-party payment platform checks the buyers' accounts consumption ability on the payment platform first. If they have enough money then deduct the money temporarily directly, if they don't have enough money, then the third-party payment platform informs the buyers to put money to their accounts and sends the payment information of the buyers on internet to the internet bank of the buyers according to the requirement of bank payment gateway technical.

(4)The third-party payment platform tells the results of payment to the sellers and authorizes them to deliver goods.

(5)The sellers provide service or goods to buyers.

(6)The buyers send message to the platform when they receive goods. If they haven't received products or the products have some quality problems they can apply for return back their money to the platform. When the platform gets the message of receiving products, it transfers the money from the accounts of buyers to the accounts of sellers.

(7)The banks implement the liquidation to the sellers through the third party payment platform, which is transfer money from the platform accounts to the bank accounts.

3. THE METHOD OF TAX COLLECTION OF THE THIRD-PARTY ONLINE PAYMENT MODE

The third-party online payment mode has promoted the development of e-commerce on a large extent by its significant advantages. The development of e-commerce has also promoted the development of international trade, so this generated much new tax revenue source. But because e-commerce has the characteristics of virtualization, paperless operation and quick sexuality, many transactions can be completed without leaving any traces. It opens the door to tax evasion which makes use of e-commerce, at the same time it brings great challenges to the traditional tax system, tax management models and anti-tax evasion. However, due to the third-party payment platform as an intermediate body between buyer and seller, all online payment transactions information can be concentrated to the third-party platform, which makes it possible for unified tax administration. Two methods of tax collection of the third-party online payment mode: bank withholding tax method, and the third-party withholding tax method.

3.1 Bank Withholding Tax Method

Bank withholding tax method is shown in Figure 3-1. The third-party online payment platform links to the Inland Revenue Department system. The third-party payment platform sends the relevant deal information to the Inland Revenue Department system when a transaction takes place. In the tax system engenders the tax bill according to different types of tax rates and volume of transactions. The tax system transfers tax bills to the banking system, and banking system withholds tax. After the banking system finish withholding tax, it sends the electronic tax documents to the third-party payment platform, as well as sends e-receipt to the tax system for retaining reconciliation. After the third-party payment platform receives the electronic tax documents, it sends to taxpayers, the entire tax process is completed.

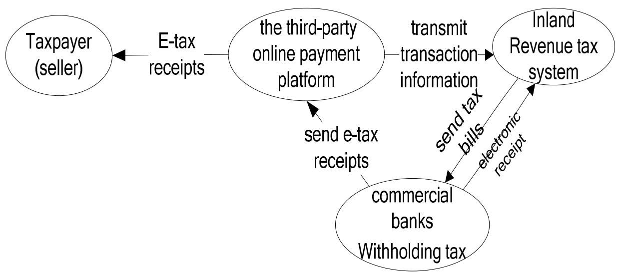


Figure 3-1 Bank Withholding Tax Method

3.2 The Third-Party Withholding Tax Method

The third-party withholding tax method is shown in Figure 3-2. According to the features of the third-party online payment mode: both transaction sides open virtual accounts, so tax system also sets up a virtual account on the third-party payment platform. This method is different from the first one; tax system commissions the third-party platform to withhold tax. In this method, after the buyer satisfies with the goods, it starts the linked deduction mechanism, first the third-party payment platform informs the taxpayer (the seller) to pay tax, and then the seller fills out tax return materials and sends to the third-party payment platform, the third-party platform implement of withholding tax. The tax is transferred from the seller's virtual account to the Inland Revenue Department's virtual account on the third-party payment platform. The Inland Revenue Department can transfer tax from the virtual account on third-party platform to the account in commercial banks on the basis of time or amount.

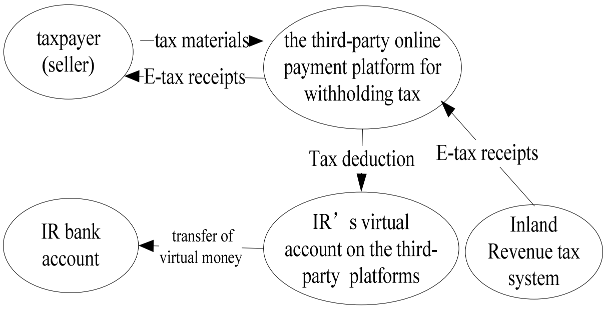


Figure 3-2 the Third-Party Withholding Tax Method

These two methods are equally effective in the conduct of tax collection of the third-party online payment mode. On the basis of the characteristics of China’s tax laws, the bank withholding tax method is mandatory implemented by the bank. But the current tax system is declaration tax system. In the third-party withholding tax method, first of all the taxpayers fill tax returns material, and then complete withholding tax by the third-party payment platform, this method reflects our current tax system. The paper designs the technical methods aim to the second method of tax collection of the third-party online payment mode.

4. TECHNOLOGY REALIZATION OF TAX COLLECTION OF THE THIRD-PARTY ONLINE PAYMENT MODE

4.1 The Liquidation of the Third-Party Online Payment Mode

In the third-party online payment mode, the cash flow of buyer, seller and the third-party payment platform is “secondary settlement”. When buyer and seller intent to deal online, the buyer will transfer money to account on the third-party payment platform through banks; when buyer and seller make a deal, the buyer confirms the information of receipt and passes it to the third-party payment platform, and then the third-party payment platform will transfer money to sellers’ account. This “secondary settlement” mode protects the security of the transaction amount and the interests of buyers and sellers.

In the process of the “secondary settlement” service, the third-party online payment is not only as the channel of linking each bank payment gateway, but also as a neutral third-party agency. It retains the effective transaction information of businesses and consumers, so that they can inquiry orders and other related information easily, especially when there are disputes between them, the information can be used as strong evidence of arbitration.

Assumption that buyer and seller open account in the same commercial bank C, and each bound with their virtual accounts on ZhiFuBao to carry out payment transactions. Inland Revenue Department commissions the third-party payment platforms to withhold tax including three stages in the entire process: pre-transaction phase, the transaction phase and post-transaction phase.

(1) Pre-transaction stag

When the buyer and the seller make a trading intent, the buyer confirms orders, and then transfers money from their bank accounts to their virtual accounts on the third-party platform. In the process of the buyer transferring funds, there is a liquidation of funds for commercial bank C. The accounting entry in commercial banks (the realization of the true meaning of the funds transfer) is:

Debit: buyer’s bank account M_1
Credit: the third-party platform’s bank deposit account M_1

At the same time, because the third-party platform and the banking system achieve the networking, payment instructions are passed to the third-party platform. So the accounting entry in the third-party platform is:

Debit: the third-party in the bank deposits account M_1
Credit: buyer’s virtual account M_1

(2) The transaction phase

After confirming business orders, the buyer sends the payment instruction to the third-party, then the third-party withholds the buyer’s fund m and notifies the seller’s to delivery goods. This process is considered as the first settlement. The accounting entry is:

Debit: buyer’s virtual account m
Credit: third-party’s current account m

After checking and accepting goods or services, the buyer notice the third-party payment platform whether he/her is satisfied with these goods or services. (If no objection is raised in agreed time, then the third-party will think that the buyer is satisfied with these goods or services) This process is considered as the second settlement. When the buyer is satisfied for goods or services, the accounting entry is:

Debit: the third-party’s current account m
Credit: seller’s virtual account m

When the buyer is not satisfied with goods or services, the third-party will refund the fund which it withholds. The accounting entry is:

Debit: third-party’s current account m
Credit: buyer’s virtual account m

If the buyer satisfies with the goods or services then start the linked withholding tax mechanism for deducting business tax, assuming that the tax rate is 1 percent, the accounting entry is:

Debit: seller’s virtual account 1% m
Credit: Inland Revenue Department’s virtual account 1%m

(3) Post-transaction phase

When buyer and seller in the transaction process don’t have dispute, the seller’s virtual accounts receive 99% m of the payment, the seller will transfer M_2 in virtual accounts to the account in commercial bank C, and at the same time the Inland Revenue Department will transfer M_3 to its account in commercial bank C.

In this process, the accounting entries in the third-party platform are:

Debit: seller’s virtual account M_2
Credit: The third-party’s bank account M_2
Debit: Inland Revenue Department’s virtual account M_3
Credit: Debit: Buyer’s bank account M_3

At the same time, there are related accounting entries liquidation of funds in the commercial bank C:

Debit: the third-party’s bank account M_2
Credit: seller’s bank account M_2
Debit: the third-party’s bank account M_3
Credit: Inland Revenue Department’s bank account M_3

At this moment, a payment transaction is finished.

4.2 The Design of Flow Chart and Database Tables of Tax Collection of the Third-Party Online Payment Mode

According to the transaction process and the way of liquidation of the third-party online payment mode, we design the flow diagram of tax collection that IRD commissions the third-party online payment platform to withhold tax. Figure 4-1.

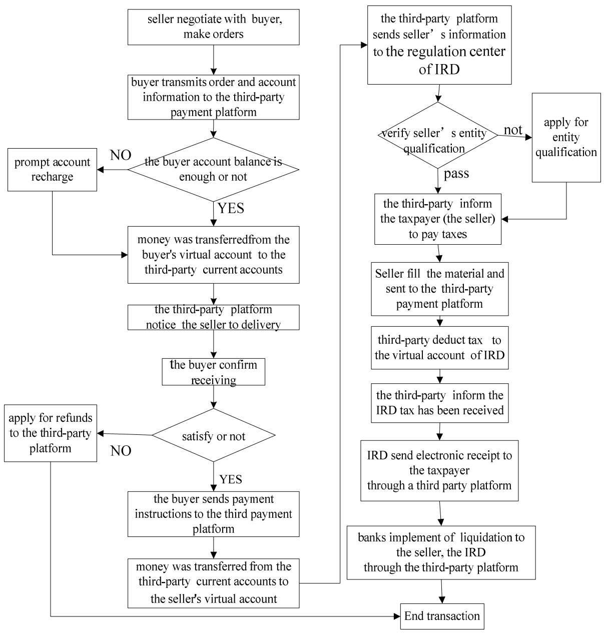


Figure 4-1 Tax Collection and Management Flow Chart of the Third-Party Online Payment Mode

It should create the underlying database in the process of transaction and tax collection of the third-party platform for recording the information of users, businesses, regulatory authorities and as a basis for tax collection. Database tables include the user's registration information, transaction information, and merchant's product information 、 registration information. Aim to the transactions process of the third-party payment platform, we design the main database tables: account master file table, transaction flow table, breakdown table, temporary table, tax collection information table and so on.

(1)account master file table (account number, account name, account type, account opening time, account balance, interest) is used to record the user's basic information, account number is primary key, account type is a personal account or business account, account balance is the amount on the platform;

(2)transaction flow table (flow number, buyer's account number, buyer name, seller's account number, seller name, transaction time, occurrence amount, order number) is used to record the transaction information of double accounts, the transaction occurred time and amount, generated order number, and it is the basis of breakdown table;

(3)breakdown table(account number, account name, order number, amount, serial number, sign of debit or credit) records the transaction specific information. Completing a transaction generates the breakdown of buyer and seller, the buyer and the seller can check their transactions over a period of time according to the breakdown table. The sign of debit or credit state decreases or increases of balance in corresponding account master file table, seller and buyer can check their account master file's balance easily; at the same time, when the buyer dissatisfies with the goods or service and applies to return money, it records the same amount in the opposite direction to clash account (to facilitate operation, 0 means the buyer and 1 means the seller);

(4)temporary table (account number, account name, recorded conditions, amount, time, sign of debit or credit)primarily saves transaction information between the seller delivery goods and the buyer send the satisfaction information, 0 and 1 means the recorded conditions. The recorded condition is 0 before the buyer send the satisfaction information to the third-party platform, and the recorded condition is 1 after the buyer send the satisfaction information to third-party platform or in prescriptive the time buyer doesn't apply return;

(5)tax collection information table (account number, account name, order number, transaction amount, time, tax rates, tax amount, and tax category) is mainly used to record information of tax collection.

Completing a transaction, the changes of database tables as follows:

(1)The buyer scans websites to select goods, and agrees price with the seller. The buyer uses his/her own account to send merchandise information to the seller, the seller determines the order information, and the buyer gains the order number;

(2)The buyer sends payment instructions and it records flow account to record the transaction serial number, the buyer and the seller's account number and name, the transaction time, transaction amount and order number;

(3)Accordance with the transaction flow, it amends the buyer's account master file table. The main amendment is the balance of buyer's account master file table;

(4)Recording breakdown. It records the buyer's transaction details, including account number, account name, order number, amount, and serial number, the direction of debit or credit. The direction of debit or credit is debit;

(5)Temporary table is used to registering the transaction. A deal generates two records in temporary table, for the third- party platform and the seller. It records transaction information in the temporary table, including account number, account name, amount, time and the direction of debit or credit is credit, at this time temporary table's of recorded condition is 0;

(6)The buyer satisfies with the goods and sends the receipt information to the third-party platform, so it modifies recorded condition (from 0 to 1).Then it modifies the balance of account master file table(here is the seller's balance of the account master file table) on the basis of the account number information and the direction of debit or credit in the temporary table. If the buyer doesn't satisfy with the goods and applies for return, the recorded condition doesn't change. At the same time it amends the buyer's breakdown, the buyer's breakdown records the same amount, but the direction of debit or credit is the credit, and amends the buyer's balance in the account master file table according to the breakdown;

(7)According to the seller's account master file table information, it records the seller's breakdown;

(8)It realizes the mechanism of withholding tax after the seller receives money. It records tax number (here is seller's account), tax name, order number, transaction amount, the tax rate, tax amount, time and revenue category in the tax collection table.

(9) It records IRD's breakdown in the breakdown table.

The whole transaction process and database tables process is shown in Figure 4-2:

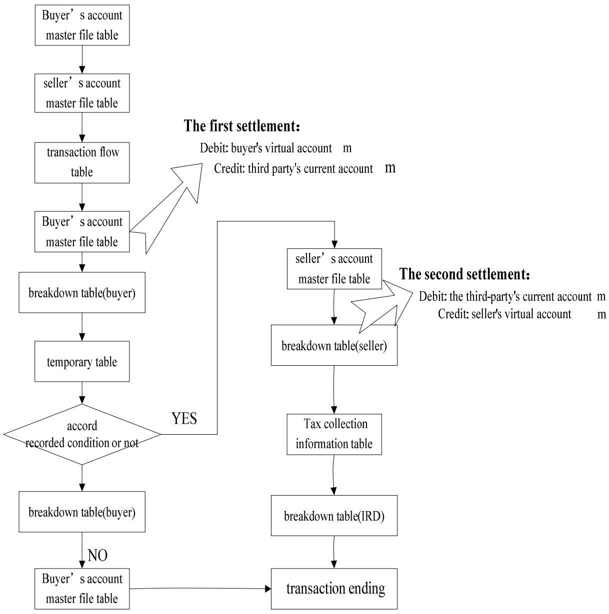


Figure 4-2 Database Flow Chart of Tax Collection and Management of the Third-Party Online Payment Mode

The method that Inland Revenue Department commissions the third-party online payment platform to withhold tax is independent withholding tax for each transaction of each seller. When too many transactions occur at the same time may cause tax system congestion, occupy amount of system resources and waste too much time, so the method can be further optimized. According to our current process of tax, the taxpayer doesn't tax for each transaction when it occurs, but tax in a fixed time (one month). According to this thinking, the method that IRD commissions the third-party platform to withhold tax can be improved; the third-party online payment platform can withholds tax according to the transaction amount per day of each seller at an agreed time. Therefore many withholding tax processes can be completed one time; so it saves much system resources and time.

5. CONCLUSIONS

The third-party online payment platform realizes logistics, business flow, information flow and capital flow to work together well by its own strengths. This saves cost for both transaction sides and protects their transaction information. It becomes the major mode of C2C online payment. But because the third-party online payment hasn't be classified to financial regulatory system and lacks of adequate legal basis for anti-tax evasion supervision, all above make the third-party online payment out of supervision for anti-tax evasion. This led to great economic loss of our country. So this article proposes two methods of tax collection based on the third-party online payment mode, and main aiming to the second method it establishes database tables to record the transaction information, and to realize of tax collection. At the same time, the mandatory feature of the tax can provide policy support for the first method of tax collection and improve the operability of this method. In the future, I will research of this method. With the e-commerce tax law establishing and perfect, in future we can implement preferential policies such as tax break or tax rebate according to the situation of sellers to encourage the development of online payment mode.

We can tax the transactions of third-party online payment mode to prevent false transactions of sellers. Some sellers often use false accounts to trade in order to get credit rating and the transactions log in reality. If it inducts tax mechanism, the false transactions will be taxed too, sellers will pay certain money for them. It can effectively prevent false transactions happening.

6. ACKNOWLEDGEMENT

This research was supported by the Ministry of Educational of the People's Republic of China in Humanities and Social Sciences Planning Fund: (1)Research on Leading Mode of Online Payment and Its Innovation Service, under Grant 07JA630028. (2)Research on Online Payment Collaborative Management Mode , under Grant 08JA630064.

REFERENCES

[1]Yu Dong. Golden Tax Project and e-commerce tax problems[J]. Information Technology and New Business,2004,10:52-55 (In Chinese)
[2]Zhang Fengling, Ceng Lei. Tax collection and management model in the environment of the E-business[J]. Business Modernization,2005,11:249-250 (In Chinese)
[3]Guan Xin.Tax policy of American's e-business[J]. Global Technology Economic Outlook,2000,7:8-9(In Chinese)
[4]Feng Bin. Research on Union Bank Online Payment Platform and Its Value-added Service. Master's thesis (In Chinese)
[5]Zhou Feng, Wang Xiaolin, Zeng Guangzhou. A Study Of Electronic Commerce Tax –Management System [J]. Journal of Shandong Normal University (Natural Science), 2003, 12:21-25(In Chinese)
[6]Gan Zaobin, Wang Yafang, Wei Dengwen. Petri Net-based Modeling of Taxation Business Process in E-Commerce[J]. Journal of Chang Sha University of electric Power(Natural Science),2004,11:59-62(In Chinese)
[7]Cheng Dan. Tax collection and management model of e-business [J].Tax Wide-angle,2003,6:38-39(In Chinese)

Yang Qifeng is a Full Professor, in school of economics, Wuhan University of Technology. He is a syndic of Hubei e-commerce institute, expert of e-government and e-commerce in development and reform commission of Hubei. He got the master degree from Hua Zhong University of Science and Technology in 1993, and got doctor degree from Wuhan University of Technology in 2006. He was a senior engineer in head office of China Construction Bank during 1993-2002, and took part in more than 10 large-scale project of computer application software development, and got 1 first prize, 2 second prize of Science & Technology progress in head office of CCB, and 1 second prize of Science & Technology progress in PBC. In recent years, he chairs or takes part in more than 10 items about Key Support Projects in the National Eleventh Five-year Plan Period, National Natural Science Foundation, State Ministry of Education Humanities and Social Sciences Planning Fund and so on. And he got 1 third prize of scientific and technological progress in Hubei Province, 2 third prize of scientific and technological progress in Wuhan City.

RESEARCH ON INDIVIDUAL REPUTATION MANAGEMENT BASED ON THIRD-PARTY ONLINE PAYMENT MODE

Ping Song, Yan Zhang, Qifeng Yang, Xuan Ma
School of Economics, Wuhan University of Technology
Wuhan, Hubei , 430070,China
Email: songpingwhut @163.com; zhangyanwhut2010@163.com

ABSTRACT

According to the shortcomings of the existing incentive mechanism of individual reputation management and sellers' mutual collusion to brush their reputation under C2C e-commerce mode, it comprehensively considers factors that impact individual reputation based on third-party online payment mode, and establishes the reliable feedback factors and their own functions, to deal with dishonest feedback and lack of incentives, and renews the existing third-party online payment mode, technically connects individual reputation management center, logistics center with third-party platform by the Internet authentication, and puts forward the individual reputation management method based on reputation guarantee with collaboration of control information flow, capital flow and logistics, and realizes the effective individual reputation management through underlying database, provides a theoretical basis and technical guidance for creating a healthy, honest atmosphere of e-commerce service.

Keywords: Third-party Online Payment, Individual Reputation Management, Collaboration, Secondary Settlement

1. INTRODUCTION

With the popularization of Internet, reputation management has become an important aspect of credit management, fraud prevention management and customer relationship management, as well as one of the means and basis of marketing. Reputation: calls prestige, credit rating, reputation rating. Entity's reputation is the expected value of future behavior with the given time interval and context environment, according to other entities' observation information or their history behaviors' information[1,2].The reputation management's basic principle is: after users complete a transaction, the opposite party evaluates oneself by many transaction aspects (such as product quality, delivery or payment timeliness and so on), forming reputation feedback, and receives reputation feedback from all transactions, builds up this user's comprehensive reputation scores in certain way to reflect his reputation condition, and to guide other users to make transaction decisions [3].

In recent years, domestic and foreign scholars have aroused interests in the reputation management's research, Resnick proposed online reputation management system based on the web, which was one kind of reputation management tool used in many electronic commerce websites currently, becoming one efficient way to solve online reputation crisis [4]. EBay is the electronic commerce website using reputation management system earliest in the world [5], which is proved that the reputation system has tremendous influence on eBay's success. The online reputation management system is a reputation management mechanism to urge strangers' cooperation behaviors to promote network trust under one network environment through collecting, counting and issuing feedback information of user's history behaviors [6], which is a new idea and pattern of reputation management in electronic commerce.

In P2P (peer to peer) system's reputation and trust management, in early times Aberer proposed management system based on

reputation in P2P system [7], without setting up the feedback reliable factors to deal with dishonest feedback and lack of incentive. At present, in Taobao website, there are false transactions with capital flow but without logistics delivery in which sellers mutually collude to brush their reputation.

Aiming at above questions in the existing online reputation management, it comprehensively evaluates factors influencing individual reputation under third-party online payment mode [8,9], sets up the feedback reliable factors, and puts forward the individual reputation management method based on reputation guarantee with collaboration of control information flow, capital flow and logistics, then confirms its validity and feasibility technically.

2. METHOD OF INDIVIDUAL REPUTATION MANAGEMENT UNDER THIRD-PARTY ONLINE PAYMENT MODE

The third-party online payment platform acts as a fair intermediary role between online customer and business, which is independent on financial institution, seller and buyer, facing business developing e-commerce, which provides foundation support and application support service of e-commerce, mainly through computer technology, network service technology, with main purpose of restraining transaction behaviors of buyer and seller, providing the guarantee for their reputation, thus bridging risk and uncertainty of online transactions, increasing possibility of transaction's reaching a deal, assuring both capital flow and logistics in transaction work [10]. According to third-party online payment mode's characteristics, it proposes the individual reputation management method based on reputation guarantee with collaboration of control information flow, capital flow and logistics, namely buyer and seller start virtual accounts in third-party platform, who carries on the real-time management of individual reputation rating changes caused by delivering goods information of logistics center and virtual accounts changes for transaction's occurring.

The basic framework of this method shown as Fig2-1: Supposing A_i (i=1,…,m) are all individuals that have transactions with B, R_B is reputation appraisal value of individual A_i to individual B. And individual A decided the individual B reputation appraisal by the appraisal of individual A_i to B. After each transaction had ended, individual A_i (i=1,…,m) and B_j (j=1…,n) submitted related reputation feedback rating of A and the B separately. Third-party online payment's reputation management system collected all historical reputation rating of A and B through third-party platform open-door website, according to positive and negative feedback statistics, after incentive mechanism reward and punishment, compiled A's and B's reputation feedback, forming A's and B's initial reputation information, calculated A's and B's reputation through certain model as reputation information, compiled it to obtain the final reputation information, at last, the reputation information was issued on third-party platform's open-door website, supplied to the potential traders to refer.

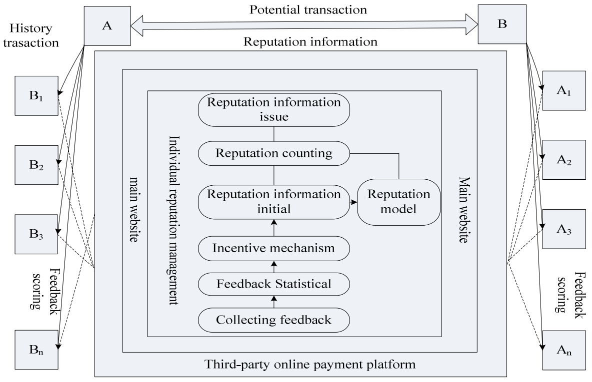


Fig 2-1 The Basic Framework of Reputation Management under Third-party Online Payment

In the incentive mechanism feedback reliable factors are set, namely comprehensively considering how online transaction amount, guaranteed accounts stored on third-party payment platform, online praised rate, online complaining/ complained ratio to influence the reputation level:

(1) Online transaction amount: After a transaction, the more ts transaction amount is, the higher individual reputation is. If only considering online transaction times, then there is possibly free-rider phenomenon, causing the reputation's nonequivalence. Therefore we calculate buyer's and seller's points according to each transaction amount, supposing the buyer initial point is C_B, seller's initial point is C_S, then C_S= C_S+k_2m, $C_B=C_B+k_3m$ (k_2, k_3 is transaction amount multiplier agreed by third-party platform, m is transaction amount). For buyer, the points may enhance own reputation. For seller, the points may enhance own reputation, guarding against the seller's reputation fraud effectively.

(2) guaranteed accounts stored on third-party payment platform: Before the third-party online payment transaction, the buyer transfers certain amount from his bank account to third-party virtual account as cash deposit, which must be equal to the purchase commodity value at least, then the more this amount is, the higher its reputation level is. If not considering this factor, then buyer who has more or less cash deposits obtains the same reputation appraisal, causing the reputation's nonequivalence. Therefore we calculate buyer's points according to the cash deposit stored in third-party account, supposing its initial point is C_B, then $C_B=C_B+k_1$ (M-m) (k_1 is the cash deposit multiplier agreed by third-party platform, M is buyer transferring amount to the third party account, m is buyer's transaction amount). Doing this has manifested the fairness of reputation appraisal, and enhanced the enthusiasm of buyer choosing online payment.

(3) Online praised rate: After transaction ended, both sides make scores mutually, to avoid their collusion to make high praise to enhance each other's reputation, through the third party verification, they will be lowered their reputation rank or warned as soon as they make malicious praise, then deducting certain integral to reduce the reputation rank, to increase the penalty dynamics, here supposing the deduction integral number for the exponential form $C_B=C_S=k_4e^{n_1-1}$ (k_4 is the penalty coefficient set by third-party, n_1 is time of the colluding praise.), and giving the warning; Hyping the reputation many times, given the punishment, when if n_1=10, deprives its transaction qualifications; If is not in cahoots, to encourage the good faith behavior supposing increases the integral number $C_B=C_S=k_5e^{n_2-1}$ (k_5 is the reward coefficient set by third-party platform, n_2 is time of the colluding praise).

(4) Online complaint ratio/complained ratio: After transaction ended, buyer will complain it with his unsatisfactory with the product or seller's service, to avoid his malicious complaint seriously impacting seller's reputation, through the third-party's verification, buyer will be lowered his reputation rank or warned as soon as he makes malicious complaint, to increase the penalty dynamics supposing the deduction of buyer's rating as $C_B=k_6e^{n_3-1}$ (k_6 is the malicious complaint penalty coefficient set by third-party, n_3 is time of malicious complaint), buyer with multiple malicious complainant, if time n_3=10, will be deprived of its transaction qualifications; If not the malicious suit, to increase the penalty dynamics supposing the deduction of seller's rating as $C_S=k_7e^{n_4-1}$ (k_7 is the malicious complaint penalty coefficient set by third-party, n_4 is time of malicious complaint).

Therefore, reputation management on third-party online payment mode can mostly constrain all quarters' behaviors, protecting against fraudulent practices, encouraging good faith behaviors, reducing credit risks, simultaneously retaining bilateral massive information good for confirming bargain partner's credit situation, to lay the foundation for customer relations management, and provide the basis for the market marketing; to seller, the reputation management can transmit information of the product quality, becoming a mechanism controlling product quality; Regardless of views from buyer and seller, the reputation management can reduce transaction cost, time and expense both trade sides searching for the product, collecting information and so on.

3. TECHNIQUE REALIZATION OF INDIVIDUAL REPUTATION MANAGEMENT UNDER THIRD-PARTY ONLINE PAYMENT MODE

3.1. Topology of Individual Reputation Management under Third-party Online Payment Mode

As shown in Fig3-1: on the basis of traditional topology under third-party payment mode, it connects the individual reputation management, logistics center with third-party platform through

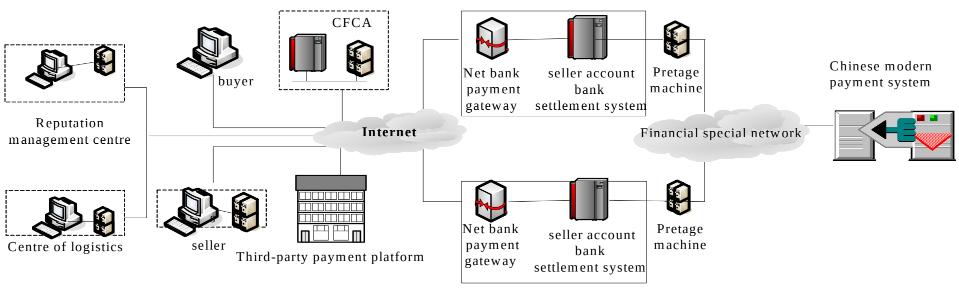


Fig 3-1 Topology Map of Individual Reputation Management under Third-party Online Payment Mode

Internet identity authentication, achieving collaboration of individual reputation management, preventing buyers converting status for another transaction after the occurrence of identity fraud, also avoiding false transactions that sellers collude with each other to do business with only capital flows not logistics delivery to enhance their reputation.

3.2. Transaction Process of Individual Reputation Management under Third-party Online Payment Mode

The collaboration of individual reputation management under third-party online payment mode stresses integration of control information flow, capital flow, logistics works together in the whole e-commerce online payment process and establishes a national information basic database of e-commerce management and comprehensively develops and uses it ,and provides innovative service of online payment.

Its feature is information transferring of all aspects completed throughout transaction process driven by the application system transaction, with all information saved in the relevant application system database. We take the C2C e-commerce online payment as an example to analyze collaboration of the individual reputation management under third-party online payment mode, as shown in Fig 3-2.

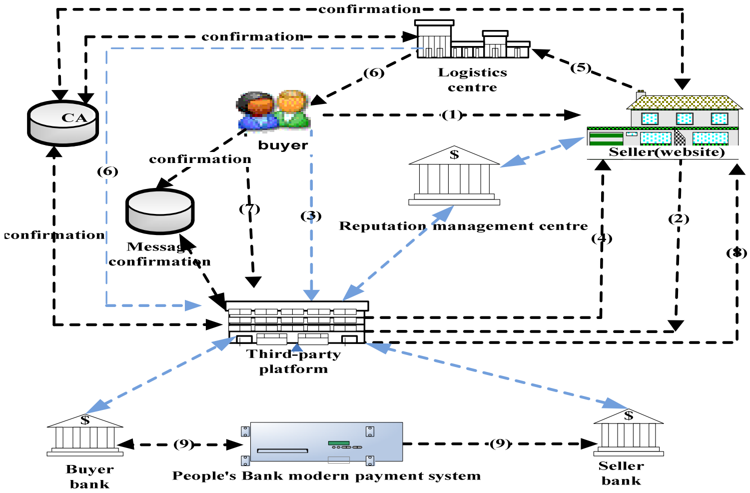


Fig 3-2 Transaction Process of Individual Reputation Management under Third-party Online Payment Mode

Before related e-commerce service and capital flow occur, customer binds information on mobile card and bank card in the security way, and transmits it to third-party online platform and business.

(1) Customer lands the business website, browsing commodity, filling information of purchase order and the bank card payment, then transmits it to the business website server;
(2) Business confirms the customer's order signature and transmits the order information to third-party online payment platform;
(3) Third-party online payment platform confirms information with consumer holding card by mobile;
(4) After confirmation, the customer transfers payment of goods from the account in the bank of the holding card to the virtual account in the third-party online payment platform, the third-party online payment platform informs the business of transferring account information about payment of goods;
(5) Business informs logistics center to delivery;
(6) Logistics center delivers goods to the customer, and feedbacks information to the third-party online payment platform;
(7) Customer feedbacks the goods confirmation information to the third-party online payment platform;
(8) Third-party online payment platform transfers goods payment from buyer's virtual account to seller's virtual account, and informs seller of transferring accounts successfully, the third-party submits the successful logistics delivery information and account transferring information to the reputation management center;
(9) Each Commercial Bank records the fund transferring information of customer, business and third-party online payment platform, and finishes the settlement through the People's Bank modern payment system. Then the transaction, the delivery and the payment entirely completed.

So it can be seen through(6)(7)(8), achieving the collaboration of control information flow, capital flow and logistics, avoiding the occurrence of fraudulent transactions with capital flow without logistics and improving the efficiency of individual reputation management.

3.3. Micro-mechanism of Individual Reputation Management under Third-party Online Payment Mode

In third-party online payment mode, the capital flow transferring among buyer, seller and third-party online payment platform is using secondary settlement mode. That is when buyer and seller reach deals intention online, the buyer will transfer goods payment to the account of third-party platform through bank; when buyer and seller come to an exchange, the buyer sends the confirmation information of receipt to third-party account, and third-party platform will transfer goods payment to the seller's account. This "secondary settlement" mode assures the security of the transaction amount and the interests of buyer and seller. In the process of "secondary settlement" service, third-party payment platform is not simply as a channel linking payment gateway of the banks, but as a neutral third-party agency, reserving the effective transaction information of businesses and consumers, so that it is convenient for both parties to search for orders and related inquiries and reputation assessment after completing transactions, especially facing a transaction dispute, the relational information can be used as strong evidence for the arbitration, which plays a role of secured reputation. It is focused on the change of individual reputation with third-party online payment funds transferring in the whole process and the situation of individual reputation management. To simplify the discussion, first we assume that there is only one seller and one buyer, the seller and the buyer's bank of account is D.

(1) The preparatory phase before transaction:

First, the buyer of third-party platform injects liquidity M_1 to the virtual account, liquidity are transferred from the buyer's bank account to third-party platform. In the commercial banks client, accounting entry is

Debit: buyer's bank deposit account M_1

Credit: third-party bank account M_1

Meanwhile, as a result of third-party platforms and the banking system to achieve the network spread to the payment instructions to pay third-party platform. As a result, third-party payment platform, client accounting entry is:

Debit: third-party bank account M_1

Credit: buyer virtual account m_1

(2) Transaction and payment stage:

Debit : buyer virtual account M_1

Credit: third-party exchange account m_1

Reputation rating management: the buyer increased points for $C_B=C_B+k_1(M_1- m_1)$.

After the buyer's acceptance of purchasing goods or service, informing third-party payment platform of their satisfaction with the goods or service (if t no objection is raised within agreed period of time, then thought buyer satisfied with goods and service), then deciding the flow of fund. This process is second settlement.

① When buyer is satisfied with the goods or service, the accounting entry is:

Debit: third-party exchanging account m_1

Credit: seller virtual account m_1

Reputation rating management: the seller increased points for $C_S= C_S+k_2m_1$, the buyer increased points for $C_B=C_B+k_3m_1$.

② When buyer is not satisfied with the goods or service, third-party platform will hedge funds received. Accounting entry is:

Debit: third-party exchanging account m_1

Credit: buyer virtual account m_1

Reputation rating management: the seller deducted points for $C_S= C_S -k_2m_1$, the buyer deducted points for $C_B= C_B -k_3m_1$.

3.4. Database Table Design

By combining with the Internet, identity authentication and third-party platform, reputation management mechanism can prevent false transactions and the buyer's fraud, on the other hand, conducting unified management for the reputation of both buyers and sellers. Evaluation of customers' reputation is built on basis of funds transferring. Reputation assess and management is started when the funds is transferred from the buyer's account to a third-party platform. The design of database table for the transaction process of third-party payment platform is as follows:

(1) User Registration Information table (account, name, password, ID, E-mail, address);

(2) Main File of Buyer's Account table (b-account, name, account bank name, type, opening account time, balance, online complaint rates, online complained rate, reputation rating, point);

(3) Main File of Seller's Account table (s-account, account name, type, opening account time, balance, online complaint rates, online complained rate, reputation rating, point); It is used to record the user's basic information. Type is divided into individual account or corporate account, balance is the amounts of platform;

(4) Order Details table (order number, logistics number, order number, order price, order details, trading time);

(5) Trading Water table (current account number, b-account, b-name, s-name, s-account number, transaction time, transaction amount, order number); it is used to record a double-account information, the time when the transaction occurred, the amounts, order number, which are the basis of breakdown;

(6) The Financial Account Stated table (account number, name, order number, transaction time, transaction amounts, number of current account); Recording specific information of the transaction and the breakdown when completing a transaction. The buyer and the seller may check the transactions for some times based on their own breakdown. Loan signs represent the change of amounts in corresponding account of the main file, which facilitate to check their balance of the main file in future time; meanwhile, if the buyer is dissatisfied with the goods after receipt and proposes return application, it will strike a balance by recording the same amounts in the opposite direction (to facilitate the operation, we set 0 represented the buyer, and 1 represented the seller in loan signs and the direction);

(7) Temporary table (account number, name, record condition, transaction account, time, third-party current account, loan direction); It is mainly used to save the transaction information after the seller deliver goods and before the buyer send satisfied information of delivery, which recorded conditions with 0 and 1. The condition before the buyer sent satisfied information to a third party platform is recorded as 0. The condition after buyer's satisfaction with the inspection or the buyer did not propose return application in specified time is recorded as 1;

(8) Red Word Hedge table (account, third-party exchange account, order number).

When carrying out such transactions, the process of individual reputation management under third-party online payment is as follows:

(1) The buyer browses the goods selected on the website, and negotiates a good price with the seller. The buyer send merchandise information to the seller through his own account, the seller determines the order information, giving the buyer order number;

(2) The buyer sends the order information and account number to third-party platform, and transfers goods payment to the virtual account in the third-party platform, then sends payment instructions and recording current account, the transaction serial number, account number and name of buyer and seller, trading time, transaction account, order number and buyer's points C_B. According to the character of secondary settlement in third-party platform, it realizes the first settlement;

(3) According to the transaction flow, modifying the balance of the main file of buyer's account, reducing m;

(4) Recording account stated. Recording buyer's transaction account stated, including account number, name, order number, amount, serial number, point and loan direction (the loan direction recorded as "debit");

(5) Temporary table registering transaction. Temporary table of a transaction generates two records, which are third-party platform and the seller respectively. The table records account number, account name, account, time and loan direction referred to the transaction, and the conditions recorded as 0 at this time;

(6) Secondary settlement. The buyer sends receiving information after satisfied with the receipt. The condition of temporary table is modified (as 1) and the balance of the main file of seller's account based on the account information and loan direction of temporary table, adding m, modifying reputation rating of buyer and seller, recording buyer points C_B and seller points C_S. If the buyer dissatisfied with the receipt, the return application is

proposed. Then it modifies buyer's account stated, and records the same amount in account stated, but changes the loan direction to "credit" direction, and modifies the balance of the main file of buyer's account based on the account stated;

(7) Recording seller's account stated according to main file of seller's account table.

4. CONCLUSIONS

In this paper, through researching individual reputation management under third-party online payment mode, it considers factors impacting individual reputation, and sets the amount of online transactions, guaranteed accounts stored on third-party payment platform, online complaint/ complained rate, and online praised rate as the reliable feedback factors, proposes the individual reputation management method based on reputation guarantee with collaboration of control information flow, capital flow and logistics, and technically verifies the effectiveness and feasibility of the method. The conclusions and their significance are as follows:

(1) According to difference of each factor impacting on the individual reputation level, various incentive functions are established. Each third-party platform sets reward and punishment coefficient based on its actual situation, enhancing the operation, scientific and flexibility of reputation management. Meanwhile it can deal with dishonest feedback and problems lack of effective incentive, giving a direction for drawing marketing strategy.

(2) Individual reputation management method is proposed, which realizes collaboration of control information flow, capital flow and logistics under third-party online payment mode technically, and effectively prevents these illegal transactions in which the sellers collude mutually to do business with only capital flows not logistics delivery which is prevalent on Taobao website to enhance the reputation, then purifying the service environment of e-business and promoting to gradually perfect the online payment mode, management and technology to form a virtuous cycle, and drafting a new scheme for improving the reputation management in the future;

(3) For buyers, with improvement of third-party online payment environment, more online users are attracted. Buyers and sellers make transactions by the online payment, making transactions faster and safer. So each buyer's utilities are higher, realizing network's externality of online payment.

5. ACKNOWLEDGEMENT

This research was supported by the Ministry of Educational of the People's Republic of China in Humanities and Social Sciences Planning Fund: (1)Research on Leading Mode of Online Payment and Its Innovation Service, under Grant 07JA630028. (2)Research on Online Payment Collaborative Management Mode , under Grant 08JA630064.

REFERENCES

[1] Farag Azzedin & Muthucumaru Maheswaran, Evolving and Managing Trust in Grid Computing Systems, In: Canadian Conference on Electrical and Computer Engineering, 2002, pp 1424-1429, IEEE Computer Society.

[2] A. Josang, R.Ismail & C. Boyd.A Survey of Trust and Reputation Systems for Online Service Provision, Decision Support Systems, 2005.

[3] Song Guangxing, Yang Deli. Study on Certain Questions of Electronic Commerce Online Reputation Management System Design [J]. System Engineering, 2004:22 (9):529(In Chinese).

[4] Resnick P, Kuwabara K, Zeckhauser R, et al, Reputation systems[J], Communications of the ACM, 2000, 43(12): pp 45-58.

[5] HouserD, Wooders J, Reputation in internet auctions: theory and evidence from eBay [J], Journal of Economics and Management Strategy, 2006, 15 (2): pp353-369.

[6] Dellarocas C, The digitization of word-of-mouth: promise and challenges of online feedback mechanisms [J], Management Science, 2003, 49 (10): pp 1407-1424.

[7] Aberer K, Despotovic Z, Managing trust in a peer-to-peer Information system [A], Proceedings of ACM Conference on Information and Knowledge Management[C], USA: ACM Press, 2001, pp 310-317.

[8] Shao Bingfang, Li Rui. Study on Credit Evaluating System to Auctions Website [J]. Statistics and Decision-making, 2006(5) (In Chinese).

[9] Wang Baoli, Yang Qifeng. Individual Credit Degrees Evaluating Index System Online Payment [J]. China Water Transportation, 2006(4) (In Chinese).

[10] Zhang Kuanhai. Online Payment and Settlement [M]. Beijing: Press of Mechanical industry, 2007(In Chinese).

Song Ping is an Associate Professor with the management doctor degree, in school of economic, Wuhan University of Technology. Her main research directions are electronic payment, international finance and operation and management of commercial banks. She chaired and participated in more than 10 items national and provincial level scientific research projects. She published a monograph and over 30 journal articles. Her scientific research got 1 third prize of scientific and technological progress in Hubei Province. Undergraduates she guided got 1 second prize and 3 third prize of outstanding thesis dissertation in Hubei Province. Masters she guided got 5 awards of Wuhan University of Technology.

COMPUTER-BASED SIMULATIVE STUDY ON ULTRA-CAPACITOR FOR ELECTRIC VEHICLE

Su Chuqi, LI Jingjing
Automobile Engineering School, Wuhan University of Technology
Wuhan, China
Email: suchuqi@163.com, qinghai107@163.com

ABSTRACT

In this paper, the characteristics of ultra-capacitor for EV (electric vehicle) are analyzed, which includes the characteristics of energy conservation, current alteration and voltage transformation during the process of charge and discharge. Considering the impact of temperature on ultra-capacitor, a physical model and a mathematical model of ultra-capacitor are raised, and models for the driving systems of all-electric vehicle and hybrid electric vehicle are established respectively. Based on the simulation result, a conclusion is derived that applying ultra-capacitor to EV contributes effectively to the protection of vehicular batteries.

Keywords: Ultra-Capacitor, Modeling, Simulation, Compound Power Source

1. INSTRUCTION

With the worldwide environment-friendly appeal, all vehicle manufacturers are exerting enormous endeavor to develop all-electric vehicle (AEV), hybrid electric vehicle (HEV) and fuel cell electric vehicle(FCEV), aiming at devising a cleaner and more efficient type of vehicle. At present EV devisal, batteries, such as lead acid battery, Ni-MH battery, lithium battery and the like, are adopted as the most common power source or auxiliary power source. Nonetheless, a major difficulty the designer confronts is the inability of these energy storage devices to fulfill the EV's power demand; on the other hand, if the designer takes the power demand into account, it would result in a waste of energy and a tremendous oversize and overweight of batteries. Meanwhile, batteries are experiencing the impulse from heavy discharging current during vehicle's startup, acceleration and climbing condition. This is responsible for the sharp decline of battery longevity and the frequent replacement of batteries, which turn out to be a distinct barrier for EV's commercialization. Besides, the failure to absorb regenerative braking energy also witnesses a serious flaw of vehicles utilizing a single chemical battery.

Consequently, based on the preceding research, more well-developed energy storage technology is in dire need if a more environment-friendly vehicle is to be invented. Fortunately, ultra-capacitor turns out to be an ideal energy source to be a part of hybrid energy supply, judging from its multiple merits like a fast charge and discharge, an intense power release, pollution-freeness, long life span and so forth.

Systematic modeling and simulation are proved to be an extremely efficient method in studying the control strategy of EV's power system. National Renewable Energy Lab in America devised ADVISOR, a well-known electric vehicle simulator, by using MATLAB; Argonne National Laboratory also devised PSAT, PNGV System Analysis Toolkit, based on test bench dynamic simulation. Since both software fail to provide a consummate model of ultra-capacitor, this paper in general probes further into the simulative model of ultra-capacitor.

2. PHYSICAL MODEL OF ULTRA-CAPACITOR

2.1 Structural Theory of Ultra-capacitor

Possessing high capacitance, double layer electrochemical capacitor is made from powdered carbon or porous carbon with high specific surface area, which amounts to 1000 to 2000 $m^2 \cdot g^{-1}$. If its nominal specific capacitance is $25\mu F \cdot cm^{-2}$, then theoretically the capacitance would be $1000\times10^4(cm^2)\times25\mu F \cdot cm^{-2}$, namely $250F \cdot g^{-1}$, or equal to the energy of $250kj \cdot g^{-1}$. A virtual electrochemical capacitor utilizing electrode/solution phase should consist of two interphases, which work independently from each other in the electrolyte solution (i.e. one positive charge and the other negative charge).In most cases, membrane should be added between two electrodes. Such a capacitor unit system comprised by two electrodes and two interphases is indicated in Fig1 (a).

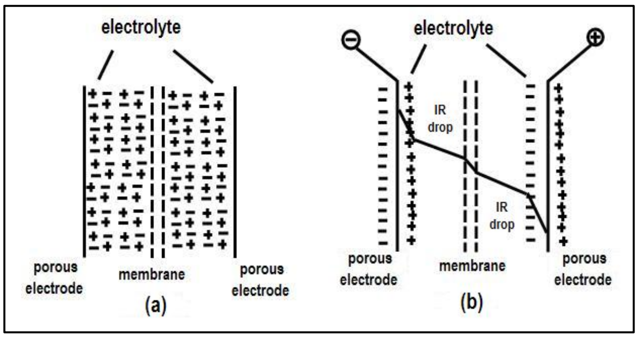


Figure1. Energy Storage Theory Of Ultra-Capacitor

When a pair of solid electrodes is immerged into electrolyte solution, the charges would be distributed and arranged within an extremely short distance between two different phases of solid electrodes and electrolyte solution, if voltage applied is lower than the decomposition voltage of the solution. As compensation, anode containing positive charge would absorb the anion in the solution, while cathode would absorb the cation for the same reason. As a result, a compact electric double layer would come into being, as is depicted in Fig. 1 (b). Accompanied by the formation of the electric double layer, a capacitor is formed at the electrode's interphase, which is called electric double layer capacitor. The energy stored in the capacitor is accumulated on the surface of electrode material in the form of charges or condensed electrons. During the process of charge, anion and cation in the solution would disjoin and move to the surface of each corresponding electrode; when it comes to the process of discharge, anion and cation would be released from the electrode surface and return to the electrolyte solution. Meantime, as is illustrated in Fig. 1 (b), two phase boundary voltage drop exist along the line penetrating the capacitor unit during the state of charge, to which each electric double layer respectively correspond; while during the state of discharge, one ohmic voltage drop reverse to that during charge exists in the solution, which varies with current and equals to IR [1-3].

2.2 Equivalent Circuit Graph of Ultra-capacitor

The equivalent models of ultra-capacitor include the classic RC model and three-phase circuit model, between which the former is simpler and more effective. In the RC model, two types of resistances, ESR and EPR, are adopted. ESR, short for equivalent series resistance, is a kind of virtual series resistance containing capacitor unit, such as electrolyte resistance or contact resistance. Generally the resistance of ESR is small, while that of EPR (equivalent parallel resistance) is large. ESR simulates the energy loss and the instantaneous voltage leap during ultra-capacitor charge and discharge, whilst EPR simulates energy loss due to ultra-capacitor self-discharge. Since ultra-capacitor possesses a small temperature rise during charge and discharge, and ESR changes slightly with temperature fluctuating, which signifies a favorable temperature characteristic, ESR could be regarded as a stable value in modeling. During charge, capacitor is at a state of higher free energy, compared with its state during discharge or partial charge. For this reason, as long as any mechanism supporting the spontaneous reduction of free energy becomes valid, the self-discharge occurs, where RF (i.e. EPR) is used to signify the energy loss from self-discharge. The established equivalent model of ultra-capacitor is illustrated in Fig. 2 [4-5].

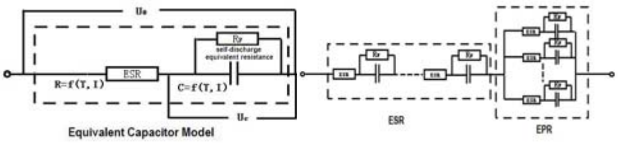


Figure2. Equivalent RC Model of Single Ultra-Capacitor and the Connection of Ultra-Capacitor Group

3. SIMULATIVE MODEL OF VEHICULAR ULTRA-CAPACITOR

Direct current of ultra-capacitor could be calculated by the balance condition between ultra-capacitor's gross power, equivalent resistance's consumed power and capacitor power:

$$P_{Rs}+P_{out}=P_o, \tag{1}$$

$$I_{out}^{2}\times Rs-VOC\times I_{out}+P_{out}=0, \tag{2}$$

$$I_{out}=\frac{VOC-\sqrt{VOC^2-4RsP_{out}}}{2Rs}\ , \tag{3}$$

where VOC is the operating voltage of ultra-capacitor, I_{out} is the operating current of ultra-capacitor, P_{out} is the output power of ultra-capacitor, and Rs is the equivalent series resistance.

Suppose at a specific time point the ultra-capacitor voltage is VOC(n), then the voltage at next time point could be deduced as:

$$VOC(\mathrm{n}+1)=VOC(\mathrm{n})-(I\times dt)/C. \tag{4}$$

Meanwhile, the operating voltage of ultra-capacitor should be restricted between the value of V_{min} and V_{max} so as to protect the ultra-capacitor itself. Moreover, since the charge stored by capacitor is in linear relation with voltage, the state of charge (SOC) could be calculated more readily:

$$SOC_{\mathrm{uc}}=\frac{Q_{\mathrm{remaining}}}{Q_{\mathrm{total}}}=\frac{C\left(VOC-V_{\min}\right)}{C\left(V_{\max}-V_{\min}\right)}=\frac{VOC-V_{\min}}{V_{\max}-V_{\min}}\ , \tag{5}$$

where SOC_{uc} is the state of charge of the ultra-capacitor, V_{max} is the maximum operating voltage permitted by ultra-capacitor, V_{min} is the minimum operating voltage permitted by ultra-capacitor.

According to the above-mentioned modeling formulas, the model of ultra-capacitor is built in ADVISOR based on SIMULINK module in MATLAB, as is indicated in Fig. 3.

The theoretical model of ultra-capacitor depicted in Fig. 3 mainly serves as the power source of the entire EV system, and as the braking energy regeneration system. The upper half of the model signifies the calculation chain of ultra-capacitor from power required to power achieved, while the lower half refers to the calculation of the heat produced by ultra-capacitor during its operation.

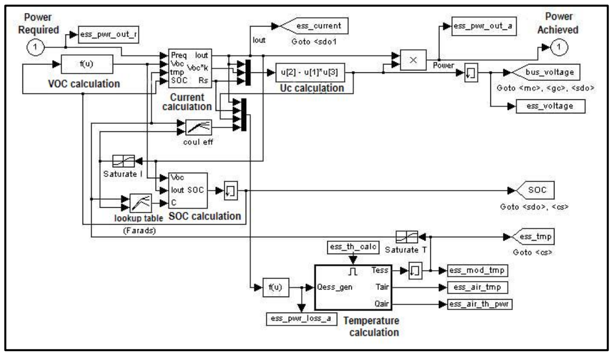


Figure3. Simulative Model of Ultra-capacitor in ADVISOR

4. REVISION OF THE ULTRA-CAPACITOR'S SIMULATIVE MODEL

4.1 Self-discharge Mechanism of Ultra-capacitor

When capacitor has been placed in open circuit for a period of time, a certain degree of self-discharge might arise, which is related to the system's chemical and electrochemical characteristics, to the purity of the reagent and electrolyte, as well as to the temperature. The probable mechanism giving rise to self-discharge mainly includes [1]:

- When the capacitor has been overcharged beyond the decomposition bound of electrolytes, self-discharge would occur spontaneously for reducing the over-potential stemmed from overcharge.
- Suppose impurity is contained in capacitor's electrolytes, then if the impurity could be oxidized or reduced under the potential difference generated during the charge of capacitor, the potential-related inductive charge's leakage current disables the polarizability of capacitor in a certain degree.
- As far as double layer capacitor is concerned, self-discharge could occur through internal oxidation-reduction reaction involving certain functional groups.

4.2 Revising Ultra-capacitor's Simulative Model

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Based on the self-discharge mechanism of ultra-capacitor, ultra-capacitor's model in ADVISOR is revised. EPR is employed to simulate the self-discharge loss of ultra-capacitor; as a result, the ultra-capacitor's gross power should include the power of equivalent series resistance, the power of equivalent parallel resistance, and the capacitor power [6]. Eventually, the direct current in ultra-capacitor could be derived by following equations:

$$P_{Rs}+P_{out}+P_{RF}=P_o, \tag{6}$$

$$I_{out}{}^{2}\times Rs-VOC\times I_{out}+\frac{VOC^{2}}{R_{F}}+P_{out}=0, \quad (7)$$

$$I_{out}=\frac{VOC-\sqrt{VOC^{2}-4Rs(P_{\text{out}}+\frac{VOC^{2}}{R_{F}})}}{2Rs}, \quad (8)$$

where R_F is the self-discharge resistance.

The simulative model of ultra-capacitor is built based on its self-discharge mechanism, as is indicated in Fig. 4.

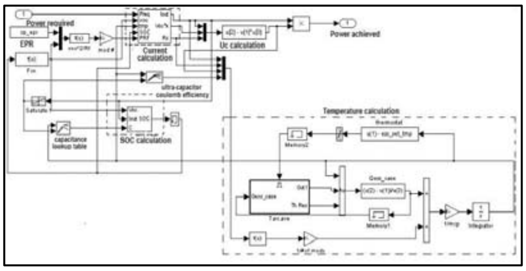


Figure 4. The Simulative Model of Ultra-Capacitor After Revision

5. SIMULATIVE STUDY ON TYPICAL ELECTRIC VEHICLE

5.1 Simulation of All-electric Vehicle

AEV (All-electric vehicle) adopts battery as its single power source. A high-current discharge is entailed during vehicle's acceleration or climbing condition, whereas during vehicle's deceleration or downgrade condition a swift charge is indispensable to facilitate the braking energy regeneration. Consequently, battery should possess some fine features such as high-rate charge-discharge capability, long life span, stable performance and the like. However, a high-current charge-discharge means that battery's life span would be considerably shortened; besides, most kinds of AEV are mini vehicle, whose interior space is relatively limited for the disposal of battery. Since a compact disposal might lead to heat accumulation and expose the battery to a high-temperature surrounding, the performance of battery would be seriously undermined due to high-temperature failure, which poses an obstacle for engineers to further enhance AEV's performance.

AEV's simulative model is depicted in Fig. 5. Its driving system contains following power components: an AC driving motor with a rated power of 75kw and a maximum efficiency of 92%, and a lead acid battery pile with a capacity of 26Ah and a maximum power of 25kw. The battery pile consists of 25 cells, with a single cell's specific power 91w/kg and weight 11kg.

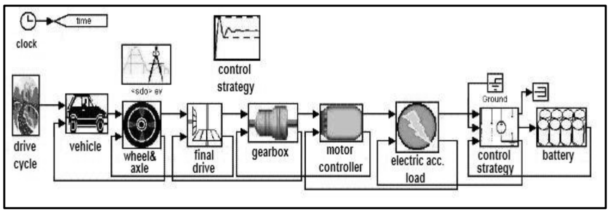


Figure 5. The Simulative Model of All-electric Vehicle

5.2 Simulation of Eletricity-electricity Hybrid Electric Vehicle

Since in AEV the battery's life span would be tremendously shortened because of the high-current discharge during acceleration or climbing conditions, a strategy of integrating ultra-capacitor and battery together is proposed, considering the high power density of ultra-capacitor could substantially reduce the impulse on battery from high-current charge or discharge. Meanwhile, the temperature rise of battery could be brought down as well, which improves the operating performance of EV.

An HEV (Hybrid electric vehicle) simulative model using ultra-capacitor and battery as the compound power system is shown in Fig. 6. In the compound power system, the power components in use include an AC driving motor with a rated power of 75kw and a maximum efficiency of 92%, a lead acid battery pile with a capacity of 26Ah, and a Maxwell PC2500 ultra-capacitor. Suppose a peak power of 25kw is demanded by the driving system, then the battery's power output is restricted to 6kw for protecting itself, and the ultra-capacitor's power output is 19kw. Since the discharge efficiency of battery is 60%, the battery's cell amount could be calculated as 6/0.6 = 10, while ultra-capacitor's cell amount is 19000/(2.5*225)=34.

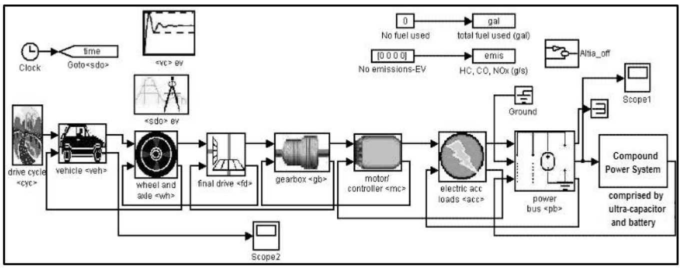


Figure 6. The sSmulative Model of the HEV

5.3 Analysis of Simulation Results

Now the established models of both driving systems of AEV and HEV are integrated into ADVISOR, the simulation software based on MATLAB/SIMULINK platform, for a simulative study.

In the simulation, the CYC_ECE_EUDC cycle, i.e. the European urban and suburban cycle, is utilized, as is shown in Fig. 7.

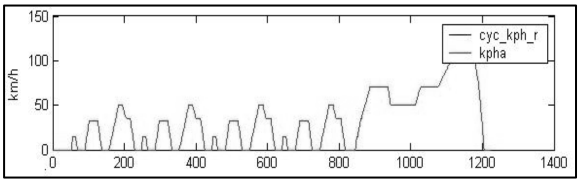


Figure 7. Simulation Output of HEV Speed

Judging from the simulation results reflected in Fig. 8 and Fig. 9, the performances of AEV which utilizes battery as its single power source and HEV whose power source consists of ultra-capacitor and battery are contrasted through the entire cycle. In AEV, as the demanded power changes, the battery current assumes an intense fluctuation, which is adverse to the battery application. On the contrary, in HEV, the compound power source ensures a relatively smooth current change, which potently safeguards the battery. While the entire cycle is taken into account, the built compound driving system is proved to successfully fulfill the EV's designing request and efficaciously guarantee the battery's safety.

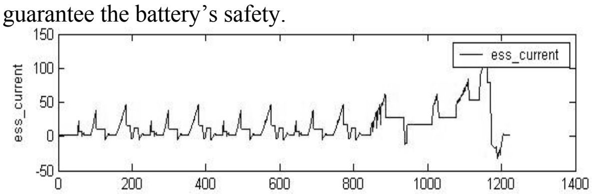


Figure 8. Simulation Output Of AEV's Battery Current

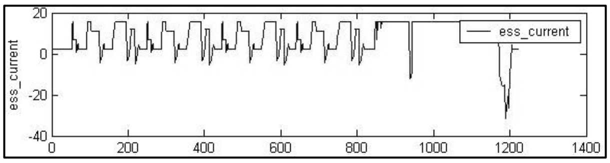


Figure 9. Simulation Output Of HEV's Battery Current

Furthermore, another conclusion could be derived from the simulation results that an appropriate control strategy is in need to facilitate a joint energy flow control of both ultra-capacitor and battery, and hence make a more sufficient use of ultra-capacitor.

6. CONCLUTIONS

The work accomplished and the conclusion derived is summed up as follows:

- Taking the influence of the multiple factors such as temperature, equivalent resistance and so forth into consideration, the simulative model of ultra-capacitor is built.
- With the above-mentioned ultra-capacitor model being applied, an electricity-electricity HEV's simulative model is entrenched, whose simulation result proves that the driving requirement could be well met.
- By comparing the simulation results of the HEV model and AEV model, the ability of HEV's compound power system to protect its battery is confirmed.
- Energy flow control on ultra-capacitor and battery still need to be achieved by establishing apropos control strategy.

REFERENCES

[1] B. E. Conway, Electrochemical Supercapacitors: scientific fundamentals and technological applications, New York: Plenum Press, c1999. 698p.

[2] Liu Zhixiang, Dong Guojun, and Zhang Milin, "Preparation and performance of electric double layer capacitor," in Chinese Journal of Power Sources, vol. 11, 2001 Beijing.

[3] Zhang Li, Song Jinyan, and Zoujiyan, "Investigations on packaging structure and temperature field of hybrid supercapacitors," in Electronic Components Materials, vol. 11, 2005 Beijing.

[4] Li Junqiu, Sun Fengchun, Zhang Chengning, and Li Honglin, "Ultra-capacitors parameter matching design and experiment of electric bus," in Chinese Journal of Power Sources, vol. 8, 2004 Beijing.

[5] Zhang Jing, "Investigation and simulation of the combined energy storage system of ultra-capacitor and battery," Master thesis, Wuhan: Wuhan University of Technology, 2005.

[6] Qiang Guobin, Li Zhongxue, and Chen Jie, "Modeling of ultracapacitor energy storage used in hybrid electric vehicle," in Energy Technology, vol. 2, 2005 Beijing.

APPLICATION OF CORRELATION ALGORITHM IN IDS EXPERT KNOWLEDGE BASE

OuYang ZhengZheng
Department of Computer and Information Engineering, Wuhan Polytechnic University
Wuhan, Hubei, China
Email: scentlilyzz@163.com; oyzz@whpu.edu.cn

ABSTRACT

This paper presents a method based on data mining technology to set up an intrusion detection system, which has characteristics of scalability, adaptability and accuracy. On the basis of analyzing the association mining algorithm combined with a number of attacks on networks, the eligible item set is identified and rule-base is set up, which has improved the ability of intrusion detection.

Keywords: Intrusion Detection, Data Mining, Apriori

1. INTRODUCTIONS

The Intrusion Detection System (or the IDS for short) is mainly used to detect the illicit attacks to computer network system or information system. Intrusion Detection System is the core of security technologies, which is an important supplement to firewalls. It can combine with the capacity of other network security products effectively, which provides the network security with an active and real-time all-round protection.IDS can be used to do the real-time detection of network boundary point data and the accessing server data stream, and then discover various attacks such as denial of service attack (DOS) effectively, to prevent the destruction caused by the invaders.

The Intrusion Detection Systems usually do the real-time detection of network behavior and the system status by pattern matching, statistical analyzing and other methods. It's shown in the Practical applications that the pattern matching is very appropriate for the detection of known attacks, with the characteristics of high accuracy and speed, but it's powerless for unknown intrusions.

Therefore the applications of data mining in intrusion detection system is to consider the intrusion detection as a process of data analysis, which can draw an unknown attack model by data mining methods, find knowledge from data of mainframe and network by data mining technology, set up rule-base of invasion and normal behavior, and then extract behavior model of the users and system in real-time detection by data mining to detect. The system can effectively identify and update the rule-base automatically, improve the scalability and adaptability of system and reduce the rate of false alarm or omitted report effectively.

2. BUILDING PROCESS OF IDS KNOWLEDGE DISCOVERY

In the intrusion detection system based on expert knowledge-base, the expert knowledge-base can be used to record and describe the normal and abnormal behavioral trait of user, it can also be used to record the behavioral patterns of known attacks and the attacks by the using of known system vulnerability. IDS do real-time and all-round invasion monitoring to the users' activities of multiple goal systems. It analyzes the network packets by data mining algorithms, match with knowledge-base of the constructed invasion system, which can be updated automatically by the using of data mining methods, and then find out the normal or malicious behavior of user .

The building of expert knowledge-base is very important; a good knowledge-base can reduce the rate of false alarm and omissions. Data mining is a particular step of setting up an expert knowledge-base, it extract a model from network data through specialized algorithm, and when the network packets arrive at the intrusion detection system, they need pre-treatment and treatment for noise reduction so as to get the important information such as time, source IP, source port, destination IP, destination port, connection ID and connection status. Such information is more conducive to data mining and qualified for expert knowledge-base. The building process of system knowledge-base is shown in Fig.1.

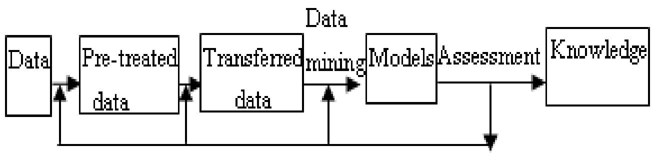


Figure 1. Discovery and Building Process of Knowledge Base

The characteristics of behavior patterns of known attacks and the attacks by the using of known system vulnerability can be added to expert knowledge-base according to specified rules, which is more targeted and accurate in detection of particular invasion, while we can extract the rules for unknown attacks by collecting enough data for training of the data mining module, then the rules be added to expert knowledge-base by data mining.

3. BASIC IDEA OF ASSOCIATION RULE APRIORI ALGORITHM

3.1 Basic Concept of Association Rule

Correlation analysis is an important method of knowledge found, if the values of two or more items repeated at a high rate, there's some association. In the building of knowledge-base, correlation analysis does the data mining by Apriori algorithm. The algorithm is used to find the rules such as 90% customers buy merchandise B when they bought A. The following are some basic concepts of association rule.

Assume that $I=\{i_1,i_2,\ldots,i_n\}$ is a collection of items, D is a collection of affairs, transaction T is a collection of items, and T□I. A is an item collection in I, if A□T, then transaction T contains A.

Definition 1: Assume that A and B are both collection, and A ∩ B = Φ. The supports of rule is S and the confidence is C. The support S means the probability that transaction D contains A∪B (both A, B), that is P (A∪B). The confidence C means the probability that the transaction contain A in D may contain B at the same time, that is, conditional probability P (B | A). Then association rule can be expressed as the following contain form A → B [S, C].

Definition 2: The collection of items is called item set, and the item set which contains K items is called K-item set. If the item set meets the minimum supports, it's called frequent item set.

There are two steps according to the basic ideas of association rule. First, find out all the frequent item sets whose support is lager than minimum supports, that is, frequent collections; The second step is to generate the desired rules by the using of frequent collections found in the first step. The core method is the recursive method based on frequent collection theory. The association rule extraction problem for unknown transaction collection D is to generate all the association rules that the supports and confidence are both larger than the minimum values specified by users. There are two steps to solve this problem:

(1) Generate all the min.sup item sets whose support is lager than minimum support. And these item sets are called frequent item sets.
(2) Generate all the rules that the confidence is larger than minimum confidence for each frequent item set.

Since the second step is easier than the first step, the current researches focus on the first step, that's to find out the frequent item set. Because the number of different item sets is large, the demanding on the algorithms is high.

So far, the most famous association algorithm of finding out frequent item sets is still Apriori algorithm proposed by R.Agawal in 1994.

3.2 Basic Idea of Apriori Algorithm

Apriori Algorithm is the basic algorithm of mining frequency item sets for Boolean association rule, it's named in accordance with the prior knowledge relevant to the characteristics of frequent item sets. This algorithm fulfills the mining of frequent item set by the using of a cycle approach with a level-sequence search. This cycle approach is to use K-item set to generate (K+1)-item set. The detail approach is: First find out frequent 1-item sets, recorded as L1; then use L1 to mine L2 item sets; continue the cycle until you can't find out more frequent K-item sets.

4. APPLICATION OF IMPROVED APRIORI ALGORITHM

4.1 Analyzing and Improvement Methods of Data Mining

The purpose of association rule is to find out frequent item sets and the important part of Apriori algorithm is to prune and exclude the item sets with low supports. But in the network attacks, detecting weaknesses of system by the using of scanning tools is commonly the prelude to attacks. In the scanning attacks, it's not prevailing in the packets that the port data that client requested to carrying out scanning but access the target IP. According to the non-mainstream activities in collected data, it's the needed aspect for intrusion detection system to discover the scanning attacks accurately. A common client accessing transaction is shown inTab.1.

Table 1. Accessing Transactions Data

Client IP	IP Port List
192.168.10.2	80,21
192.168.10.22	80
192.168.10.68	21
192.168.10.22	80
192.168.10.22	80,21
192.168.10.68	80
192.168.110.1	21
192.168.120.68	1034,1035

As we can see, in the accessing transaction to different IP ports, the s=62.5% for 80-port accessing, s=50% for the 21-port accessing, while the support is only 12.5% for 1034 and 1035 accessing, the confidence is 100%. It's possible to be an attack similar to a scanning behavior according to correlation analysis. Now we can change the Apriori algorithm appropriately, that is mainly to change the function which generates a new candidate set. The improved algorithm is:

Procedure apriori_gen(Lk-1,min_sup)
C_k=Φ;
C_k'=Φ;
For each itemset li∈Lk-1
For each itemset lj∈Lk-1
If (li[1]=lj[1])∧(li[2]=lj[2])∧…∧(li[k-1]=lj[k-1]) then
{ c= li join lj ;
if has_omfreqient_subset(c, lk-1) then
add c to C_k; //generate non-frequent item sets
else add c to C_k'; //frequent item sets
}
return C_k';

C_k'can be reused for mining by other data mining methods and it is the target item set for this mining. The selection of training sample and support is important for attacks of this type,high support will result in missing judgment, while low support is not possible to collect samples.

4.2 Further Explore to Algorithm

The accuracy of new algorithm used for common unknown attacks is high, but it turns out to be powerless when used for attacks to same common ports, such as Dos attack. Dos attack is one type of main network attacks, whose feature is to send a lot of packets in a period and the common method is SYS Flood. SYS Flood forges many false addresses in a period to attack a particular IP, which results in the long-term semi-join waiting status of many server resources, and ultimately the resources run out.

In the association analysis of Dos attack, if we definite a transaction as access by each IP to a server IP, the mined supports and confidence may be low, which results in missing judgment, so it needs further improvement.

The new algorithm definites a transaction as each couple of property values of records, data needs to be pre-treated such as discarding, discretion and formatting, purpose of the discretization of some continuously variable is to discover association rule. Here we choose five properties (time, sip, protocol, desip, state) for data mining, the properties of records mainly including the time, source address, protocol, destination address, state. All matters in accordance with time stamp (time stamp here is a specified unit of time), can be on an accessing sequence to a IP. If the state is waiting for connection or refuse, the frequent item set of Dos can be discovered. By the using of the sequence models class Apriori algorithm, we can find out the frequent sequence whose state is waiting for connection or refusing in a period. The class algorithm discovered by sequence model is:

L1={i|i∈I∧P(i)/N>= min_sup};
For (k=2;Lk-1≠Φ;k++)
{C_k=apriori_gen(Lk-1, min_sup); // C_k is a K-item candidate set
For all transactions t∈D do
{C_t =subset(C_k, t); // C_t is all the candidate included in t
For all candidates c∈Ct
c.count+ + ;
}
Lk={ c∈C k|c.count>=minsup_count}
}

The algorithm above is almost the same to Apriori. This algorithm will generate new candidate 1-sequence alternatively, cut off the non-frequent candidates of (1-1) sequences, and then count the left candidates.

4.3 Application of Matrix

It's time-wasted and inefficient to analyze from the network accessing database. The application of data mining algorithms in IDS for data analyzing is a better method. Apriori algorithm needs to scan database each time when generating candidates with different numbers of items. And this algorithm will consume much time when the scale of candidates is large. In addition, since the data of transaction database is increasing continuously, Apriori algorithm has to do the two tasks — to generate frequent item sets and association rules both—once again each time when the data is increased, which means the previously generated frequent item sets and association rules turn out to be of no value and this obviously is not conducive to rapid and efficient discovery of association rules.

Application of matrix can optimize the Apriori algorithm, which reduces the number of scanning times in transaction database and improves efficiency. The basic idea of this algorithm is to give matrix expression to the database. The specific methods are: Order each member to a sequence and so does the transaction set. Members mean row vector and transaction means column vector.If the ith member is in the jth transaction, then the value in row I and column J is 1, if not, the value is 0, which is called the Boolean matrix of database. The database above can be expressed by the following matrix.

$$\begin{pmatrix} 11001101 \\ 11110101 \\ 10101110 \\ 00000001 \\ 01010000 \end{pmatrix}$$

In this matrix, the sum of the row vector is the emerging times of members, and then the support of an item set can be obtained. As to the two item set {I_i, I_j} we only need to scan the ith row and the jth row, the number of 1 in the same column is the supports of two item set {I_i, I_j}, and so on. We only need to scan the I_{i1},I_{i2},…,I_{ik} th row in the matrix, the number of 1 in the same column is the supports of K-item set {I_{i1}, $I_{i2,\ldots,}I_{ik}$}. As can be seen, Ci scan only part of the database, but Apriori algorithm need to scan the full database, so the efficiency of algorithm is improved.

5. EXPERIMENTAL RESULTS

During the experiment, we extracted 10000 records as a sample for training, including the access data to IP addresses of two servers, the results during the attack scanning are s = 36.86%, c = 82.32%. As we can see, the possibility of scanning attack is high, and as the training data is added to 15000, s reduced to a value of 23.22%. Similarly, during the Dos detection testing, according to the sample data, the s has reached 96.28% and c = 93.32% in unit time.

6. CONCLUSIONS

The intrusion detection systems are required to solve the problem of improving the detection rate of unknown intrusion and lower the false alarm rate. On the basis of analyzing the data correlation, this system can improve the intrusion detection accuracy effectively by the using of data mining algorithms. The automatically mining of effective rules by data mining algorithms and the in-time updating of rule-base enhanced the automation of the system. With the deep research of data mining technology, especially the mining algorithms, provided channel to optimize the performance of system, and at the same time, it fully integrated the anomaly detection techniques and misuse detection technology, which realized the mutual support of relative advantages, improved the scalability, adaptability and accuracy of system.

The application of data mining in intrusion detection is a new topic and there are still some questions to be solved. With the development of computer network technology, the detection of attacks of new types effectively needs to be researched right along.

REFERENCES

[1] K.Ali,S.Manganaris,and R.Srikant.Partial Classification using Association Rules.In Proc.of the 3rd Intl. Conf.on Knowledge Discovery and Data Mining,pages 115-118,NewPort Beach,CA,August 1997.
[2] C.C Aggarwal and P.S. Yu. Mining Large Itemsets for Association Rules.Data Engineering Bulletin,21(1):23-31, March 1998.
[3] Han J,et a1.Data Mining:Concepts and Techniques[M]. BeiJin:Machinery Industry Press,2001:149-222.
[4] Margaret,H-Dunham.Data Mining：Introductory and Ad．vancedTopics[M].BeiJing: Tsinghua University Press, 2003:164-192．

INTELLECTUALIZED BIDIRECTIONAL TIRE PRESSURE MONITORING SYSTEM BASED ON CAN BUS

YANG Ye[1,2], ZHANG Xi-huang[1]
[1]Southern Yangtze University School of Information Technology
Wuxi 214122,China
[2]Wuxi Professional College of Science and Technology
Wuxi 214028,China
Email：big_ye@163.com

ABSTRACT

To improve the performance of the independent tire pressure monitoring system, based on wireless sensor SP30 and CAN (control area network) bus, the intellectualized TPMS was realized. This system could monitor the real-time pressure and temperature in the tire. The tire position could be identified automatically. It could exchange the data with other nodes in the CAN bus. The bidirectional communication could be realized between the monitor module and the wireless sensor module. Compared with the indirect monitor system, the difference of direct-monitor was listed. The installation and the failure identification of the sensor module were analyzed. The control scheme, hardware and software design were introduced. The bench test in the lab showed that the TPMS can monitor the tire pressure and temperature accurately. The data can be exactly exchanged between the TPMS and CAN bus node.

Keywords: Sensor, TPMS, CAN bus, Self-identifying.

1. INTRODUCTION

Current annual traffic accident caused by a puncture in all highway accidents account for a large proportion. Tire Pressure Monitoring System (TPMS) can be used on the tire pressure and temperature monitor, when the tire pressure is too high, too low or the temperature is too high, give the alarm information, then ensure traffic safety. Proper tire pressure, also can improve the drive comfort, and extend tire life, reduce fuel consumption.

The current system is divided into indirect pressure monitoring systems and direct pressure monitoring system. Indirect system costs are lower than direct air pressure monitoring system, but with one or two diagonal tire pressure is too low at the same time or the speed is faster than 100Km / h, the system can not determine [2]. Direct tire pressure monitoring system is installed in each tire using internal wireless sensor module to measure pressure and temperature, the measurement results to encode, through the radio frequency chip is sent to the monitoring receiver. When the tire pressure or temperature anomalies and rapid leakage occur, TPMS will automatically make a warning message, start the voice of alarm, fault location and tire air pressure value are displayed in the screen flicker.

2. TIRE PRESSURE MONITORING SYSTEM DESIGN

Direct TPMS systems include wireless sensor modules, wireless receiver and monitoring modules, low-frequency communication module, TPMS display device, CAN bus control. The structure of the system as shown in Figure 1.

Wireless sensor module through the low-frequency (LF) communication interface receives the data acquisition command by monitoring system, tests tire pressure, temperature and battery voltage. Manchester encoding for test results, data through the high-frequency RF wireless interface is sent to the monitoring module. Monitoring module installed in the dashboard display underneath receive wireless signals, decode and analyze of testing data. If the temperature or pressure beyond the control limit, the alarm signal output. Data processing through the CAN bus to the instrument display panel display. Through the CAN bus interface, TPMS monitor module will get other vehicles information from the bus node, such as whether to activate the vehicle speed and ambient temperature information. If a vehicle stop, wireless sensor modules will be maintained power sleep state. There are four low-frequency communications modules, each module is installed as near as possible places around the tire. In order to save costs, TPMS monitor module control four low-frequency communication module through the RS485 bus.

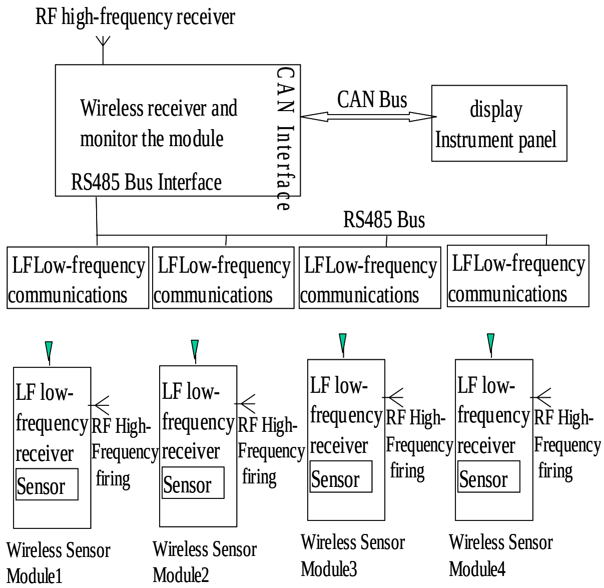


Figure 1. TPMS System Architecture Diagram

3. MONITORING SYSTEM HARDWARE DESIGN

3.1 Integrated Sensor

SP30 is a high level integration, newly developed tire pressure sensors by Infineon company , pressure sensors，temperature sensors, acceleration sensors (optional), voltage monitoring circuits, signal conditioning circuits, an 8-bit micro-chip unit, LF low-frequency interface, power management circuit are integrated.

3.2 Wireless Sensor Module

Wireless Sensor Module include integrated sensors SP30, battery, RF radio frequency and antenna TDK5110 control, LF low-frequency antenna, as shown in Figure 2. SP30 to detect the tire air pressure, temperature and sensor battery voltage power supply module, the test results through the radio frequency RF chip TDK5110 control transmission to the wireless modem is sent to the monitoring receiver. LF low-frequency monitoring device interface for receiving 125kHz low-frequency control signals. Use this LF interface device can monitor the sensor control module.

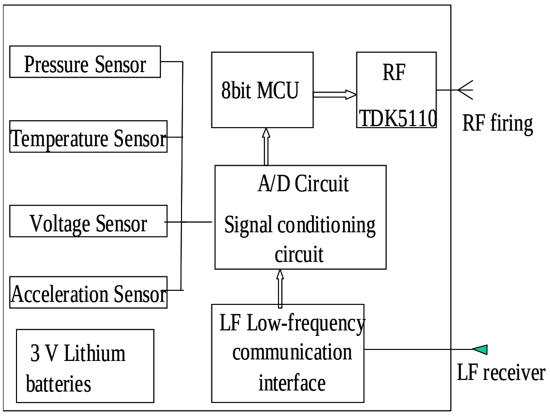


Figure 2. Wireless Sensor Module

3.3 Monitor Device

Monitoring device include RF receiver circuit, 8-bit MCU, CAN controller, RS485 communication modules and power components. Figure 3 Schematic diagram of receiving equipment of monitor.

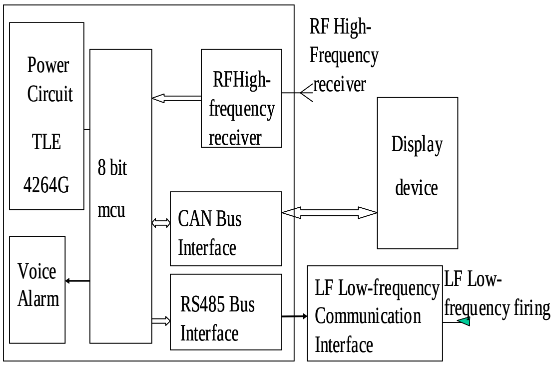


Figure 3. Monitor Receiver

High-frequency RF receiver circuit receive wireless signals sent by the wireless sensor module. Single-chip microcomputer, according to the pressure and temperature limit values, compared it. If the measured value exceeded the control range, the monitor output alarm signal to start the voice alarm function. Receiving device through the single-chip CAN controller, the measurement results will be sent to instrument panel display.

Monitoring device through the RS485 bus to send the command to LF low-frequency communication module of the wireless sensor module. Use this 125KHz signal, the monitor and the wireless module can realize bidirectional exchange of information. Monitoring devices can set the parameter on the wireless sensor module, inform the wireless sensor module for data collection and transmission. Based on vehicle speed, control module can control the data acquisition frequency. Monitoring through LF communication module to start the various wireless sensor module, sensor module return the single ID number that can automatic recognize the location of every tire. When a wireless sensor module failures, it can be found immediately. At the same time, there is only one sensor module exchange the information with the monitor, so it completely avoided in the general wireless TPMS sensors send data at the same time, resulting in the issue of communication conflict.

4. MONITORING SYSTEM SOFTWARE DESIGN

4.1 Wireless Integrated Sensor Module Software Design

Sensing software modules including system initialization, start monitoring the pressure, pressure and temperature measurement, data processing, the battery voltage detection, data, wireless upload. Figure 4 show the wireless sensor module software control flow.

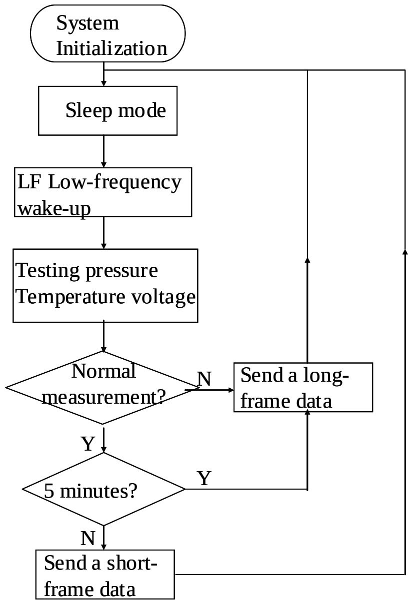


Figure 4. System Software Processes of Wireless Sensor Module

Wireless sensor system initialization module, the module then go to the power-saving mode, waiting for LF low-frequency monitoring module interface to wake-up. Wireless sensing modules receive the order form the monitor. It immediately detect pressure, temperature and battery voltage. In order to reduce the chance of measurement error, the use of median filtering algorithm for digital filtering technology. Data on the detection and control of the comparison, through wireless means, using Manchester coding, the measurement results will be sent to the monitoring module. And then re-enter the power-saving mode, waiting for the next wake-up call. Because under normal circumstances, the tire pressure and temperature will not vary much, in order to reduce the wireless sensor module power consumption at the time of the send. In normal circumstances, the test data will be send at intervals of about five minutes. Then send the value to update the displayed value. Other times only to send the ID number of sensor modules to inform the monitor module that the sensor module work state is normal.

Wireless sensor modules and control modules are achieved through the wireless data frame. Wireless sensor module uses two types of data frame length, divided into short and long frame.

Table1. A short frame format

ID	CRC Check Code	End code
32 bit	8 bit	8 bit

Table2. A long frame format

ID	Pressure	Temperature	Battery voltage	CRC check code	End code
32 bit	8 bit	8 bit	8 bit	8 bit	8 bit

Starting frame for the 32 pre-digital, ID of SP30 chip is 32 numbers, two identical ID numbers appear very low probability. Pressure and temperature values for the 8-bit unsigned number. Battery voltage is normal or under-voltage of the flag.

Wireless sensor communication module and monitoring module of the main functions are:

1） int sensor_init(void)

The function is used mainly for wireless sensor monitoring and initialization, if detected the wireless sensor module initialization, return 1, otherwise return 0.

2） int command_send(char *comm)

The function is used to send commands form monitor module to wireless sensor module， so that the wireless sensor data are acquired and sent expediently .

3） int data_recv(char *buf)

Data_recv function receive data from wireless sensor modules and frame format for data analysis.

In order to ensure the reliability of RF high-frequency communications, error detection must be used. CRC error checking methods because of their ability, coding and error detection algorithm achieve using because it is relatively easy.

4.2 Monitoring Control Module

After power on, the monitoring module to complete initialization, automatically identify the location for the tire, and then enter the normal operating mode. Figure 5 is the software flow chart of control modules.

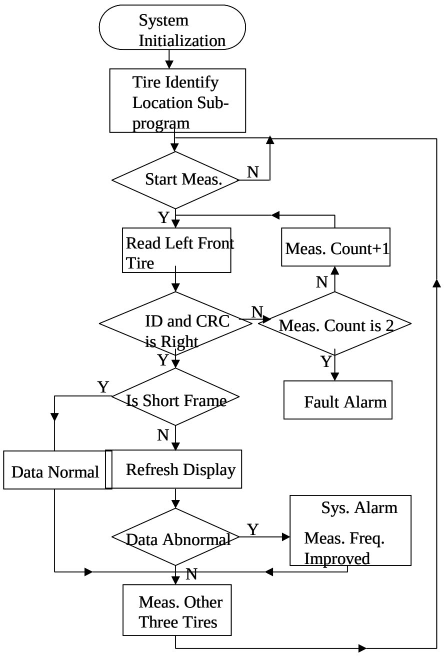


Figure 5. The Software Flow Chart of Contro

Automatic Identification of the location of the tire: monitoring module through the LF low-frequency communications interface, inform ordinal the left front wheel the wireless sensor modules, the right front wireless sensor modules, the left rear wireless sensor modules and the right wireless sensor modules and send the ID. In order to avoid interference from wireless sensor module for other vehicles, the entire identification process repeated five times.

In normal working mode, the control module through the low-frequency communication interface followed by polling of the wireless sensor module. When receiving a short frame the tire pressure is normal, display data is not refreshed. When receiving a long frame then determine whether measurements is out of the control limit and display data is refreshed. Monitoring module polls frequency of sensor module is one time in one minute. If tire pressure data is abnormal it will increase then the frequency of polling, continuous rotation of the module information, until the tire pressure data is normal or vehicle stop running.

5. EXPERIMENTAL RESULT

Tire inflation pressure of the first two 2bar, two tire inflation pressure 2.2bar. Laboratory ambient temperature is 26 degrees, simulating the exercise of the speed of vehicles from 60Km / h to 120Km / h, running 200 hours of bench test. Pressure measurement accuracy of ± 0.09bar, temperature measurement accuracy of ± 3 degrees, sending wireless signal reception reliability 98%, CAN bus data transmission error-free frame.

Tire internal pressure to change the static simulation test. When the tire pressure is 25 percent higher than the standard pressure, or the tire pressure is less than the standard 12.5 percent, or when the rapid leakage (pressure per minute declined by more than 0.2Bar), TPMS will automatically issue an alarm message, start the voice of alarm, fault tire location and pressure value displayed in the instrument panel display. Less than the standard tire pressure when the pressure of 25% and 50%, TPMS will automatically issue a secondary and tertiary information alarm, warning of high frequency sounds.

6. CONCLUSIONS

The system is confirmed by the experiment, monitoring pressure and temperature data are accurate. TPMS is an important guarantee for traffic safety, the United States has passed legislation to force new vehicles to install TPMS system. As people learn more about TPMS system and improve the requirements of active safety for vehicles, TPMS will have a broad space for development and good market prospects.

REFERENCES

[1] Chen Yong, Li Quan, Li Yuan. SP12-based tire pressure sensor chip, real-time monitoring of system design [J]. The application of electronic technology，2007，33（7）：83~85

[2] TAN De-rong, Zhang Feng, Wang Yanyang. Signals based on the ABS Tire Pressure Monitoring System (TPMS) [J]. Agriculture equipment and vehicle engineering, 2007, (4): 16 ~ 18

[3] Cold-yi, Zhao Genbao, Li Xia, etc. Based on wireless sensors and CAN bus of the automatic identification of car tire pressure monitoring system [J]. Instrumentation technology and sensors, 2007, (6): 63 ~ 66

[4] Wu Bao, Zhou Qing, Yang Liu Bin. TPMS transmitter Reliability of products [J]. Mechanical design and manufacturing, 2008, (9): 230 ~ 232

[5] ZHAO Yong-ping, ZHANG,王启松. CRC algorithm in TPMS application [J]. Journal of Harbin Institute of Technology, 2008,40 (5): 759 ~ 762

[6] Li-xia, where min, cold-yi TPMS system applies to the low-frequency communications research [J]. Computer Engineering and Applications 2006,42 (17): 191 ~ 193

[7] Yue Huan, following a strong summer**.** MC68HC908EF2 by AT91M55800A and composition of TPMS [J]. Microcontroller and Embedded Systems, 2008, (1): 52 ~ 54

[8] Shao Beibei. Single-chip embedded applications online development methods [M]. First edition. Beijing: Tsinghua University Press, 2004:326 ~ 340

FORECASTING THE STATE OF METAL OXIDE VOLT-SENSITIVE PROTECTORS WITH IMPROVED RECURSIVE NEURAL NETWORKS

Shuang-de Zhang, Long Zhou
Wuhan Polytechnic University
Wuhan 430023, P. R. China
Email: zhangshd@whpu.edu.cn

ABSTRACT

The MOV is important extra-voltage protection equipment in electric power system. It is very important to forecast the state of electrical power equipment. In this paper, a method is put forward to forecast the Metal Oxide Volt-sensitive protectors (MOV's) state with the improved recursive neural network. Recursive neural network is a dynamic neural network, the basis of dynamic neural network learning algorithm is sequential partial differential coefficient. By comparison with the real example, it is better than the normal recursive neural networks algorithm. In order to realize the intellectual diagnosis and forecast of the MOV's state, we provide an effective method.

Keywords: Improved Recursive Neural networks, Metal Oxide Volt-sensitive Protectors (MOV), Forecast.

1. INTRODUCTION

In order to realize the intelligent diagnosis and forecast the electrical power equipment's state, an effective method is suggested. The MOV is an important extra-voltage protection equipment; its liability directly decides the running safety of electrical network. Because of the system voltage's effect, there is some leakage current passing the MOV. Therefore, it is important to supervise and diagnose the MOV on line, it will provide a method to avoid the malignant accident. [1]

Research results and working experience indicate that the majority of MOV's fault symptoms put up to be the increase of leakage-proof 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. [2] The Reference [3] provided a method using BP network to forecast the resistance leak current, and it resolves effectively the nonlinear forecasting problem of resistance current relatively, and provides a better method for pre-management and pre-examination of MOV.

Because the BP network is a static neural network, without taking into the consideration of the relative relations among samples in forecasting process. The static network only carries out nonlinear mapping between input and output, without dynamic feature. But the problem of MOV's state forecasting 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 MOV's state with the recursive neural network. And the result indicates that this method is effective. [4-7]

2. IMPROVED RECURSIVE NEURAL NETWORK AND ITS ALGORITHM

The recursive neural network is formed with the sequential system idea. The hidden units and output units' neurons in BP network are regarded, as a sequential system, 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 to it. [8-10]

Suppose that $\{z_1, z_2, \cdots, z_i, \cdots, z_j, \cdots, z_n\}$ is a set of variables, if z_i is only the function of variable set $\{z_1, z_2, \cdots, z_{i-1}\}$, then the variable set is called sequential set. In order to be distinguished from the common partial differential coefficients, $\partial^+ z_j / z_i$ represents the z_i 's partial differential coefficients towards z_i, and the constant is $\{z_1, z_2, \cdots, z_{i-1}\}$, and the variable is $\{z_1, \cdots z_j, \cdots, z_n\}$, the sequential partial differential coefficients have two properties defined as follow:

$$\frac{\partial^+ z_{i+1}}{\partial z_i} = \frac{\partial z_{i+1}}{\partial z_i} \tag{1}$$

$$\left.\begin{aligned} & when \cdot j > i, \frac{\partial^+ z_j}{\partial z_i} = 0 \\ & when \cdot j > i+1, \frac{\partial^+ z_j}{\partial z_i} = \frac{\partial z_j}{\partial z_i} + \sum_{k=i+1}^{j-1} \frac{\partial^+ z_j}{\partial z_k} \bullet \frac{\partial z_k}{\partial z_i} \end{aligned}\right\} \tag{2}$$

In the dynamic recursive neural network, the relationship among the training samples is shown between the neurons of the network structures, that are 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.

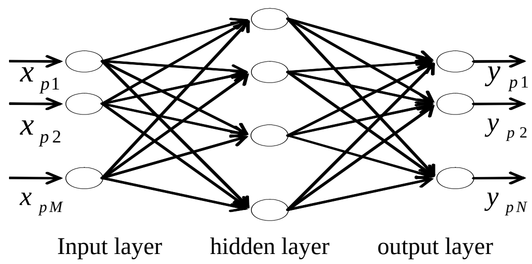


Figure 1. Three-layer Recursive Neural Network Structure

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 as follows:

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):

$$\left.\begin{aligned} & net_{pk} = \sum_{i=1}^{k-1} W_{ki} O_{pi} + b_k \\ & M+1 \le k \le M+k \\ & O_{pk} = f(net_{pk}) = 1/\left[1+\exp(-net_{pk})\right] \end{aligned}\right\} \tag{4}$$

The input and the output of the output node are respectively as follows:

$$\begin{aligned} & net_{pj} = \sum_{k=M+1}^{k-1} W_{jk} O_{pk} + b_j \\ & M+k+1 \le j \le M+k+N \\ & y_{pj} = f(net_{pj}) = 1/\left[1+\exp(-net_{pk})\right] \end{aligned} \tag{5}$$

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

$$E_p = \frac{1}{2} \sum_{j=1}^{N} \left(y_{pi} - y_{pj}\right)^2 ,$$

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

$$\begin{aligned} \delta_{pj} &= \frac{\partial\ E_p}{\partial net_{pj}} = \frac{\partial E_p}{\partial net_{pj}} \bullet \frac{\partial y_{pj}}{\partial net_{pj}} \\ &= (d_{pj} - y_{pj}) \bullet y_{pj} \bullet (1 - y_{pj}) \end{aligned}$$

When $M+k+1 \le j < M+k+N$,

$$\frac{\partial net_{pr}}{\partial net_{pj}} = \frac{\partial net_{pr}}{\partial y_{pj}} \bullet \frac{\partial y_{pj}}{\partial net_{pj}} = W_{rj} y_{pj} (1 - y_{pj}),$$ so

$$\begin{aligned} \delta_{pj} &= (d_{pj} - y_{pj}) \bullet y_{pj} \bullet (1 - y_{pj}) + \\ & y_{pj} (1 - y_{pj}) \sum_{r=j+1}^{M+k+N} W_{rj} \delta_{pr} \end{aligned}$$

- As for the connection weight W_{kj} and the threshold b_k between hidden node and input node, we can get that

$$\left.\begin{aligned} & \Delta_p W_{kj} = \eta \delta_{pk} O_{pi} \\ & \Delta_p b_k = \eta \delta_{pk} \\ & where \cdot \delta_{pk} = O_{pk} \bullet (1 - O_{pk}) \sum_{j=k+1}^{M+k+N} W_{jk} \delta_{pj} \end{aligned}\right\} \tag{7}$$

The training learning process mentioned above indicates that the algorithm of recursive neural network embodies the idea of time sequence, without taking into the consideration of 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. The learning speed is compared between normal recursive neural network and improved recursive neural network in Fig.2. From the Fig.2, learning speed of improved recursive neural network is much faster than normal recursive neural network.

3. THE STATE FORECASTING OF MOV

The actual running theories and experience indicate that the main reason, which causes the increase of MOV'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 forecasting 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 affect the forecasting 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 Figure 1, the inputs are leakage-proof 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 MOV, 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:

$$\Delta W(t) = -\eta \frac{\partial E}{\partial W(t)} + \alpha \Delta W(t-1) + \beta \Delta W(t-2) \tag{9}$$

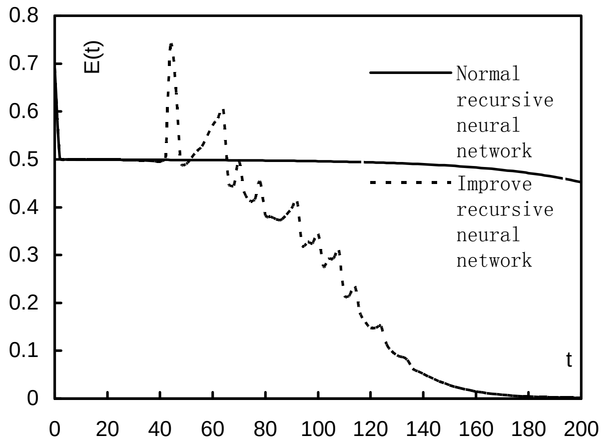


Figure 2. Learning Speed Comparison between Normal Recursive Neural Network and Improve Recursive Neural Network

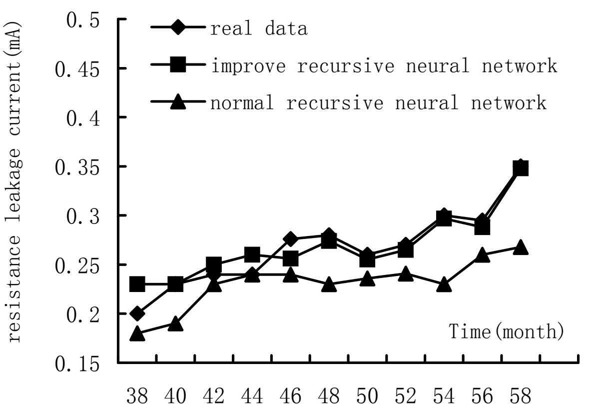


Figure 3. Comparison between the Actual Value and Forecast Value of Lleakage-proof Current

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 $(t = 0,1,\cdots)$, because the network structure is not complex, one-order momentum is only added to revise the weight, $\alpha = 0.6$, $\eta = 0.8$, the results are shown in Fig.3.

In order to further compare them, we use the equation (8) to compute the forecasting precision of the two methods, forecast precision of normal recursive network algorithm is 0.08552, but that of improved recursive network is 0.00235, the result indicates that recursive network is more adapted to the state forecasting of MOV.

4. CONCLUSIONS

In this paper, a method is put forward to forecast the MOV's state with the improved recursive neural network. The result indicates that recursive network is more adapted to the state forecast of MOV. Because the running state of MOV is closely related to the system voltage and the environment, the state forecasting method affected by multi-factors should be further considered.

5. ACKNOWLEDGEMENTS

This paper is supported by National Natural Science Foundations under Grant 69874018 of China, Hubei Provincial Science Foundation under Grant 2003ABA053, Hubei Provincial Department of Education under Grant Q200618002.

REFERENCES

[1] Li Xueshi et al., The finding analysis on high voltage metal oxide arrester's condition. Insulators and Surge Arresters, 1992(2)，46-48.

[2] Linyi, The accident analysis and prevention on 110kv and upwards of MOV in China. Power System Technology,1995(3)，57-58.

[3] Zhou long et al., The forecast method of MOV's condition based on neural network, High Voltage Apparatus,1997(3)，78-80.

[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,545~549,1992，37-39.

[6] Parlos A G et al., Application of the Recurrent Multilayer pereeption in Modeling Complex Process Dynamies, IEEE Trans. Neural Networks,5(2),2003，40-43.

[7] Huazhong electrical power testing institute, The analysis on spoil reason of 500KV metal oxide arrester in Gezhouba.1994(11)，66-68.

[8] Ku C C,Lee K Y.System,identification and control using diagonal recurrent neural networks.Proc American Control Conf.Chichago,545~549,1992。

[9] Parlos A G et al., Application of the Recurrent Multilayer pereeption in Modeling Complex Process Dynamies, IEEE Trans. Neural Networks,5(2),2003，17-19.

[10] Huazhong electrical power testing institute, The analysis on spoil reason of 500KV metal oxide arrester in Gezhouba.1994(11)，25-27.

1.

INTEGRATED SERVICE AND DESKTOP GRIDS FOR SCIENTIFIC COMPUTING

Robert Lovas[1], Tamas Kiss[2]

[1]Computer and Automation Research Institute, Hungarian Academy of Sciences
Kende u. 13-17, Budapest, 1111, Hungary
Email: rlovas@sztaki.hu
[2]Centre for Parallel Computing, University of Westminster
115 New Cavendish Street, London, W1W 6UW, UK
Email: kisst@wmin.ac.uk

ABSTRACT

Currently two different approaches can be distinguished in the research and development of Grid systems. In case of traditional *Service Grids* (such as operated in the EGEE project), the infrastructure consists of several well maintained clusters and data centers, which are provided by institutes and universities in order to serve the demands of large user communities. On the other hand, *Desktop Grid* solutions —such as Internet-based distributed or volunteer computing infrastructures— usually collect non-reliable resources from the donors (desktop PC owners) for some selected, grand challenge projects. In this paper we discuss a generic bridge solution between these different computation platforms, and demonstrate its capabilities and advantages by applications. As one of the key issues, the software build, test, and validation procedures in such heterogeneous environment will be outlined based on the ETICS-2 services as well. The presented solution tries to offer best practices for other Grid initiatives and projects; how to tackle the digital fragmentation with gateways, quality assurance and reliability issues concerning scientific Grid systems.

Keywords: Desktop grid, Service grid, Gateway, Applications, Build and test procedures, Validation, Software quality

1. INTRODUCTION

The targets of the Enabling Desktop Grids for e-Science (EDGeS) project [1] are user communities that require large computing power not available or accessible in current scientific e-Infrastructures. In order to support the specific needs of these research communities and commercial partners the consortium interconnected the largest European Service Grid infrastructure (EGEE) with several existing Desktop Grid systems based on the widespread BOINC [3] or XtremWeb [6] solutions.

Service Grids (SG) are more flexible and can accommodate a broader variety of applications than Desktop Grids, however, their setup and maintenance require more efforts, highly skilled IT specialist, and dedicated resources. On the other hand, Desktop Grids (DG) are currently restricted solely to a subclass of compute-intensive applications but these easy-to-scale systems are able to collect one or sometimes even two orders of magnitude more computational power by utilizing the involved spare and volunteer IT resources at a fraction of the cost. Making a bridge between these two types of Grid systems will enable the users to transparently execute applications on any arbitrary platform involved in the new infrastructure. Taking the advantages of both approaches the EDGeS infrastructure can represent a step towards a worldwide scientific grid where extremely large number of resources could be integrated to support grand-challenge scientific and other applications.

* This research work is supported by the EDGeS and ETICS-2 projects funded by the European Commission under contract numbers INFSO-RI-211727 and INFSO-RI-223782.

2. RELATED WORKS AND THE CORE COMPONENT: 3G BRIDGE

In order to understand our bridge solution between Service and Desktop Grids, we have to take into consideration and distinguish two approaches.

The ***superworker*** approach, proposed by the Lattice project [4] and the SZTAKI Desktop Grid [5], is the first solution. The superworker is a bridge between the DG server and the SG resources. From the DG server point of view, the Grid or cluster appears as one single resource with large computing capabilities. The superworker continuously fetches tasks or work units from the DG server, wraps and submits the tasks accordingly to the local Grid or cluster resources manager. When computations are finished on the SG computing nodes, the superworker sends back the results to the DG server. Thus, the superworker by itself is a scheduler which needs to continuously scan the queues of the computing resources and watch for available resources to launch jobs.

Since the superworker is a centralized agent, this solution has several drawbacks: performance bottleneck, single point of failure in the system, latency. On the other hand, the superworker approach does not require modification of any connected infrastructure but e.g. the proper delegation of security credentials is necessary.
The ***gliding-in*** approach to cluster resources spread in different Condor pool using the Global Computing system (XtremWeb) was first introduced in [6]. The main principle consists in wrapping the XtremWeb worker as regular Condor task and in submitting this task to the Condor pool. Once the worker is executed on a Condor resource, the worker pulls jobs from the DG server, executes the XtremWeb task, and returns the result to the XtremWeb server. As a consequence, the Condor resources communicate directly to the XtremWeb server. Mechanisms similar to those described just above are now commonly employed in Grid Computing [7].

The gliding-in or pilot job approach has several advantages; simple, fault tolerant, and efficient load balance, and good performance. On the other hand, e.g. strict firewall settings can block the direct communication, which is necessary for this approach.

The EDGeS 3G bridge encompasses the superworker and gliding-in approaches, and four different bridges in order to integrate EGEE, BOINC and XtremWeb systems: BOINC → EGEE, XtremWeb → EGEE, EGEE → BOINC, EGEE → XtremWeb. Instead of developing four different solutions for the four bridges, the project partners rather created a Generic Grid to Grid (3G) bridge that can be easily adapted for the different cases of the required bridges. In fact, the 3G bridge is generic enough to easily adapt not only for the three Grid systems we tackle in EDGeS, but also for other Grid systems like GT2, GT4, Xgrid, etc.

The basic architecture of the 3G Bridge and its application in EDGeS is shown in Figure 1.

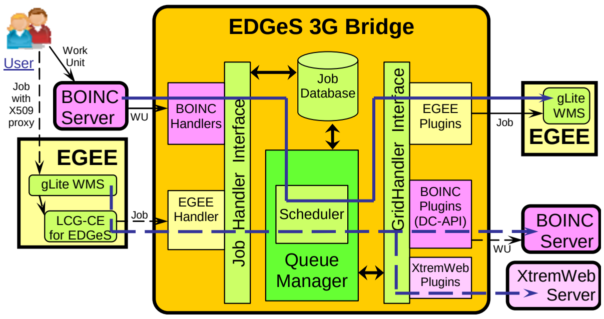


Figure 1. EDGeS 3G Bridge

The Job Database stores DG Work Units and SG jobs as generic descriptions. The Source Grid Handler Interface is used to place DG work units and SG jobs into the Job Database and to query their status. The Source Grid Handler Interface is implemented via MySQL and as such it accepts SQL queries, inserts and updates. Besides the task of placing generic descriptions into the Job Database, its other task is to get job status information from it. Jobs/WUs coming from various source Grids are received by specific handlers that transfer the incoming jobs/WUs to the Source Grid Handler Interface. So, in order to connect a Grid as a source Grid to the 3G Bridge, a Grid-specific handler should be written for this source Grid.

The Queue Manager periodically reads jobs from the Job Database and transmits them to the Target Grid Plug-in Interface. The Target Grid Plug-in Interface enables to connect various target Grids via their plug-in. The Target Grid Plug-in Interface provides a generic set of interface functions that should be implemented by the target Grid plug-ins. In order to connect a Grid as a target Grid into the 3G Bridge, a Grid-specific plug-in should be written for this target Grid. Note that the Grid plug-in is also responsible for querying the status of job/WU execution in the target Grid and retrieving the output of submitted jobs/WUs.

3. INFRASTRUCTURE

The EDGeS 3G bridge is in operation between EGEE and Desktop Grids, as shown in Figure 2.
The DG → EGEE bridges of the EDGeS system have been prototyped in June 2008 and put into operation in September 2008.

The BOINC → EGEE bridge is currently in operation at MTA SZTAKI (Hungary). It connects the SZTAKI Desktop Grid (Hungary), the UoW Desktop Grid (UK), and Correlation Systems Ltd's Grid (Israel) to a dedicated external VO of EGEE.

The XtremWeb → EGEE bridge is now included in the standard distribution of the XtremWeb middleware. It is in operation at the IN2P3 and INRIA Desktop Grids (France), and at the AlmereGrid Desktop Grid (Netherlands).

In the operational DG → EGEE infrastructure of EDGeS, the EDGeS VO currently has access to three computing elements as EGEE resources: CNRS/IN2P3 (France), SZTAKI (Hungary), CIEMAT (Spain).

It also contains the basic EGEE core services like BDII, VOMS, MyProxy, WMS and LB.
The EGEE → DG bridges of the EDGeS system have been prototyped in December 2008 and put into operation for the EGEE users in March 2009.

In the operational EGEE → DG infrastructure of EDGeS, the EDGeS VO currently has access to eight desktop Grids.

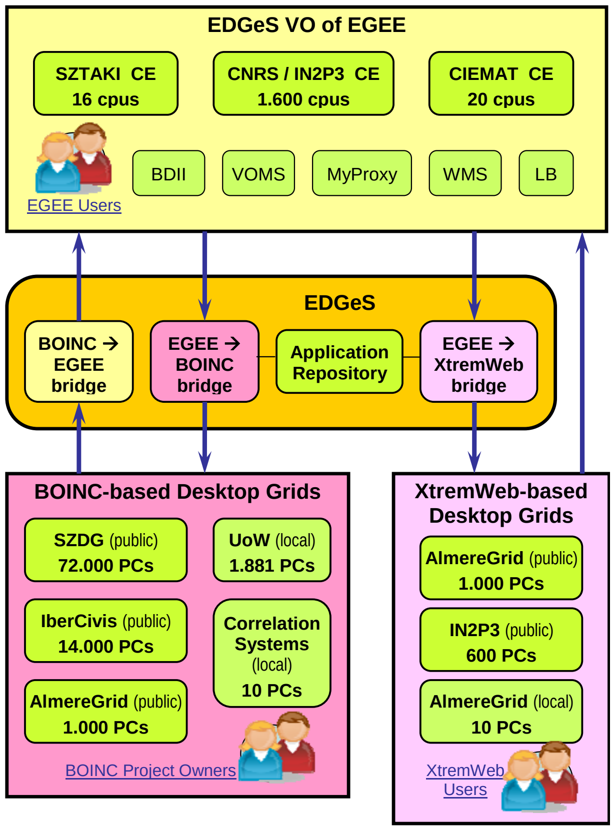


Figure 2. EDGeS Infrastructure

4. APPLICATIONS

As the most important goal, EDGeS ports existing applications to the integrated infrastructure and provides a seamless job execution mechanism among the interconnected Service and Desktop Grid systems. **Table 1** summarizes the already ported applications from various areas; e.g. biotechnology, signal and image processing, or high energy physics.

Table 1. EDGeS Applications

Applications already ported to EDGeS	Organisation	Platform (DG/SG)
Video Stream Analysis in a Grid Environment (VISAGE)	Correlation Systems Ltd. Israel	Both

Digital Alias-free Signal Processing	Uni. of Westminster, UK	Both
Protein Molecule Simulation using Autodock [9]	Uni. of Westminster, UK	Both
E-Marketplace Model Integrated with Logistics (EMMIL)	MTA SZTAKI, Hungary	Both
Anti-cancer Drug Design (CancerGrid) [10]	MTA SZTAKI, Hungary	Both
Cellular Automata based Laser Dynamics (CALD)	Uni. of Seville and Uni. of Westminster, UK	Both
Signal and Image Processing using GT4 Tray	Forschungs-zentrum Karlsruhe, Germany	DG
Analysis of Genotype Data (Plink)	Atos Origin, Spain	DG
Distributed Audio Retrieval using TRIANA (DART)	Cardiff University, UK	Both
Fusion Plasma Application (ISDEP)	BIFI, Spain	DG
3-D Video Rendering using Blender	Uni. of Westminster, UK	Both
Profiling Hospitals in the UK based on Patient Readmission Statistics [9]	Uni. of Westminster, UK	Both

As an example; the main purpose of the *3-D Video Rendering* application is to create images and videos from models that were designed in Blender, which provides a broad spectrum of modeling, texturing, lighting, animation and video post-processing functionality in one package. Through its open architecture Blender provides cross-platform interoperability as well as support for broadcast and cinema quality videos. On the other hand, the rendering process is computationally expensive. In order to reduce the execution time the application has been ported to the EDGeS infrastructure based on the EDGeS Application Development Methodology (see Section 5). The new version has potential benefits for very large user communities (including artists, designers, etc.) thus, it has been selected as one of the supported applications. The result of testing is discussed in Sections 6 and 7.

5. EDGES APPLICATION DEVELOPMENT METHODOLOGY

Contrary to the generic software development methods, the EDGeS Application Development Methodology (EADM) [9] focuses on the specific requirements of porting applications to a combined SG/DG platform. The EADM identifies well defined stages that have a suggested logical order. However, the overall process is in most cases non-linear allowing revisiting/revising the results of previous phases at any point.
The *Analysis of current application* phase describes the existing application in detail. This phase identifies the target user community, the problem domain, and the typical use cases and functionalities of the system. It also captures technical characteristics, such as the type of computing platform and the way of parallelism (if any) utilized by the current application, data access volume and methods, memory and hard disk usage, programming language, operating system, or security solutions. An Application Description Template has been developed to capture the above information mainly from the operators of the existing application.

The aim of the *Requirements Analysis* stage is to identify how the target user community will benefit from porting the application to the SG/DG platform. The requirements towards the ported application concerning efficiency of execution and data access are analyzed from a user perspective. The target computing platform (either SG or DG) that the user wants to access as an entry point when executing the application, and the desired user interface are also identified in the User Requirement Specification document.

The *Systems Design* phase outlines the proposed structure of the system. The target computing platform, the type of user interface, and the parallelization and data access principles are designed taking both user requirements and technical feasibility into consideration. The outcome of this stage is a Systems Design Specification that identifies at a high-level how the ported application will work regarding the above aspects. The use of structural diagrams, such as a system block diagram and UML diagram techniques is highly recommended at this stage.

The *Detailed Design* stage provides class level specification of any required modification in the original application when porting it to the SG/DG platform. This stage results in a Technical Design Specification that forms the basis of the *Implementation*. EADM recommends specific tools for the detailed design and implementation phases that were specifically created or enhanced to serve SG/DG migration tasks. These tools [1] include the Distributed Computing (DC-API) and the XtremWeb APIs for application development, an extended version of the P-GRADE Grid portal to support the transparent exploitation of the SG/DG infrastructure at workflow level, and the GEMLCA legacy code repository to deploy validated applications for end-users.
During *Testing* both the functionalities and the performance of the ported application are evaluated and compared to the identified user requirements.

The aim of the *Validation* phase is to assure that the application causes no harm to the computers of Desktop Grid donors and also that it conforms to the generic aims of the target Desktop Grid system. This stage is inevitable in order to deploy the application on a DG platform where individuals or institutions offer their volunteer resources for the computation.
For the above described implementation, validation, and testing phases the EDGeS project has started utilizing some high level state-of-the-art solutions as they are described in Section 6.

Finally, the application is published in an application repository and deployed to be utilized by end-users. It is also desired to give organized support for end-users after deployment to support and maintain the application.
The major motivation behind the specification of the EADM was to identify the most important players and aspects when porting applications to an SG/DG platform.

6. BUILD, TEST, AND VALIDATION

The ETICS-2 [8] project provides an international e-Infrastructure with more than 450 CPU cores and 15 different platforms to support the build, test and certification of complex Grid and distributed software. The consortium developed a software Quality Certification Model (A-QCM) to evaluate and monitor the quality of software according to current standards. ETICS-2 offers software applications and tools either in the form of web applications running on its own servers or command-line tools installed on user computers to perform the configuration, build, test and quality analysis of software code developed by the users. The project partners provide support in the usage of the applications and tools, in the integration of the tools in the user software engineering procedures and in the development of user-customized tools and procedures.

Combining the efforts of EDGeS and ETICS-2 in some areas can be mutually advantageous. On one hand, the success of EDGeS project depends partially on the software quality and reliability of the Desktop Grid enabled applications to be executed on donor’s computers, which provide a highly heterogeneous computational platform from PCs with limited capacities. In the EADM (see Section 5) during the testing and validation phases of these applications the EDGeS application support teams can rely on well-established and automated tools in order to build, test, and validate the mostly third-party applications (see Section 4) before their execution on the production DG infrastructures of EDGeS (see Section 3). The ETICS-2 services help accelerate and make more efficient these essential steps in the EDGeS application validation procedure. On the other hand, the migration of some key components of EDGeS Grid middleware, such as SZTAKI Desktop Grid and 3G Bridge (see Section 2), under the ETICS-2 Build and Test system can ease the build, deployment, test on a heterogeneous multinode environment, and maintenance of these components. Moreover, obtaining an A-QCM certificate for the developed Grid middleware core components could contribute to the successful promotion of EDGeS achievements and solutions towards other industrial partners. On the other hand, the ETICS-2 System efficiency, correctness and completeness depend on the involvement of software engineering professionals and of project developing distributed software for as many communities and applications as possible; EDGeS can contribute to these efforts.

7. MULTINODE DEPLOYMENT AND PERFORMANCE TESTS

One of the major goals of ETICS-2 is to provide a distributed complex build and test framework, as well as a Workflow Designer as part of the ETICS-2 test services [12]. In that respect, the new Workflow Designer [11] –developed by MTA SZTAKI based on the WS-PGRADE portal [10]– extends the existing distributed testing mechanisms with advanced functionalities by supplying a high level front-end interface being able to orchestrate distributed deployment and testing scenarios in order to reduce the complexity of designing these complex tests for applications.
Two main use cases of the testing procedure have been identified in ETICS-2 in order to help testers to define complex methods more efficiently with high level workflow tools:

(1) multi-node test design, when the required services (see *workflow nodes* on Figure 3 as an example), are deployed on different machines, where the data exchanges between different phases of complex multi-node deployment on different nodes is crucial (see the *flow connectors* represented by small numbered rectangles, and the arcs between workflow nodes on Figure 3)
(2) test process design, when the focus is on the orchestration of different testing tools, and where the dependencies and the necessary exchange of reports are crucial among registered and deployed test tools/plugins

The reason behind these scenarios is the evolution of web and Grid applications: the software is becoming more and more distributed and systems require a complex deployment and test services that must interact, communicate and synchronise among themselves.

In that sense, distributed testing is different comparing to traditional one node testing and simultaneous testing, because a distributed test case consists of two or more parts that interact, communicate and synchronise with each other. Each part is being processed on a different node. All of the test cases processed on all of the different nodes contribute towards a single common result.
With that in mind, when setting up complex multi-node deployments, a series of issues regarding the synchronisation, monitoring and information exchange should be resolved.

In order to make more efficient the testing and validation phases in the described EADM, a multinode deployment test of any Desktop Grid application, such as the 3-D Video Rendering (see Section 4), can be specified at high-level with a workflow as shown in Figure 3.

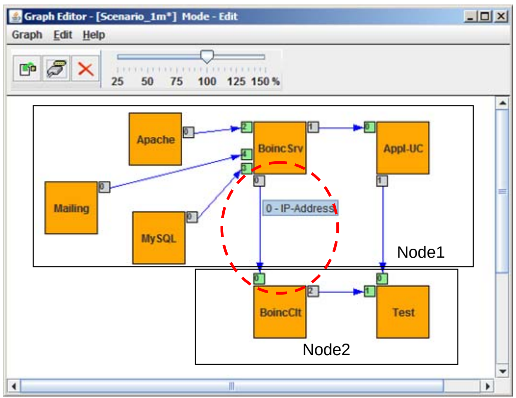


Figure 3. Multinode Test for Desktop Grid Application

In this example the application tester can set various parameters, e.g. the *Node 1* will require a Linux node, and software component on *Node 2* will have to be deployed on a Windows node (both automatically provided by ETICS-2 from its distributed virtual machine-based resource pool). On these computers, the following services must be deployed:

BoincSrv: BOINC server (developed by Uni. of Berkeley)

- Packaged by MTA SZTAKI for Debian
- Required services (dependencies): **MySQL**, **Apache**, **Mailing**

Appl-XX: any BOINC-based DG application (developed by third-party application providers – see Section 4):

- Client side must be compiled on various platforms, master side only for the server using ETICS-2 build services
- Must be registered in the BOINC server
- Requires optionally: DC-API (for the more user-friendly programming of desktop Grid applications, developed by MTA SZTAKI)

BoincClt: BOINC Client (developed by Uni. of Berkeley)

- Available for various platforms as source/binary package
- The client has to connect to a server; it requires the IP address of the server, which must be passed in run-time (this is one of the exchanged parameters between nodes, see Figure 3)

Test: Test component (developed by EDGeS support team)

- Responsible for joining a DG project/application with the client, i.e. register the client on the server, in order to fetch work units belonging to the application (the project name is another exchanged parameter between nodes)
- Validation of the entire test scenario and the application based on the EADM tester's expectations; e.g. check the successful processing of a certain number of work units without failure and, as more important, the possible harmful or unwanted effects on the node (suspicious memory/disk consumption, etc).

The presented workflow can be easily scaled up; more (other) nodes with different platforms and more applications can be involved in the tests even simultaneously based on the abstract and concrete workflow facilities of the WS-PGRADE portal.

Beside the multinode deployment tests, the performance tests of the 3-D Video Rendering application have been also executed on the Westminster Local Desktop Grid of the University of Westminster (see Section 3). At the time of the testing more than 700 computers were continuously available. Most of the computers have a dual core processor, i.e. from the BOINC server's point of view they behave like more computers because they fetch and can compute more work units at the same time. Note that the measured speedups are approximated. All of the test cases (except of the longest one) were executed 3 times. Table 2 summarizes the median test results; the gridification was relatively successful. The application has been tested and monitored using the 3G bridge in the BOINC→EGEE direction, too.

Table 2: Performance Test Results

Number of frames	Number of work units	Duration (minutes)	Number of PCs	Speed up
100	10	38	6	*3.95*
1.000	100	65	58	*23.08*
10.000	1.000	161	268	*93.17*
100.000	10.000	409	691	*366*

8. CONCLUSIONS AND FUTURE WORKS

The involvement of low-cost volunteer Desktop Grids into the scientific Grid infrastructure contributes to the establishment of sustainable Grid infrastructures. Currently the EDGeS infrastructure and the 3G bridge are in production level operation, and demonstrated with several applications. However, during the utilization of such new and combined platforms the application developers face challenges and require high-level methods and tools in order to tackle several problems. According to our first experiences, the EDGeS Application Development Method and its tools together with ETICS-2 build and workflow-based test facilities can provide efficient solution for these efforts. As the future work, we plan to apply these new methods in case of further application, and finalise the workflow-based solution in order to release the tool to the wider Grid application developer community. To obtain an A-QCM certificate for the bridge is also scheduled.

REFERENCES

[1] Z. Balaton, et al.: EDGeS: the Common Boundary between Service and Desktop Grids. Parallel Processing Letters 18(3), 2008, pp. 433-445

[2] EGEE Project website: http://www.eu-egee.org

[3] D. Anderson: BOINC: A System for Public-Resource Computing and Storage. In Proceedings of the 5th IEEE/ACM International GRID Workshop, Pittsburgh, USA, 2004.

[4] D. S. Myers, A. L. Bazinet, and M. P. Cummings. Expanding the reach of Grid computing: combining Globus- and BOINC-based systems, chapter Grids for Bioinformatics and Computational Biology. Wiley Book Series on Parallel and Distributed Computing, 2008.

[5] Z. Balaton, et al.: SZTAKI Desktop Grid: a modular and scalable way of building large computing grids. In Proc. of the 21th International Parallel and Distributed Processing Symposium, 26-30 March 2007, Long Beach, California, USA, 2007.

[6] O. Lodygensky, et al.: XtremWeb & Condor: Sharing Resources Between Internet Connected Condor Pools. In Proceedings of CCGRID'2003, Third International Workshop on Global and Peer-to-Peer Computing (GP2PC'03), Tokyo, Japan, 2003, pages 382–389

[7] Douglas Thain and Miron Livny. Building reliable clients and services. In The GRID2, Morgan Kaufman, 2004, pages 285–318

[8] A. D. Meglio, et al.: ETICS: the international software engineering service for the grid, Journal of Physics: Conference Series 119, 2008, pp. 042010

[9] T. Kiss, et al.: Porting Applications to a Combined Desktop Grid/Service Grid platform using the EDGeS Application Development Methodology, INGRID 2009 workshop, Sardinia, Italy, 2009

[10] P. Kacsuk, et al.: WS-PGRADE: Supporting parameter sweep applications in workflows, Proceedings of the Third Workshop on Workflows in Support of Large-Scale Science, Austin, TX, 2008, pp. 1-10

[11] DJRA2.4 - User Interface and Test Execution Generator (Prototype), ETICS-2/JRA2 project deliverable, https://edms.cern.ch/document/990909/1.0

[12] L. Dini: Technical note on multinode tests specification (ETICS-2/SA1 project document), https://edms.cern.ch/document/964202

Dr. Robert Lovas is the deputy head of the Laboratory of the Parallel and Distributed Systems, Computer and Automation Research Institute of the Hungarian Academy of Sciences. He received his MSc and PhD degrees at the Budapest University of Technology and Economics, Hungary. He is a member of the Technical Committee of the Hungarian Grid Competence Center, the Grid Application Support Center (GASuC), the deputy member of the Project Management Board of the European ETICS-2 consortium, the Scientific Steering Committee of the CANCERGRID project; and the activity leader of the European EDGeS research infrastructure project. From 1997 he has been involved several national (HAGRID, WEB2GRID), intergovernmental bilateral, and European research projects and networks in the Framework Programmes 4-7 (EGEE projects, CoreGrid – EU Network of Excellence). In the HARNESS project (US/DoE) he worked as an exchange scientist at the Department of Math and Computer Science, Emory University, Atlanta. He has long-running experience in ICT collaborations with academic organizations, universities, and enterprises from various application areas of meteorology, biotechnology, telecommunication, and computational chemistry. He is a co-author of more than 35 scientific papers on parallel software engineering and Grid computing particularly from design, debugging and application aspects.

PARALLELIZATION STRATEGIES FOR DISTRIBUTED GRID COMPUTING

Souheil Khaddaj
Faculty of Computing, Information Systems and Mathematics, Kingston University,
Kingston upon Thames, KT1 2EE, UK.
E-mail:S.Khaddaj@kingston.ac.uk

ABSTRACT

Grid and distributed computing systems have been used successfully in many different applications that range from numerical modelling to web applications. However, in order to maximise the benefits of using such systems different partitioning and parallelisation strategies must be employed, which depends on the applications as well as the underlying architectures. In order to demonstrate the various strategies, examples based on modelling and discrete event stochastic simulation techniques commonly used in science and engineering are considered. The problems that are encountered in each case are discussed, and the advantages and disadvantages of using the different strategies are presented.

1. INTRODUCTION

In the last few decades there have been substantial improvements in computer performance, due not only to advances in hardware but also to innovations in computer architectures which has resulted in the advancement of parallel computer architectures, wherein the processors are replicated and organised such that they can act in unison on one application. However, there are still many complex scientific and engineering problems which are beyond the current generation of parallel supercomputers. This has led to the consideration of a very large number of heterogeneous geographically distributed resources to act as a powerful computing engine, an approach which is now known as Grid Computing. The universal acceptance of this technology has been slow for a number of reasons, including the wide differences of the available parallel architectures and grid models, which mean that there is no unifying software/hardware environment [1,6]. Moreover, the performance of parallel algorithms depends on the parallelisation strategies which are tied to the target parallel architectures [3].

It is widely accepted that the efficient use of distributed architectures depend very much on the adopted parallelization and decomposition strategies. In this paper we describe a number of these strategies which have the potentials to present a "natural" decomposition or "natural" parallelism to each specific implementation. The term “natural” is used since in many cases more that one type can be applied but one seems to be natural. In order to discuss and analyse the suitability of different strategies a number of discrete event stochastic simulations are considered which are very attractive for simulating and tracking particles and atoms in many areas of physics, atmospheric, ocean surface modelling etc., and can offer good experimental environments for various parallel strategies [4, 5, 10].

The organization of this paper is as follows. We start with a general description of the parallelisation and decomposition strategies. We follow this with a presentation of an event based strategy with its Monte Carlo implementation and observed results. Similarly, we then describe a geometric decomposition strategy. Finally we present some conclusions.

2. PARALLELISATION AND DECOMPOSITION STRATEGIES

Parallelism is an intrinsic feature of many physical systems [2, 9]. With the majority of high performance computers making use of this concept, the development of corresponding algorithms, and the interaction between parallel architectures and parallel algorithms development, become extremely important for the efficient use of the available hardware. However, many simulation problems that map into distributed architectures follow one or more of the following strategies:

Data Decomposition: this represents the simplest form of decomposition since it is based on the application of a multi-serial approach where the same serial/sequential code is running independently on different nodes dealing with different data/parameters with no communications although the overall picture might need to be reconstructed for some applications. This approach is commonly used in many grid applications.

Event Decomposition: is another relatively simple form of decomposition where also the same serial/sequential code running on different nodes, but requires occasional communications, perhaps at synchronisation points, and with each node dealing with its own allocated data.

Geometric Decomposition: using this approach each node is executing more or less the same code but now the data is distributed according to some geometric/space decomposition and in such a manner which require extensive communications between the nodes, for example each node might be used to simulate one part or more of a large system of similar objects interacting with each other. This approach is commonly used on closely coupled systems with dedicated communication hardware since it usually involves intensive communications and many synchronisation points.

In this paper we mainly concerned with the parallelisation strategies using two types of discrete event simulation which are considered from the areas of particle and quantum physics [7, 11]. The cases have been chosen to represent two different schemes of Monte Carlo based simulations which typically require high performance computers in order to achieved any reasonable results with good statistical properties. In the first case during the simulation individual particles are followed independently and the scheme is inherently parallel. Here, a number of particles are attached to each node and the process farm strategy is used. In the second case, the simulation of material growth it is natural to divide the surface, where atoms

are deposited and growth occurs, into distinct spatial regions, each being assigned to a particular node.
However, the stochastic nature of the problems and the types of parallelism that are used introduce some new problems and features that do not occur for sequential computers. These include time synchronization, making careful use of the fault tolerant nature of the diffusion process to obtain the correct dynamical evolution, and using ideas from queuing theory to control information flow across boundaries of neighboring regions.

3. EVENT BASED DECOMPOSITION

In this case the Monte Carlo simulation is used for solving three dimensional problems in particle and reactor physics. Conventional Monte Carlo code simulate the history of the particles by tracking the particles, one at a time through the problem geometry, and by random sampling from the probability distribution functions that are considered to represent the collision physics of the particles' interactions with the media [12].

In the particular form of the Monte Carlo method there is a close analogy between the physical particles, and the 'mathematical' particles followed by the program. The code simulates the history of a large number of particles from 'birth' to 'death', one particle at a time (i.e., sequentially). The particle is 'killed' if it escapes from the system, or, due to successive scatters, its energy becomes less than a preset minimum; in either event, a new particle is started from the source. The only sophistication employed is the concept of survival weight, by means of which the effect of absorption is accounted for by modifying the particle weight after each collision. In order to accelerate the termination of the unimportant (i.e. low weight) particles, when a particle weight becomes less than a pre-assigned minimum value, a game of Russian Roulette is played to determine whether or not the particle survives; should it survive, its weight is increased appropriately so that the resulting scores are unbiased [13].

3.1. Parallelisation and Decomposition

The conventional Monte Carlo simulation is inherently parallel due the statistical independence of particle histories. Even in some applications where the particle histories are not independent, such as where the particles might interact with background medium and change its properties, the histories can be still be treated independently within sufficiently small time steps. Therefore, if one follows the physics and uses a "natural" partitioning of particles to nodes, the inherent parallel nature of the physical process will be manifested in the algorithm. On the other hand, this mandates that each node know the entire geometrical mesh through which the particles are moving which can lead to a prohibitive memory demand.

3.2. Parallel Procedure

Since, in this type of Monte Carlo computation, the life history of the individual particles must be scored independently, the processor arm strategy is well suited to this type of computation. A straightforward implementation of serial fixed-source particle transport Monte Carlo code can be made subject to two main constraints:

1) Enough memory is available to accommodate the whole program on each node.

2) The ability to generate independent random number sequences on each node. Of course, these sequences of random numbers must have suitable statistical properties within themselves.

For the implementation of this simulation the relationship between the master and the slaves is relatively straightforward where the master reads in the basic data for the calculation and passes it down to the slaves, then the master and the slaves carry the computation and the results are then combined. It is clear from the above that little communication occurs during the parallel implementation, just when data are passed to the slaves and when the results are combined (global communication with the master), when all calculations have finished, thus a blocking synchronous communication is used. In order to keep the work load balanced between the nodes an equal number of particles are followed on the master and slaves.

3.3. Results and Observations

The variation of the speedup factor achieved with the number of nodes used is shown in Figure 1. This shows that the speedup factor varies linearly when the number of nodes is varied between 1 to 8, and we would expect it to remain so even if the number of nodes were considerably increased. This is because the processor farm strategy requires minimal interprocessor communication between master and slaves. Figure 2 demonstrates that the efficiency achieved is over 95%.

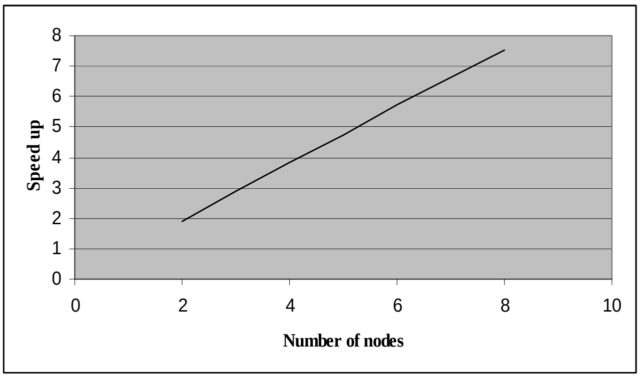


Figure 1. Speed up Ratio

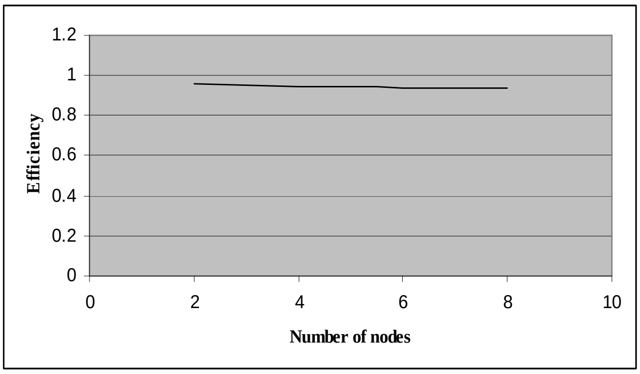


Figure 2 Efficiency

4. GEOMETRIC DECOMPOSTION

The way to address the excessive memory demand is to use geometric decomposition, where specific regions/zones are assigned to nodes, which only treat particles which are in their assigned region. The advantages of this scheme are primary that it saves memory; the node only needs to maintain a data base for its particular region.

They are two problems associated with this approach: laod-balancing and communications. It is difficult to assure load-balancing without some a priori knowledge of the problem because particles may tend to congregate in certain regions of the phase space. Interprocessor communication also becomes an issue because particles will cross boundaries. Crossing a region boundary necessitates that the particle description be communicated from one node to the next. Thus, geometric decomposition may result in a substantial saving in memory due to the need for each node to only keep track of one region, but it has the consequence of increased interprocessor communication and the potential of unequal workloads.

The second case considered in this paper is based on geometric decomposition and is used for material growth Monte Carlo simulation. This is based on a technique called molecular-beam epitaxy (MBE) which aims to create new classes of artificially structure materials involving semiconductors, metals and insulators [8, 14]. The process involves randomly depositing atoms onto a surface after which they acquire energy from the thermal vibrations of the substrate and hop from site to site while undergoing attachment and detachment processes involving other atoms, clusters and step edges before final incorporation. The current inability to probe into the details of events on such scales is one of the primary motivations behind development of computer simulations.

4.1. Parallelisation and Decomposition

The geometric decomposition seems to be the natural way to decompose the lattice. There are various schemes for decomposing a lattice geometrically and mapping it onto a number of nodes. The obvious configurations involve dividing the lattice into an equal number of slices or to tessellate the lattice into a checkerboard pattern. We have adopted the former configuration because of simpler communications required for accessing nearest-neighbors than in the latter configuration.

Again a master/slave approach is adopted. The role of the master process is to control the global status of each slave process. Under this geometric decomposition, each slice or sub-region of the lattice is attached to a unique node (slave). The number of nodes that are employed can be varied; the more nodes that are invoked the smaller is the area of the lattice on each node. This, however, does increase the ratio of the length to the area of the lattice and means that more communications will occur.

Since the rules for the dynamics are approximated by sequential nearest-neighbor hopping processes, each sub-region needs only to communicate with its nearest-neighbors. In such a case the ideal network map or topology for the system would be a pipeline system connected in a modular fashion to form a ring with spokes to the central master node. Periodic boundary conditions are employed in the direction along which there is no decomposition of the lattice. For the other direction, there is an overall toroidal boundary condition amongst the nodes.

The range of the interaction in the barrier to hopping is modeled by nearest-neighbor interactions, i.e., a site requires the height of its nearest-neighbors to determine its environment. This causes difficulties at boundary sites since neighbor sites are not local to the node. In principle this information can be obtained via interprocess communications each time it is needed. This however is an extremely expensive overhead in computational time. To minimize the overhead, each node *i* contains a copy of the boundary site information from nodes *i+1* and i-*1*, i.e., the nearest neighbor nodes. These boundary sites are stored onto ghost sites on each node, so-called because they are not under the control of the local process, but are used only to calculate environments of sites local to the node.

4.2. Parallel Procedure

Now that the geometric scheme for mapping the lattice onto a variable number of nodes and the communications network have been defined, we can run the simulation code on each node which deals with its own geometry. For each generated event, the following is done in addition:

1) If the event is local to the node, update the relevant site and calculate its new environment; otherwise update ghost sites.

2) If the event occurs on a boundary site then send the boundary site information to the nearest-neighbor node, or stack this information in a buffer for transmission at a later time.

3) Initiate a background receive process to check if any stack of boundary site information has arrived. If it has, update the relevant boundary sites.

4) Check if the time for the next input/output (i/o) event has been exceeded. If it has, process the i/o request. Synchronize prior to each i/o event. Each slave process handles its own file i/o.

5) Check if the time for synchronization has been exceeded. If it has, stop all locally generated activity and check for any stack of incoming boundary site information. Generate a new synchronization time and proceed with local activity once all slave nodes have synchronized.

An important part of the simulation is the control of the communications between nodes since this greatly affects the efficiency of the implementation. Due to the stochastic nature of the problem all interprocessor communications are non-blocking with the data transfer occurring in a synchronous mode. Two types of communications are required in the simulation. Local communications which occurs between neighboring nodes at random intervals of time and is required when an event occurs on a boundary of a sub-region, and global communications which occurs at regular fixed intervals of time when the system enters into the synchronization mode or prior to i/o and finally before exiting the program. Here, the master node receives broadcasts from each slave node regarding its current status. Upon receiving the status of all the slave nodes, the master node broadcasts a return status to each slave indicating it can now proceed.

The sending of boundary site information is deterministic on each node since it is known on the local node where the event occurred. However, the contrary applies for the process of receiving. Data can be sent to its neighbor only if the previous transfer was completed successfully. Due to the non-blocking nature of the transmission new boundary events can be generated before the first has been successfully transmitted. Synchronous transmission is necessary to prevent the earlier data being over written. The result of this is that the program could be blocked waiting for the successful transmission of the earlier data. To avoid this, if the boundary site information cannot be sent immediately, then it is stacked in a buffer to await transmission. This leads to the problem that data transfer occurs out of time step. However, these problems can be minimized by correction at synchronization times and it greatly enhances efficiency. Synchronization simply flushes the stacking buffer by forcing the transmission, ensuring the lattice structure throughout the nodes is commensurate (i.e., the nodes are in agreement at the boundaries).

The number of sites that are stacked awaiting transmission is dependent on the geometry of the lattice and the growth conditions employed. If the width of the boundary is large with respect to the other dimension of the lattice, then there will be a significant number of events which occur on the boundary. In addition, conditions of high mobility will also lead to a large number of boundary events. If the rate of successfully completing a receive process is much less than the required transfer rate, then the number of sites stacked increases. As the stack length increases, the error in the configurations at a particular time increases accordingly. These errors are removed at synchronization

4.3. Results and Observations

Figures 3 and 4 show the speed-up and the efficiency obtained from the parallel implementation for variable number of nodes. There is almost a linear monotonic increase in the speed-up as a function of the number of nodes, particularly when the number of nodes is small. However, a change is detected in the speed up as the number of nodes increases.

For greater number of nodes, the communications across the boundaries involve different nodes, which reduce the bottleneck for the stacking of sites for transmission giving a gain in the speed-up. However, the efficiency of the implementation is not very good and decreases as the number of nodes in the system increases. This is because each node performs less computation, as the number of nodes increases, but the parallel house-keeping facilities (communication time, etc.) also increases. However, in order to avoid a saturation in the speed-up, implying that further gain cannot be made by increasing the number of nodes, the amount of computation (geometry) needs to be increased per node.

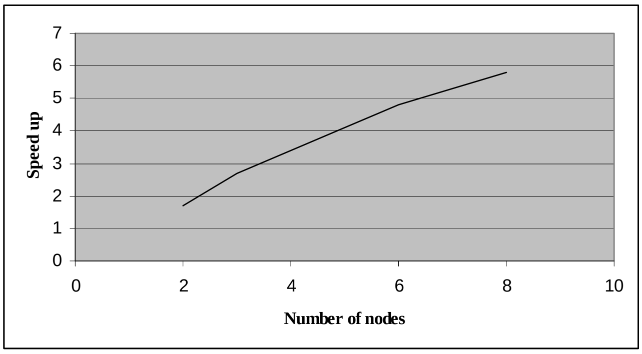


Figure 3 Speed up Ratio

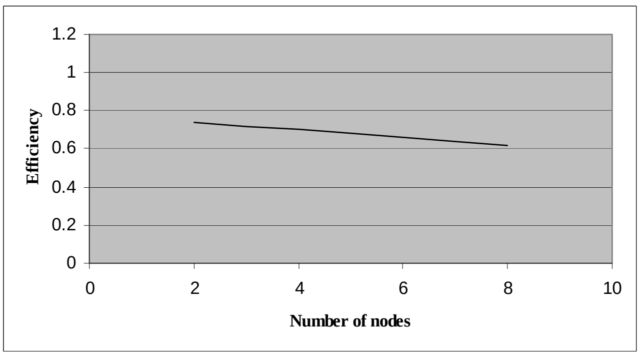


Figure 4 Efficiency

5. CONCLUSIONS

Although the considered simulations have focused on some particular applications they are examples of generalized event driven algorithms and so the used decomposition strategies can be easily applied to other systems and applications. These include simulated annealing, queuing systems, etc. As the applications under consideration indicate, our interest is in general purpose systems with highly computation intensive applications. It is however difficult to generalise the suitability of one particular strategy for all applications. It is clear that certain applications are ideally well suited to parallel environments. While others do seem to require a shared state. In this case it is essential to construct a coherent shared sate of some fashion on top of the message passing model.

The work described above has shown that history-based particle transport Monte Carlo is well suited to parallel computer architecture, and that codes based on the method written for conventional computers can be readily be adapted to run on distributed systems. It has been shown that good speed up factors can be obtained from the parallel implementation of the method, which means that saving in computing time can be achieved when compared with the method in the serial mode. Communications are non-blocking due to the stochastic nature of the problem, with the data transfer occurring in a synchronous mode. In contrast the second case study the impact of communications and sychronisation is significant due to the decomposition nature of the program, although speed up can be achieved the ratio between the benefits and overhead needs to be considered carefully, however, for larger problems it is expected that a large amount of time spent on computation thus minimizing the impact of the communications and synchronisation overheads.

REFERENCES

[1] S. Khaddaj, B Makoond, R Oudrhiri, "Distributed Computing Techniques for Wireless Messaging Systems" in 'Journal of Algorithms and Computational Technology', pp 429-446, Multi-Science Publishing Co Ltd, 2008.

[2] B Makoond, S. Khaddaj, R Oudrhiri, "Modelling and Simulation of Distributed Communication Agents", The Seventh International Conference on Distributed Computing and Applications for Business, Engineering and Science, pp. 52-60, 2008.

[3] S. Khaddaj, "The Impact of Component and Distributed Systems on Software Quality" in 'Semantic Enterprise Computing', pp. 15-21, 2008.

[4] B Makoond, S. Khaddaj, M.J. Tunnicliffe, "A Study on the Development of Petri Net Methodology for the Development of Distributed Applications", PGnet 2004, Liverpool, 2004.

[5] B Makoond, S. Khaddaj, "The Design of a Multi Agent Systems for the Management of Distributed Disaster Logistic", BPML, Cyber city Ebene, Mauritius, 2004.

[6] S. Khaddaj, "Stochastic Parallel Discrete Event Simulation Applied to Materials Design", High Performance Computing Systems, IEEE Canada/Carlton University Press, pp. 391-399, 1996.

[7] N Haider, S. Khaddaj, M Wilby, D D Vvedensky, "Parallel Monte Carlo Simulations of Epitaxial Growth" in 'Computers in Physics', 1 90, pp. 85-96, 1995.

[8] S. Khaddaj, "Parallel Programming and Graphics Techniques Applied to Artificially Structured Materials", 4th Meiko User Conference, Southampton, U.K, 1993.

[9] S. Khaddaj, "Scientific computation and visualisation on a hybrid computer architecture", Parallelism and Distributed Computing in Materials Science: Trends and Challenges, Nancy, France, 1992.
[10] S. Khaddaj, "A parallel computer performance comparison as applied to a Monte Carlo atmospheric dispersion model", Parallel Computing workshop, Imperial College, 1992.
[11] S. Khaddaj, "A comparison of Monte Carlo photon transport on two types of parallel Computer" in 'Annals of nuclear Energy', 18(3) pp. 155-166, 1991.
[12] S. Khaddaj, "The solution of radiation engineering problems on a transputer based system" in 'Concurrency: Practice and Experience', 3(4) pp. 423-434, 1991.
[13] S. Khaddaj, "Monte Carlo photon transport on parallel computers" in 'Progress in Nuclear Energy', 24, pp. 153-164, 1990.
[14] S. Khaddaj, "Parallel processing techniques for solid state physics applications" in 'Parallel Computing and Transputer Applications', Edited by M. Valero et al, IOS Press/CIMNE, pp. 1029-1038, 1992.

FRAMEWORK OF SERVICE DISCOVERY FOR COMPOSITE PROCESS WITH RESPECTING BUSINESS AND TIME CONSTRAINTS *

Yiqiang Li[1], Yu Dai[1], Zhiliang Zhu[1], Bin Zhang[2]
[1]College of Software, Northeastern University
[2]College of Information Science and Technology, Northeastern University
Shenyang, Liaoning, 110004, China
Email:neu_daiyu@126.com

ABSTRACT

This paper proposes an approach for solving the problem of service discovery for composite process. In such approach, in order to respect the business relation between operations of services, business relation between tasks is added to composite process. In order to generate requirement towards service, an algorithm for generating such requirements which constrains the time and business relation on service is proposed. Besides this, a multilevel matching relation model is given which includes matching rules on the level of syntactic, functional semantics and behavioral semantics to test whether service is similar to the requirement. And based on such relation model, an algorithm for discovering services for the composite process is proposed. The experimentation shows that the proposed approach can assure the time and business relation effectively.

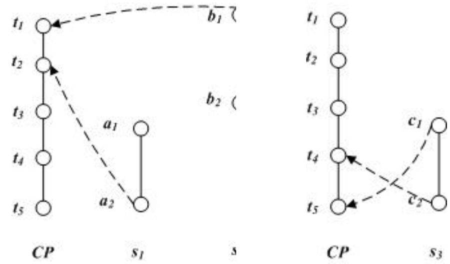
Keywords: Service discovery, Composite process, Service, Functional semantics, Behavioral semantics.

1. INTRODUCTION

Web services are loosely coupled software components, published, located and invoked across the Web. With the prosperity of web services, composing available services to form a new value-added one becomes a new business paradigm. Web service composition has been studied variously and a variety of frameworks are proposed [1]. These frameworks include AI planning based composition [2, 3], workflow based composition [4-6], and interactive composition [7-9].

In this paper, we formulate a service composition problem which starts from a composite process specification and arises naturally from studies on interactive composition. A composite process is a process for a composite service with each task in the composite process as a placeholder for a concrete service. One of the essential technical elements in this framework is service discovery to find appropriate concrete service for each task in order to accomplish the composite process. Approaches in [7, 10] have reported problem of service discovery for composite process while simplified such service discovery problem by assuming that composite process only have stateless services, where there exists no business relation between tasks and each task can be performed by different service. However, this assumption may not always hold. Composite processes with stateful services are often widely used in the domain area such as online shopping.

Let us illustrate the problem of service discovery for composite process with stateful services using the example in Fig 1, where there are three web services s_1, s_2 and s_3 as well as one composite processes CP. Service s_1 and s_2 are provided by two companies for providing online air ticket booking. CP is a composite process for booking air ticket and hotel. In CP, t_1 is a task for querying whether the air ticket is available, t_2 is a task for ordering desired ticket for users, t_5 is a task for receiving payment for the hotel booking and t_4 is a task for giving the booked room number. a_1 and b_1 are the operations which can fulfill the functional requirement of t_1 while a_2 and b_2 are the operations which can fulfill the functional requirement of t_2 as well as c_1 and c_2 are the operations which can fulfill the functional requirement of t_4 and t_5 respectively. Fig 1 (a) shows that task b_1 is assigned to t_1 and a_2 is assigned to t_2. In this situation, if s_1 has no air ticket of s_2, CP cannot run properly as the output of b_1 cannot feed to a_2; if the price of the air ticket booked by s_1 is above the price of the air ticket queried by s_2, the performance of CP may not be accepted by end users. The main reason for such phenomenon is that there exists no business relation requirement between tasks in the composite process CP. If it is required that t_1 and t_2 must be assigned by the same service, such phenomenon will not happen. Fig 1 (b) shows that c_2 is assigned to t_4 and c_1 is assigned to t_5. In this situation, CP cannot run properly also as the time relation of c_1 and c_2 in s_3 violates the time relation of t_4 and t_5 in CP. The main reason for such phenomenon is that current service discovery approach overlooks the time relation between operations in certain service. If it is required that operations which can assign to t_4 and t_5 must have the same time relation as that of t_4 and t_5, such phenomenon will not happen.



(a) Business Disruption (b) Time Disruption

Figure 1. Problems of service discovery for Composite Process

In order to preserve the time and business relation, this paper proposes an approach. The main contributions of this paper are three-fold. First, a novel approach is proposed to discovering services for composite process. Based on the requirement constraining the time and business relation among tasks, services with operations satisfying such relation will be discovered. Second, business relation between tasks is added to composite process and based on the composite process, an algorithm for generating requirements towards service is proposed. Third, a multilevel matching relation model including matching rules on the level of syntactic, functional semantics and business behavioral semantics to test whether service is similar to the requirement is defined. And based on

* Supported by the National Natural Science Foundation of China under Grant No.60773218; the National High-Tech Research and Development Plan of China under Grant No 2009AA01Z122.

such relation model, an algorithm for discovering services for the composite process is proposed. Compared with the works [11] we have done for this problem, this paper presents the framework of the approach and gives the detailed algorithm of generating the local requirements towards service. The former work has a limitation as a result of an assumption that each service can only implement one local composite process. For this problem, this paper extends the former work by a method of verifying whether the service can implement multiple local constraints. The experimentation shows that the proposed algorithm can assure the time and business relation effectively.

2. APPROACH OVERVIEW

Fig.2 gives the framework of our approach:

- When the composite requestor needs a composite service, he will firstly give a composite process.
- According to such composite process, the requirement towards service will be generated.
- Then, according to each requirement towards service, based on the proposed multilevel matching relation model, services for each requirement can be discovered.
- After that, a compatible verification will be done in order to generate the ultimate composite service.

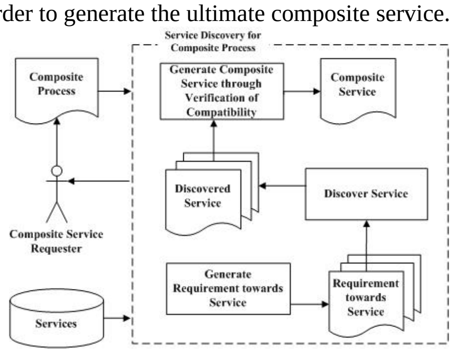


Figure 2. Framework of Approach

3. DETAILS OF THE APPROACH

3.1 Generating Local Requirement towards Services

Before given the description of proposed business relation added composite process, we will firstly give the definition of composite process.

Definition 1. Composite Process. A composite process CP can be defined as a 2-vector: CP=<Taskset, C >, where:

- Taskset={t_1, t_2,…, t_k} is a set of task nodes in the composite process. For each $t \in$ Taskset, t=<ID, Input, Output>, where ID is the identification of t, Input and Output are the input and output parameters and for each parameter p, it can be defined as p=<ID, Meaning, DataType> where Meaning signifies the semantics of p and DataType is the data type of p.
- $C \subseteq$ TaskSet$\times$Taskset$\times$condition is used to signify the time relation. For each e=(t_i, t_j, c_u)$\in$ C, it means that if condition c_u is true, after task t_i is finished, t_j will be invoked. We call t_i and t_j are respectively the source and target task of e which can also be signified as $t_i \prec_{c_u} t_j$.

Currently, most of the process description languages such as BPEL all support the above definition. However the above definition of composite process only identifies the control flow which constrains the time relations between tasks. In order to support service discovery for composite process, it is needed to add business relation between tasks to the composite process. Business relation constrains that tasks having such relation will be provided by the same service.

As current approach lacks of the business relation between tasks in the composite process, it will result in the disruption of business and time relation. Through adding business relation to the composite process, in this paper, the definition of business relation added composite process is given as follows:

Definition 2. Business Relation Added Composite Process. A composite process BCP can be defined as a 4-vector: BCP=<Taskset, C, B, f>, where:

- The meaning of Taskset is as same as Taskset in Defintion 1.
- The meaning of C is as same as C in Defintion 1.
- $B \subseteq$ TaskSet $\times$ Taskset is used to signify the business relation. For each g=(t_i, t_j)$\in$B, it means that t_i and t_j have business relation and they must implemented by a same service which can also be signified as $t_i \Delta t_j$ or $t_j \Delta t_i$. Especially, for each task t_i, (t_i, t_i)$\in$B;
- f:B $\rightarrow$ BRD is a mapping relation from B to BRD. BRD={br_1, br_2,…,br_v}is a set of business descriptions and for each br_i in BRD, it can be defined as br=<ID, Fc> where ID is the identification and Fc signifies the functional category. For each g=(t_i, t_j)$\in$B, f(g)=br means that the business relation between t_i and t_j is for the function br.Fc.

In business relation added composite process BCP, there exists a set of tasks t_1, t_2,…, t_n . If $t_i \prec_{c_1} t_2 \prec_{c_2} \cdots \prec_{c_{n-1}} t_n \prec_{c_n} t_j$, we call that in BCP, t_i can achieve t_j which can be signified as $t_i \rightarrow t_j$. (t_i,t_2,…,t_n,t_j) is called a path from t_i to t_j. In business relation added composite process BCP, there exist a set of sets x_1, x_2,…, x_u. For each x_i, $x_i \subseteq$BCP.Taskset and $x_i \cap x_j = \phi$ ($x_i \neq x_j$) and $x_1 \cup x_2 \ldots \cup x_u$ =BCP.Taskset. For each $t_{ik} \in x_i$ and each $t_l \in$BCP, if $t_{ik} \Delta t_l$ and $t_l \in x_i$, we call that (x_1, x_2,…, x_u) is a business partition of BCP. And for each x_i in the business partition, we call it a gene.

In the rest of this paper, in order to simplify, when we mention composite process, we mean that the composite process is a business relation added one.

In service discovery for composite process, as requirement towards service is implicated in the composite process, it is needed to generate such requirement automatically or manually. In order to facilitate the discovering process, this paper gives the requirement generating algorithm as follows.

One of the important issues in the service discovery for composite process is to make the requirement towards service full of time and business relation in order to solve the problem of disruption of time and business relation illustrated in Fig 1. For each gene in the business partition, it is a set of tasks which have business relations and must be implemented by a same service. Gene constrains the business relation between operations of certain service. If taking the gene as the requirement, it will not disrupt the business relation. However, as the gene has no time relation, it is needed to add the time relation to gene and the time relation between tasks in the gene must be consistent with the time relation between tasks in the composite process. In the following, we will give the definition of local composite process.

Definition 3. Local Composite Process. For business partition (x_1, x_2,…, x_u) on the composite process CP and certain gene x_i in the partition, if the following conditions are true, we call that composite process $LCPx_i$ is the local composite process of x_i under CP.

- $LCPx_i$.Taskset=x_i.Taskset;
- $\forall$ t_i, $t_j \in$ CP.Taskset and $t_i \prec_{c_u} t_j$, if t_i, $t_j \in LCPx_i$.Taskset, then in $LCPx_i$, $\exists$ $t_i \prec_{c_u} t_j$; if $t_i \in LCPx_i$.Taskset and $t_j \notin LCPx_i$.Taskset, then if $\exists t_k \in LCPx_i$.Taskset and in CP, $t_j \prec_{c_u} t_k$, then in $LCPx_i$, $\exists t_i \prec_{c_u} t_k$; if $t_i \notin LCPx_i$.Taskset and $t_j \in LCPx_i$.Taskset, then if $\exists t_k \in LCPx_i$.Taskset and in CP, $t_k \prec_{c_u} t_i$, then in $LCPx_i$, $\exists t_k \prec_{c_u} t_j$.

The first condition in the above constrains that the local composite process have the same tasks as the gene. The second condition constrains that if there exists a path between two tasks in the local composite process, there also exists a path between these two tasks in the original composite process which means that the time relation between tasks in the local composite process must be consistent with the time relation between tasks in the composite process.

For that in the local composite process, all the tasks must be implemented by a same service and the operations of the service must have the consistent time relation as the local composite process, the local composite process can be served as the requirement towards service.

Definition 4. Local Requirement towards Service. For business partition (x_1, x_2,…, x_u) on the composite process CP, for each gene x_i in the partition, under x_i the requirement towards service can be defined as R=<Fc, LCP>, where Fc is the function description Fc=CP.f(x_i.t, x_i.t) and LCP is the local composite process of x_i under CP.

The basic idea of generating such requirement towards service is: firstly, based on the business relation between tasks, get the business partition of the composite process; finally, for each gene in the business partition, generate the local composite process whose time relation is consistent with the composite process. Algorithm 1 gives the algorithm of generating the requirement towards service.

Algorithm 1. Generating Local Requirement towards Service.

```
RS GR(CP) //CP is the composite process, RS is a set of requirements
begin
    X=GetBusinessPartition(CP);//Get business partition of CP
    for i=1 to X.Getlength do
        RS[i]=GenerateR(X[i], CP);
        //for each gene, get the local composite process
    return RS;
end
X GenerateBusinessPartition(CP)
begin
    for i=1 to CP.Taskset.GetLength do
    begin
        if visited[i]=false then
        //task_i has not been visited
        begin
            for j=i to CP.Taskset.GetLength do
            begin
                visited[j]=true;
                if (t_i, t_j) ∈ CP.B then
                X[u][++u_k]=t_j;//put t_j into gene u
            end
        end
    u++;
    end
    return X;
end
R GenerateR(gene, CP)
begin
    put each task in gene into R.LCP.Taskset;
    R.Fc=CP.f(gene.t_1, gene.t_1);
    for each e in CP.C do
    begin
        if e.source in R.LCP.Taskset and e.target in R.LCP.Taskset then
            put e into R.LCP.C;
        else if e.source in R.LCP.Taskset then
            begin
            if ∃e' in CP.C and e'.source=e.target and e'.target in R.LCP.Taskset then
            put (e.source, e'.target) in R.LCP.C;5
            end
        else if e.target in R.LCP.Taskset then
            begin
            if ∃e' in CP.C and e'.target=e.source and e'.source in R.LCP.Taskset then
            put (e'.source, e.target) in R.LCP.C;
            end
    end
end
```

3.2 Discovering Services for Requirement

We define a service as a set of operations and these operations have time relation with each other. In this paper, we assume that for each operation in certain service, it has business relation with all the other ones. Finding services for the composite process is to find a service for each requirement towards service. The requirement towards service constrains the functional description, the operations in the service and the time relation between operations in the service. In the following, we will give the definitions of service, ontology for facilitating the functional semantic matchmaking and the multilevel matching relation model.

Definition 5. Service. A service can be defined as s=<ID, Fc, ServiceModel>, where ID is the identification of s; Fc is the functional category and ServiceModel=<OPset, C> is used to identify the operations and the time relation between operations in the service (for each op in OPset, it has the same definition as task in Definition 1 and C has the same meaning as C in Definition 1).

Definition 6. Ontology. An ontology Ω = {$concept_1$,…, $concept_n$} contains a set of classes. Each class $concept_j$ has an associated set of properties P_k={p_1,…, p_m}. An ontology relates more specific concepts to more general ones from which generic information can be inherited. Such links have been variously name “is a”, “subset of”, “member of”, “sub concept of”, “supper concept”, etc.

The proposed multilevel matching relation model for matchmaking between service and the requirement contains rules organized into 3 levels. Each rule MR_{pq} at a level ML_p (p=1…3) compares a specific feature of requirements within ML_p. The first level ML_1 compares syntactic attributes such as the number of tasks (MR_{11}). The second level ML_2 compares functional semantic attributes. We define 3 groups of rules at this level. The first group compares the functional semantics of functional category of the requirement and that of the service. The second group compares the functional semantics of parameters of each task of the requirement and that of parameters of each operation of the service. The third level

ML_3 compares the behavioral semantics of the requirement and service. We define one group of rules at this level which compares the time relation among tasks of the requirement and service.

For service s and requirement r, the syntax, functional semantics and behavioral semantics rules is given in Table 1.

Table1. Rules in Multilevel Matching Relation

MR	Rule Description
MR_{11}	$\|s.OPset\| \geq \|r.Taskset\|$
MR_{21}	$\forall$ p$\in$task.Output, $\exists$p'$\in$op.Output, p.Datatype is the sub-type of p'.Datatype and Ω(p.Meaning)=Ω(p'.Meaning); $\forall$ p' $\in$op.Input, $\exists$p$\in$task.Input, p'.Datatype is the sub-type of p.Datatype and Ω (p'.Meaning)= Ω (p.Meaning). (task$\in$r.Taskset, op$\in$r.OP)
MR_{22}	Ω (s.Fc)=Ω(r.Fc)
MR_{31}	$\forall$e$\in$r.C, in s, there exists a path which makes e.target$\rightarrow$e.source.

Through testing if the service has matching relation with the requirement based on matching rules in Table 1, it can discover service for each requirement towards service of the composite process. Thus, for the problem illustrated in Fig 1, as t_1 and t_2 have business relation, they must be implemented by a same service. As in service s_1, the time relation between b_1 and b_2 is consistent with that of t_1 and t_2 in the composite process. s_1 is a service that can implement t_1 and t_2.

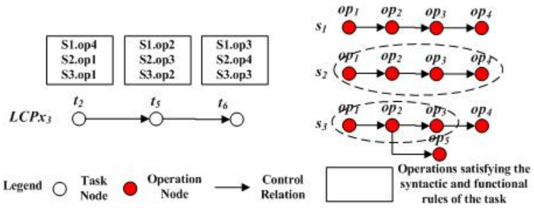


Figure 3. Example for Service Discovery

Let us illustrate a more sophisticated example. In Fig 3, as the time relation between $s_1.op_4$ and $s_1.op_2$ is not consistent with the time relation between t_2 and t_5, s_1 is not an appropriate service. In Fig 3, as there is a path from $s_2.op_1$ to $s_2.op_3$, the time relation between them is consistent with the time relation between t_2 and t_5, s_2 will be an appropriate service. Also, service s_3 will be the appropriate service. However, service s_2 and s_3 are different from $LCPx_3$ as that they have more operations than $LCPx_3$. This is because that the service model of each service is heterogeneous.

Algorithm 2. Discovering Service for Requirement towards Service

```
DS Discovery(RS, Serviceset)// RS is the set of requirements generated based on Algorithm 1
begin
    for each r in RS do
        begin
    DS[i].r=r;
        for each s in Serviceset do
            begin
                if MR (s, r) then
                    put s into DS[i].S;
            end
        i++;
            end
end
```

The above algorithm has a limitation as a result of an assumption that each service can only implement one local composite process. However, in the reality, each service can implement multiple local composite processes. In this situation, multiple local composite processes that can be implemented by one service should be merged together in order to test whether the time relation between the operations in the service is consistent with that of the merged local composite processes. Thus, after the discovery, a work of testing whether the service which can implement more than one local composite process is able to implement these processes.

Given that there exist requirements towards service $r_1, r_2, \ldots, r_m$, service s is the candidate service for them and CP is the composite process from which these requirements are generated. From the definition of local requirement towards service, the task set of local composite process of r_i is r_i.LCP. Taskset. We use T_Merge to signify r_1.LCP.Taskset $\cup r_2$.LCP.Taskset…. $\cup r_m$.LCP.Taskset. Then, T_Merge is a new business partition and according to Definition 5, the local composite process for T_Merge can be generated. If service s is really can implement r_1 to r_m, the time relation between operations of s must be consistent with the time relation between tasks of the local composite process of T_Merge. The operations of s must comply with rule of MR_{31}. Based on such basic idea, in the following, we will show the algorithm for testing whether the service which is a candidate service for more than one gene is really can implement these genes.

Algorithm 3. Testing Whether Service Can Implement Multiple Local Requirements.

```
DS Test(DS, CP)// DS is the set of candidate services discovered by Algorithm 2; CP is the composite process.
begin
    for each ds_i in DS do
    begin
        for each s in ds_i.S do
            if ∃DSM={ds_j|s in ds_j.S and j<>i} then
            begin
                T_Merge= ∪_{ds_j∈DSM} ds_j.r.LCP.Taskset;
                r=GenerateR(T_Merge, CP);//GenerateR in Algorithm 1;
                if MR_31(s, r)=false then
            remove s from DS;
    end
      end
end
```

For composite process CP, the requirements generating from CP forms the set R={r_1, r_2,…, r_n}. There exists a set S of services, for each service s_i in S, there exists one and only one r_j in R, which can be implemented by s_i. Since that the service model of services may be heterogeneous from the local composite process of r, there may be additional interactions between services in S which may cause deadlock among services. Thus, for services in S, a compatible detection of these services will be done in order to determine whether the service set is a candidate composite service. As lots of works have been done in this field [12], in this paper, for the limitation of this paper, we will not give algorithm for detecting the compatibility among discovered services.

4. EXPERIMENTS

As there have no standard testing data, this paper simulates sets of service behaviors as the testing case. And 4 testing sets are generated where the numbers of services are 200, 400, 800 and 1600 respectively. We will compare the performance of current service discovering algorithm for composite process [10] and the proposed algorithm.

On the testing set where the number of service is 1600, randomly generating 10 requirements towards service, compare the performance of current service discovering algorithm for composite process and proposed service discovering algorithm in preserving the business relation. The experimentation is shown in Table 2 where if the discovered service satisfies functional semantic rules of MMR, it will not disrupt the business relation signifying as 1 and if such service does not satisfy the rule, it will disrupt the business relation signifying as 0.

Table 2. Comparison of Disruption of Business Relation

Service's Requirement	1	2	3	4	5	6	7	8	9	10
Proposed Algorithm	1	1	1	1	1	1	1	1	1	1
Current Algorithm	0	0	1	0	1	0	1	1	0	0

Table 2 shows that current service discovering algorithm for composite process will be likely to disrupt the business relation while the proposed algorithm can assure the business relation. This is because that in the proposed algorithm, the requirement towards service constrains that service must provide operation for each task in the requirement while in the current algorithm, the requirement only constrains that for each task in the requirement, the task can be implemented by operation from different services.

On the testing set where the number of service is 1600, randomly generating 10 requirements towards service, compare the performance of current service discovering algorithm for composite process and proposed service discovering algorithm in preserving the time relation. The experimentation is shown in Table 3 where if the discovered service satisfies behavioral semantic rules of MMR, it will not disrupt the time relation signifying as 1 and if such service does not satisfy the rule, it will disrupt the business relation signifying as 0.

Table 3. TComparison of Disruption of Time Relation

Service's Requirement	1	2	3	4	5	6	7	8	9	10
Proposed Algorithm	1	1	1	1	1	1	1	1	1	1
Current Algorithm	0	0	0	0	1	0	1	0	0	1

Table 3 shows that current service discovering algorithm for composite process will be likely to disrupt the time relation while the proposed algorithm can assure such relation. This is because that in the proposed algorithm, the requirement towards service constrains the time relation between operations in the service while in the current algorithm, the requirement lacks of such constraint on relation.

5. CONCLUSIONS

This paper proposes an approach for solving the problem of service discovery for composite process. Compared with the works we have done for this problem, this paper gives the detailed algorithm of generating the requirements towards service. The former work has a limitation as a result of an assumption that each service can only implement one local composite process. For this problem, this paper extends the former work by a method of verifying whether the service can implement multiple local constraints. The experimentation shows that the proposed algorithm can assure the time and business relation effectively.

REFERENCES

[1] R. Hull and J. Su. Tools for composite web services: A short overview. SIGMOD Record, 34(2), 2005: 86–95.

[2] S. Narayanan and S. McIlraith. Analysis and simulation of web services. Computer Networkds, 42(5), 2003:675–693.

[3] S. Narayanan and S. McIlraith. Simulation, verification and automated composition of web services. Int. World Wide Web Conf., ACM Press, 2002:77-88.

[4] W. M. P. van der Aalst. On the automatic generation of work-flow processes based on product structures. Computer in Industry, 39(2), 1999: 97–111.

[5] S. Lu. Semantic Correctness of Transactions and Workflows. PhD thesis, SUNY at Stony Brook, 2002

[6] A. Tsalgatidou, G. Athanasopoulos, and et al. Developing scientific workflows from heterogeneous services. ACM Sigmod Record, 35(2), 2006: :22–28.

[7] D. Berardi, D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Mecella. Automatic composition of e-services that export their behavior. Int. Conf. on Service Oriented Computing, 2003: 43–58.

[8] C. Gerede, R. Hull, and et al. Automated composition of e-services: Lookaheads. Int. Conf. on Service-Oriented Computing, 2004:252-262.

[9] D. Berardi, D. Calvanese, and et al. Automatic composition of transition-based semantic web services with messaging. Int. Conf. on Very Large Data Bases, 2005: 613–624.

[10] J. Cardoso, A. Sheth. Semantic e-workflow composition. Journal of International Information System, 21(3), 2003:191-225.

[11] Y. Dai, L. Yang, B. Zhang. Composite Process Oriented Service Discovery in Preserving Business and Timed Relation. International Workshop on CESSIA in conjunction with WWW2008, Beijing, China, Apr 2008, Article 7.

[12] S. G. Deng, Z. H. and et al. Modeling Service Compatibility with Pi-calculus for Choreography. Int. Conf. on Conceptual Modeling. Arizona, USA, 2006: 26-39

THE APPLICATION OF STRIP-SHAPED STORAGE MECHANISM IN VIDEO GRID *

Zheng-liang Yao,Hong Zhao, Jie Cao, Lei Yin
School of Computer and Communication, Lanzhou University of Technology
Lanzhou, Gansu, 730050, P.R.China
Email: zhaoh@lut.cn, yaozhengliang@sina.com

ABSTRACT

There are lots of service requests to apply for videos with secondary hot degree, whose video files are only located in the central server because of the limited disk capacity of the local servers, it will cause remote service and long transmission delay. Aiming at the problem, the strip-shaped storage mechanism is produced to save more video files in the local servers with the same disk capacity, so the service rate provided by the local servers is raised and the transmission delay is shortened. In the mechanism, a video file is divided into two segments which have same volume, named the front segment and the end segment respectively, the two threshold values, α and β(α<β) are produced by experimental method. None of segment of a video file will be saved in the local servers while the hot degree value of the video is less than the value α, only the front segment of the video file will be saved in the local servers while its hot degree value is between α and β, or both segments of the video file will be saved in the local servers while hot degree value is greater than β. The result of the simulation in VGSim which is a video grid simulator shows that the mechanism can improve the QoS and service capacity of the video grid obviously.

Keywords: Video grid; CDN; Hot degree; Zipf distribution; Strip-shaped storage mechanism

1. INTRODUCTION

More and more attention has been focused on the video grid that can combine the varied technologies of grid[1, 2] and multimedia organically following the development of Internet. It adopts the CDN(Content Distribution Network)[3, 4] storage mechanism. The video files with high hot degree would be pushed to the local servers distributed on different sites which lots of users reside in. Therefore it covers with large-scale area and has tremendous number of users.

A service request that a user apply for will be satisfied preferentially by the local server located in the site that the user resides in while the local server has the video files needed by the service request and available capability to take on the service, otherwise, the service will be provided by another local server or the central server if the server has the video file and available capability, or the service request will be refused because none of server can afford the service.

A video may form lots of video data streams while different users enjoy it simultaneously. Because video data stream are in real time and time sequence, the video QoS(Quality of Service) lies on the route path which the video data may travel. The *CDN* storage mechanism can improve the QoS through shortening the route path of video data stream by storing the video file the user usually needs in the user's local server.

The effectiveness of the *CDN* storage mechanism in video grid has been proved by many deployed systems [5, 6, 7]. There is a comprehensive description about video grid by Liu[5]. Zhao et al[6] produced the mechanism to improve the service capacity and expansibility of video grid by making some client computers as temporary servers to provide video services to other users. In Wang's report[7], a *VOD* system is developed by using *OGSA-GT3*, the hot degree of video files are discussed for loading balance and schedule schemes. But these papers neglect the QoS effect caused by the video file storage mechanism in the video grid system base on the *CDN* .

This paper which introduces the strip-shaped storage mechanism improve the QoS of video grid, its architecture is organized as follows: firstly, the video grid architecture and the *CDN* storage mechanism is described briefly; secondly, the strip-shaped storage mechanism is discussed detail; at last, the simulation experiment and the future work is presented.

2. THE VIDEO GRID ARCHITECTURE

The backbone of video grid is composed by a central server and some local servers, shown as Figure 1. The central or local server may be a single server for special use with high capacity of processing and storage, or a cluster composed by lots of computers. In general, we assume that the central or local server is composed with multi-CEs (Computing Elements) and multi-SEs (Storage Elements), the CE is responsible to process the service requests coming from the users and the SE is responsible to save the video files. Scheduling of tasks within multi-CEs will be processed by the local scheduler and SE which will be used to support the saving service is also determined by the local scheduler. The chief scheduler is only responsible to assign a task to a server.

Every local server and the central server contains *MFBS*(Module For Business Statistic), *MFTC*(Module For Transmitting Context), *MFMS(*Module For Managing System) and *MFBL*(Module For Balancing Load) and so on[8]. The central server with powerful processing ability and huge storage capacity provides the uniform grid portal and saves all video files. The local servers have lower process ability and smaller storage capacity than the central server generally. A part of video file duplicates with high hot degree are saved on these local servers.

The hot degree value of the videos follows Zipf distribution, the formula of the Zipf distribution can be shown as following:

$$P_k = \frac{k^{-\theta}}{c} \qquad c = \sum_{k=1}^{n} k^{-\theta} \tag{1}$$

Here P_k is hot degree value of the kth video in n videos and θ is a constant, which is called depth factor. In this article, the value

* This research is supported by the Science Foundation, Gansu Province, P.R China project No.2007GS04826

of θ is chosen as 0.8 according to previous report [5].

There is a little of videos with high hot degree according Zipf distribution, so the characteristic of *CDN* storage mechanism is to store the duplicates of video files with high hot degree to the local servers distributed on different sites which lots of users reside in. The central server stores all the video files in video grid, and provides video services to the users who can not receive the video service from local servers. Usually 80% of the users visit 20% of video[5, 6], which means the video amount that users usually apply for is small. In general, there are small duplicates of videos with high hot degree in local server, but it will cause transmission delay when users apply for the large number of videos with low hot degree in non-local servers.

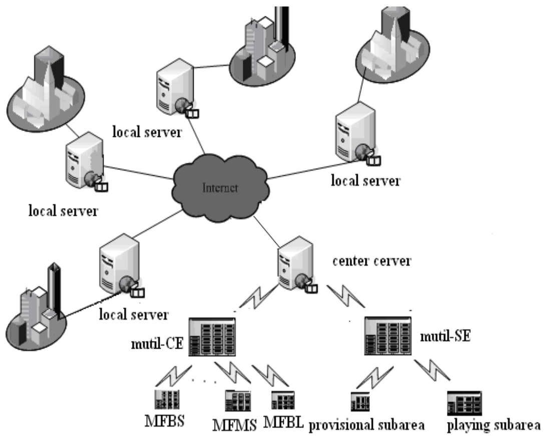


Figure 1. The structure of video grid

3. STRIP-SHAPED STORAGE MECHANISM

3.1 The Design for Polymorphic Structure

Generally, the evolution of the hot degree of a video is shown as Figure 2, at beginning when a new video file is stored in center server, the hot degree value of the video will show increasing tendency; when it has increased up to the peak, it will show declining tendency; after some time more, it becomes flat.

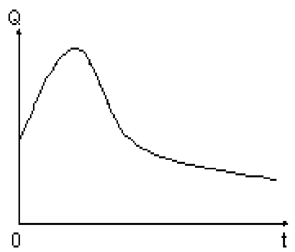


Figure 2. The evolution of the hot degree of a video

According Figure 2, suppose that *V* is a video, the video file of *V* is *F* which is divided into two segments with the same volume, named the front segment and the end segment respectively. The value of hot degree of *V* is P_v, *α, β* are thresholds, the relationship among the Mark Of Hot Degree(*MOHD*), the Range Of P_v(*ROP*), the Change Of P_v(*COP*) and the Volume Of *F* Stored In local Server (*VOFSIS*) is shown in Table 1.

3.2 The Classification of Storage Room in Local Server

According to Table 1, the storage room of local server will be divided into two subareas because of different functions in order to store more video files and facilitate to schedule and update video files.

(1)Playing subarea. It takes most of the storage capacity of local server, the usage of which is to store video files whose video with a certain value of hot degree.

Table 1. The corresponding relationship among *MOHD*, *ROP*, *COP* and *VOFSIS*

MOHD	*ROP*	*COP*	*VOFSIS*
Initial state	$P_v<\alpha$	$\frac{dP_v}{dt}>0$	None of *F*
growth state	$\alpha<P_v<\beta$	$\frac{dP_v}{dt}>0$	The front segment of *F*
Peak state	$P_v>\beta$		The whole *F*
decline state	$\alpha<P_v<\beta$	$\frac{dP_v}{dt}<0$	The front segment of *F*
Death state	$P_v<\alpha$	$\frac{dP_v}{dt}<0$	None of *F*

(2)Provisional Subarea. It takes about a little of the storage capacity of local server and has two roles: firstly when users apply for videos with non-peak state in local server, the complete video files which of video with initial state or death state and the end segment of video files which of the video with growth state or decline state will be downloaded from other servers and provided to the user, at the same time it will be stored in provisional subarea temporarily, which would be provided to other users in local server when these users apply for the videos at the same time period; secondly, it would be used to store these video files which have been washed out from playing subarea and combined with playing subarea as a source of updating to provide to other local server when the video file in video grid are been updating.

3.3 The Scheduling Strategy about the Strip-shaped Storage Mechanism

If the local server has received request from the user which locates in the local server, *MFBS* will calculate the number which the users apply for *F* with automatically, then mutil-CEs would take different scheduling strategy of video files according to the different states of *F*.

(1)Initial state or death state ($P_v<\alpha$). If there is complete *F* which is been downloading or has been downloaded, it would be provided to the user; if not, *F* would be searched from other local servers or central server by *MFBL* and be downloaded to provide to the user by *MFTC*, at the same time, the complete *F* will be stored in provisional subarea in the local server and provided to other users by *MFMS* when they need the whole *F* at the same time period.

(2)Growth state or decline state ($\alpha<Pv<\beta$). *MFMS* provides the front segment of *F* stored in playing subarea to the user. when the video service has been received or the playing point is advanced over 1/4 of *F* by the user, that means the end segment of *F* is confirmed to be needed by the user at this time, if there is the end segment of *F* which is been downloading or has been downloaded to provisional subarea in local server, *MFMS* will use the logic point to link the end segment of *F* to the end of the front segment of *F* to provide to the user in order to ensure the continuity of playing the video; if not, the end segment of *F* will be searched by *MFBL* from other local servers or center server, and will be downloaded to provisional subarea in local servers by *MFTC*, at the same time *MFMS* do the same operation as that when there is the end segment of *F* which is been downloading or has been downloaded to provisional subarea in local server. The front segment of *F* stored in provisional subarea will be provided to other users by *MFMS*, when they need the whole *F* at the same time period.

(3)Peak state ($P_v>\beta$). The complete *F* stored in local server will be provided to the user, the copy of *F* will be provided to other users located in non-local servers need when they need *F*.

3.4 The Updating Strategy about the Strip-shaped Storage Mechanism

The provisional subarea is cleaned up for prepare to store the video files which is washed out from playing subarea. After *MFBS* has calculated the hot degrees of the videos in local server at a certain period of time. Mutil-CEs will update the video files one by one in local server because of different video files with different hot degree in different local servers. Suppose that *Si* is the *i*th of local servers; *Ci* is the *i*th of the video files in *Si*. *P* is the hot degree of the video. The detail is shown as follows:

(1)The hot degrees of the videos in playing subarea of S_1 will be sorted in descending order, the sequence is P_Q: P_{C1}, P_{C2},..., P_{Cn}, ..., P_{Cm}, and the relevant sequence of video files is Q:C_1, C_2,... C_n,..., C_m;

(2)The subsequence G: C_1, C_2, ..., C_n is chosen from Q when $P_{Ci}>\alpha$. *MOHD* of C_i is set to peak state. The video files in $Q\cap\overline{G}$ are sent to provisional subarea;

(3)If $(\alpha<P_{ci}<\beta)\cap(Ci\in G)$ and the complete *Ci* is stored in playing subarea of S_1,the end segment of C_i will be deleted and the *MOHD* of C_i will be set to growth state or decline state according to Table 1 after calculating *COP* of C_i;

(4)If $(P_{ci}>\alpha)\cap(Ci\notin G)$, C_i will be searched from playing subareas and provisional subareas in other local servers by *MFBL*, then it will be sent to playing subarea of S_1 by *MFTC* in other local servers, if $P_{ci}<\beta$, turn to (3);

(5)If ($P_{ci}>\beta$) and there is only the front segment of C_i in S_1, C_i will be searched from playing subarea and provisional subarea in other local servers or center server by *MFBL*, the end segment of C_i is sent to S_1 by *MFTC* in other servers;

(6)The video files in playing subarea in other local servers are dealt with by the same way above;

(7)The provisional subareas of all local servers are cleaned up.

3.5. The Updating Strategy of Provisional Subarea

The update of provisional subarea is in real time. As lots of users' service requests apply to the videos with non-peak state, there will be lots of the complete video files or the end segment of files stored in provisional subareas. If the whole volume of these files has exceeded the rated capacity of provisional subarea, these video files will be updated by using *LFU* algorithm[5].

4. SIMULATION AND RESULTS

An environment with 40 nodes (node 0 is the central server, and the other 39 nodes is local server nodes) is constructed by the *VGSim* [9] after summing up the some video modules[7, 8, 10]. The storage capacity of node 0 is 800 *Gb*, The storage capacity of every local server node is 200 *Gb*, the bandwidth of backbone is 1*Gb/s*. there are 1800 video files in video grid. The volume of every video file is about 375 *Mb*. Generally the playing speed is 512 *kb/s*, the capacity of the playing subarea could contain 100 complete video files. The first experiment simulated the *CDN* storage mechanism, and the video files in local server were updated by using *LFU* algorithm. The second experiment simulated the strip-shaped storage mechanism. In the second experiment, according to Zipf distribution and calculating a lot of experimental data after using Eq(1), if the value of β is set too big, there will be too much the front segment of files in local server, it will consume too much backbone bandwidth when users apply for these video files; if the value of β is set too small, the number of video files in local server will be too little, the probability for users applying for the video services in local server won't increase obviously, so the value is chosen, $\alpha=P_{145}\approx0.001039$, $\beta=P_{50}\approx0.002436$ according to Eq(1). At the beginning, all the video files are stored in node 0. With the advancing of time, these video files are pushed to the nodes of local servers gradually under the strip-shaped storage mechanism. The change of the number of the video files stored in playing subarea in node 1 is shown in Figure 3. The consumption of the backbone bandwidth is shown in Figure 4.

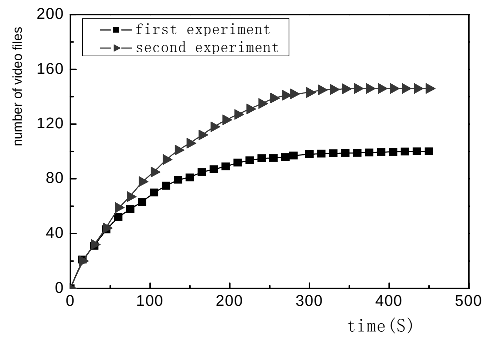


Figure 3. The change of the number the video files stored in playing subarea in node 1

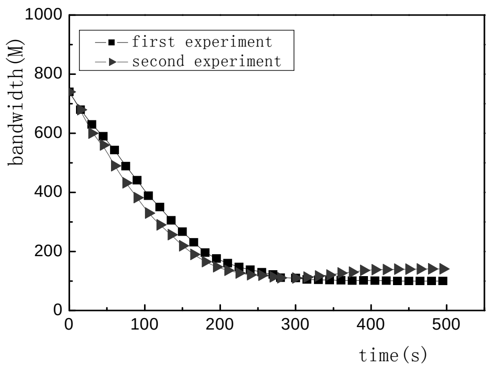


Figure 4. The change of backbone bandwidth

The number of the complete video files and the number of the front segment of video files in local server are fixed by setting threshold α and β. In Figure 3, when the change of the number of the video files in playing subarea of node 1 becomes stable, the number of video files in the strip-shaped storage mechanism is 1.5 times as many as that in the *CDN* storage mechanism. In the second experiment data of Figure 4, at the beginning, all the video files are stored in center server, users request videos from center server, it will consume backbone' bandwidth greatly. With the passage of time, some of the complete video files and the front segment of the files are transmitted to local server, therefore, the users' requests begin to apply for videos from local server, so the consuming of backbone bandwidth is reduced, at the same time, too many users would get the end segment of the files via backbone, but by reusing the video files stored in provisional subarea, the bandwidth consumed by these users will be reduced, so there is no significant difference between the bandwidth consumed in the strip-shaped storage mechanism and the *CDN* storage mechanism at this time. After 300s, the capacity of provisional subarea is used to its rated capacity but the number of the users applying for the same video files is still increasing, so the backbone bandwidth consumed in the strip-shaped storage mechanism is a little more than that in the *CDN* storage mechanism at this time.

5. CONCLUSION AND FUTURE WORK

According to the characteristic of the hot degree of video files following Zipf distribution in normal condition, the strip-shaped storage mechanism has increased the probability for users request video services in local server by increasing the number of video files with secondary hot degree in local server. Therefore, this storage mechanism would decrease the transmission delay when users request the videos with secondary hot degrees.

When users apply for the same videos in non-peak state at a certain period of time, the strategy of reusing video files in provisional subarea will reduce the consumption of backbone bandwidth. If the video file (explosive news) is just put in video grid, it can avoid congestion of backbone network when too many users request the same video service via backbone from center server, so the service quality of video grid is improved.

The strip unit which is the half of the video file is studied in this article, the effect of the strip-shaped storage mechanism by changing the volume of the strip unit in video files will be researched and the best strip-shaped storage mechanism will be constructed in the future work.

REFERENCES

[1] FOSTER I, KESSELMAN C. The Grid: Blueprint for a New Computing Infrastructure [M] 2th edition. Jin hai, Yuan Ping-peng, Shi Ke translation. BeiJing: Publishing House Of Electronics Industry, 2004.

[2] DU Zhi-Hui, CHEN Yu, LIU Peng. Grid computing [M] BeiJing Tsinghua University Press, 2004.

[3] Pallis G,Vakali A. Insight and perspectives for content delivery networks [J].ACM Communication, 2006, 49(1):101–106.

[4] PENG GANG. CDN: Content distribution network [R]. New York: Experimental Computer Systems Lab, Department of Computer Science, State University of New York, 2003.

[5] Liu Peng. Research on High Performance and High Availability Mass Storage Technologies [D], Beijing:Tsinghua University Doctor Degree Dissertation, 2006.

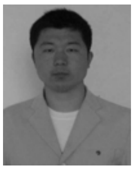
[6] ZHAO Hong, WOXUR Silamu, HOU Yong, LIU Xin-wen, YAO Zheng-liang. A Scalable Architecture of Video Grid Based on CDN [J] MICRO ELECTRONICS & COMPUTER Vol. 25, No. 10, October 2008.

[7] WANG Anfu. Program scheduling and implementing in a Grid-VOD system [J]. Engineering Journal of Wuhan University, Vol. 40 No. 4, Aug 2007.

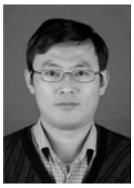
[8] NA Bao-yua, ZHANG Yu-Sena, LIU Peng. Design and Realization of Video Service System Based on Grid [J]. Journal of Jilin University (Information Science Edition) Vol. 26 No. 2. Mar 2008.

[9] Hong ZHAO, WOXUR Silamu, Yong HOU, Xinwen LIU. Zhengliang Yao. Architecture and Realization of Simulator of Video Grid In Proceedings of PACIIA'08, WuHan China. December 2008.

[10] Yan Huang, Tom Z.J.Fu, Dah-Ming Chiu,John C.S.Lui, Cheng Huang. Challenges, Design and Analysis of a Large-scale P2P-VoD System. In Proceedings of ACM SIGCOMM'08, Seattle Washington USA. August 2008.



Zheng-liang Yao is M.E candidate of the School of Computer and Communication, Lanzhou University of Technology China. His research interest is grid computing.



Zhao Hong is an associate professor of Lanzhou University of Technology and a doctor candidate of Xinjiang University. He graduated from Northwest Normal University in 1993, since then he has been working in Lanzhou University of Technology as a faculty. Since 2006, he has studied in the college of information science and engineering of Xinjiang University for doctor degree. He is in charge of nine projects and over 20 journal papers. His interests of research are grid computing, intelligent information processing and distributed parallel processing.

A GRID-BASED PLATFORM FOR DISTRIBUTED MULTI-SOURCE REMOTE SENSING DATA SHARING *

Fan Li [1], Xu Zhang [1], Guang Deng [1] ,Yong Shan [2],Hongrong Wang [3]
[1] Institute of Forest Resource Information Technique, CAF,Beijing, P.R.China
[2] Daniel B. Warnell School of Forestry & Natural Resources,Georgia,GA30602 U.S.A
[3] Information Institute of Beijing Forestry University,Beijing,P.R.China
Email: [1] lifan@caf.ac.cn; [2] shanyongsy@hotmail.com; [3] rongcloud@163.com

ABSTRACT

Grid is emerging as a way to provide users with information integration and application services (computing, storage, accessing, etc.) by the use of the Internet to connect various resources (including computing and storage resources, bandwidth resources, software resources, data resources, information resources and knowledge resources) that are geographically broad as a logical unit. In this work, Grid technology was used to realize data organization management and sharing services on the basis of the existing multi-source remote sensing data under a unified framework. A distributed multi-source remote sensing data platform is designed and implemented with Grid-based portal construction technology. This platform provides visual services of heterogeneous distributed remote sensing data. It also provides flexible and dynamic functions of publishing its services, thus offering a performance guarantee for rapid construction of heterogeneous remote sensing data integration platforms.

Keywords: Grid, Distributed, Remote Sensing Data, Web Services.

1. INTRODUCTION

Remote sensing (RS) is an integrated detection technology, which has been widely used in resources and environment investigation and monitoring, military applications, urban planning and other fields during forty years because of its objectivity, timeliness, macroscopy, integrity and economy. With improvement of resolution and increase of satellite types, now image data has been steadily increasing of a wide range of visible light, infrared, microwave and other electromagnetic waves collected by various remote sensors. These image data in space, time, spectrum, direction, polarization and other aspects form multi-source data in the same area[1][6]. As aerospace technology, sensor technology and database technology are developing rapidly, remote sensing data are growing at an amazing rate. The sharp growth of remote sensing data provides the basic conditions for many areas of application research and new approach of problem analyzing and solving for people. Meanwhile, the characteristics of remote sensing data also bring challenge to real implication. Currently, the explosive increase of remote sensing data has made inconvenience in use due to the great capacity, distributivity, multi-band characteristic, low correlation and storage equipment heterogeneity of the data. Take the China-Europe project Dragon Plan for example, at present, satellite data that have been collected in China consist of Beijing-1, CBERS, HY-1 A/B,FY-3,HJ-1-A,HJ-1-B,HJ-1-C, a total of 15315 sites, and that out of China consist of ENVISAT,ERS,ALOS, a total of 49070 sites. At the same time, we receive MODIS data every day, the current amount of which has reached about 15TB. These remote sensing data are huge and distributed geographically in various heterogeneous storage equipments which have their own different management platform of both hardware and software, management and accessing modes and therefore form “islands of data resources” [2][6]. These features lead to the reality that the existing data management architecture, methods and techniques has no longer been able to meet the need of high-performance and high-capacity ability of distributed data storing and processing. How to store, distribute, organize and manage remote sensing data, as well as finding an effective way of processing, analyzing and mining these huge distributed data, have been considered to be one of the primary issues.

The development of Grid provides an effective means of solving this problem, since the most prominent feature of Grid is the ability to maximize the sharing of resources. By applying Grid technologies to remote sensing data processing, the effective data storage and sharing in distributed heterogeneous environments can be achieved, thus a large number of valuable remote sensing data distributed within different geographical areas can be made use of instead of remain idle, which really eliminates the “islands of data resources”[3][5].

2. THE CONCEPT OF GRID AND OGSA FRAMEWORK

Grid is a brand new computing platform established on the existing base of the Internet[2]. It combines high-speed Internet, high-performance computers, large databases, sensors and remote devices together as an interactive unit that can provide more resources and functions. Grid makes the effective integration of distributed resources available. Users are furnished with all kinds of media of data sharing to improve resource utilization. Grid carries out a reasonable match between requests from a user and the ability of resource supply, and then chooses the appropriate resources and services for the user, realizing resource sharing within wide areas. Resources that are connected by means of communication are integrated seamlessly into an organic whole[3]. These distributed resources are centralized as a huge supercomputer, which offers overall sharing of computing resources, storage resources, data resources, information resources, knowledge resources, expert resources and equipment resources. Resource centralization and sharing are the most important features of Grid. More characteristics of Grid include transparent access to remote resources, support of multi-site management domains and self-government, support of highly active security and fault-tolerant, as well as dynamic state and flexibility.

OGSA（Open Grid Services Architecture）is referred to as the

*This Work was supported by National 863 Project "Research of Distributed Heterogeneous Forest Information Resource Management based on Forest Form Data" (2006AA10Z240), Peking University Wiser Research Fund project "A Study of MODIS data sharing system with Web Services" (W08SD03), China Forestry Research Institute fund project "Research Institute Integrated Service System" (RIFRITZJZ2007009), Chinese Academy of Forestry fund project "Distributed Multi-source Remote Sensing Data Sharing Service System Technology" (IFRIT200801)

next generation of Grid architecture. It was proposed on the basis of the original "Five-Level Sandglass Architecture" and the combination of the latest Web Services technology. OGSA consists of two key technologies—Grid and Web Services. The most prominent idea of OGSA is to take "service" as the center. In the OGSA framework, everything will be abstracted as service, including computers, programs, data, apparatus, equipments, etc. This concept is conducive to manage and use Grid by the adoption of a unified standard interface[4][7][9]. OGSA Architecture is as shown in Figure 1.

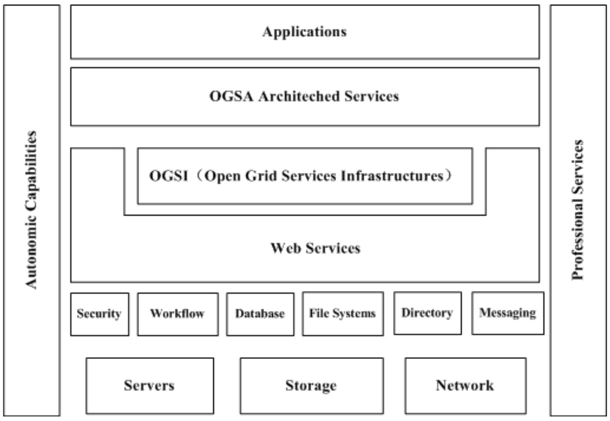


Figure 1. The Architecture of OGSA

3. GRID AND THE PLATFORM FOR DISTRIBUTED MULTI-SOURCE REMOTE SENSING DATA SHARING

The reasons why to use the technology of Grid to build the distributed multi-source remote sensing data sharing platform are as follows: (1) Different from general information, remote sensing data are complex, unstructured, fast changing, wide in range, large in amount, multi-scale and diverse in formats, which brings a lot of difficulties to data users and makes data sharing hard to go on[10]; (2) In application, because remote sensing data required in forestry are often stored in different departments, data production and maintenance are carried out respectively in different units, thus overall integration seems hard, islands of data resources form. How to govern those departments and make them relatively concentrated but actually divided, and offer services for forestry information application through reasonable technological means, is still a problem needs to be solved. (3) In the procedure of analyzing and processing remote sensing data, it is necessary to do some calculation and operation on large amounts of data. However, due to the existing computing environment that requires much time and manpower, it is indispensable to improve the efficiency and ability of computing to meet the increasing computing needs. Hence the technology of Grid is expected to solve series of problems in remote sensing data sharing and service modes.

4. ARCHITECTURE DESIGN AND RESEARCH OF THE DISTRIBUTED REMOTE SENSING DATA SHARING PLATFORM

4.1 The Basic Characteristics of the Grid-Based Distributed Data-sharing Platform

This work is conducted on the basis of Grid to construct a data management and sharing platform with OGSA architecture, which contains the functions of multi-source remote sensing data processing and application. By using technologies that help to solve problems in data auto-discovery, auto-cataloging, quick release, etc., this platform forms a flexible, high performance, open, scalable, collaborative environment for multi-source remote sensing information sharing services, and lays the foundation for conducting future researches within a greater scope of distributed multi-source remote sensing data sharing and services[11].

This platform mainly deals with the storage, backup, pre-processing and other operations of multi-source remote sensing data. It implements the integration of three components: remote sensing data sharing and utility, a series of software and hardware equipments concerning remote sensing data resource management, and relevant information regulation, information technology, information professionals and data standard system. In order to achieve a distributed multi-source remote sensing data service portal, three links of construction should be carried out: database groups building, multi-source heterogeneous data processing and integration, data resources storage construction[12].

(1) Database groups building. Database groups will set up subject databases and meta databases in accordance with present demands, a process of aggregating and managing intensively the existing data.

(2) Multi-source heterogeneous data processing and integration. According to the pre-established standards, norms and corresponding policies, different data are identified, classified and sorted, computed and studied, recorded and indexed, cataloged and organized, and then based on those processes above are combined and integrated to become secondary information.

(3) Data resources storage construction. It is a process in which the data resourced processed are recorded in corresponding media in accordance with certain rules, and then these media are organized as a retrieval system according to certain characteristics, contents and properties.

4.2 Design of the Platform Architecture

In order to achieve multi-source remote sensing data sharing, a data sharing and service system should have functions of data application service that follow the course of data submission—data query—data download, and also functions of data management that follow another course of data storage—data management—data security control[13]. At the mean time, in implementation, for the specific characteristics of remote sensing data, a development strategy "centralized metadata management, distributed data body service" is adopted to replace the traditional data storage and management with metadata storage and management, and realize sharing between remote and heterogeneous dataset entities by transparent metadata accessing.

According to the function model above, the remote sensing data sharing platform on the vertical level can be divided into four layers: the presentation layer, the business logic layer, the resource management layer and the distributed heterogeneous data structure layer. The presentation layer is the medium of interaction between users and the platform. It accepts all kinds of data requests from users, and then submits these requests to the business logic layer to complete specific business processes. The results will be returned to the presentation layer[14]. The resource management layer is for system administrators to check up, release and organize remote sensing data. It also provides functions like data security control and log analysis. The distributed heterogeneous data structure layer provides storage of metadata and data bodies, including the databases and resources of all the nodes.

The basic working process of the platform is as follows: The user poses a data request through a browser to the server. This request then will be accepted by the business logic layer and be submitted in accordance with the type of service to corresponding data processing module in which the business object will invoke certain business logic to analyze and process the response of this data request. Finally, the information of processed data will be displayed through a view in the user's browser. The architecture of the Grid-based distributed remote sensing data sharing platform is shown in Figure 2.

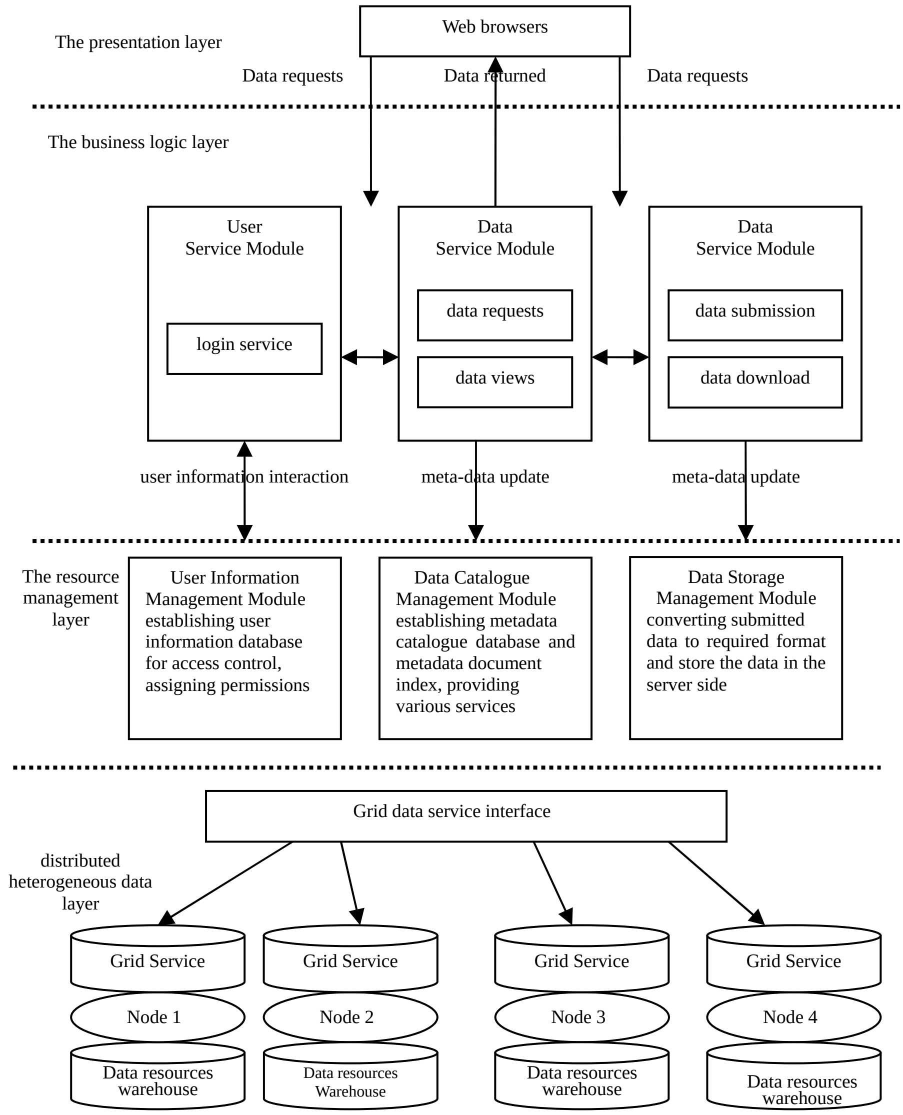


Figure 2. The Architecture of the Grid-based Distributed Remote Sensing Data Sharing Platform

4.3 Layer Functions of the Platform

4.3.1 The Presentation Layer

The presentation layer provides users with a corresponding data reception platform. The data user poses requests through a Web browser to the lower layer. The data service modules in the lower layer then convert the requests received to certain data processing operations, process the required data in business logic layer, and display the final data in way of views to the user.

4.3.2 The Business Logic Layer

This layer can be divided into four modules: user register and login module, data retrieval module, data submission and publication module, online data analysis and download module.

Through data retrieval module, the user can make remote sensing data queries by keyword, subject, time and other forms. Query results are sent back to the user in forms of metadata. With the connection parameters of metadata and data bodies, and the data browser middleware, the user can read the data online. Within the user's data accessing permission and data level, all the datasets can be downloaded.

The function of data submission is conducted by the data submission module. But the data submitted by users must be reviewed and checked up by the administrator before published[15]. Data review consists of metadata review and data bodies review. It is necessary for the administrator to do further processing work on the data bodies according to corresponding information that metadata describe.

The data download module can be able to choose relevant datasets under the data query results, and get the relevant information through http or ftp services. Data subscription service requires users to fill in some information about the data wanted, including the name, scope, format, time and detailed purpose of use of the dataset, and personal contact information as well. After the information above is filled in and submitted to the server, the subscribed dataset will be waiting for the administrator to be reviewed before arrives at the user in due time.

4.3.3 The Resource Management Layer

The purpose of the resource management layer is to protect the security of both data and the system. This layer mainly contains four modules: dataset retrieval and storage management, data catalogues management, data log management and user information management. The Grid-based distributed heterogeneous resource management layer model is as shown in Figure 3.

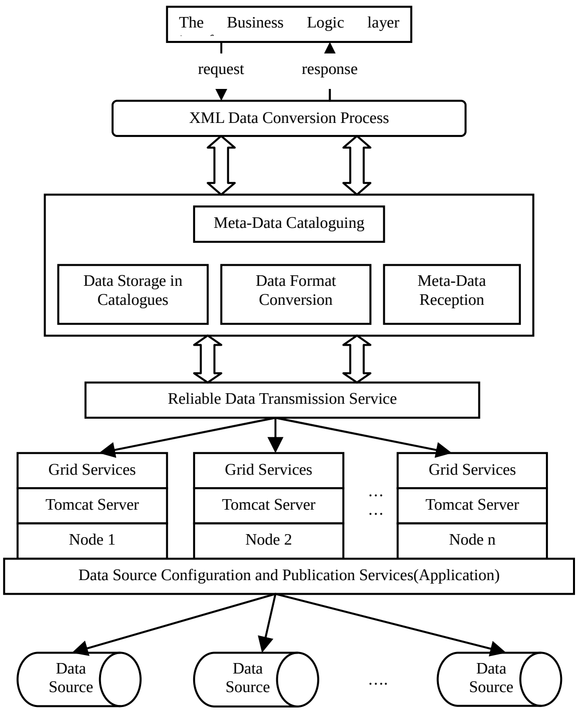


Figure 3. The Model of Grid-based Distributed Heterogeneous Resource Management Layer

Under the heterogeneous storage modes of remote sensing metadata and data bodies, the data storage management module makes a selection among different datasets and rearranges them. When the user submits different types of data to the server through a Web browser, the XML engine inside the server will convert these data into standard XML documents. And then data service components will take over the processing of these XML documents, such as establishing some indexes. Thus, the multi-source heterogeneous data are converted into unified standard XML documents as a fixed-structured middleware, by the integration of which the unified storage of multi-source heterogeneous data can be achieved [7][16].

The data catalogues management module mainly contains basic structures of catalogue generation, data resources classification and metadata generation. When the user publishes a data resource document to data catalogue services, the metadata documents should be stored into the meta-database. Metadata catalogue services can use the database management system to build a meta-database. Indexes of the information that metadata documents contains should be built as a interface for data publication and accessing, from which the user can quickly and easily make a search by keywords.

4.3.4 The Grid-Structured Heterogeneous Data Layer

Combined with Grid technologies such as Grid Services, databases in grid nodes can define database operations themselves that they allow. Data provided by each node are published in forms of Grid Services. Each data node has a Tomcat (Apache) server as the basic operating platform of Grid Service, thus all the nodes form a "physically distributed and logically unified" data resource unit[8].

4.4 Relationships and Organizations of Modules in the Platform

According to the overall framework of the platform architecture, interactions among all the modules are realized through the data requests from users to the platform. The user submit a data request through a Web browser to the platform, and then the platform passes the data request to the lower business logic layer, invoking relevant modules to process the request[17]. After a series of data analysis and rearrangement processes, the required data are presented to the user in certain form of interaction. The collaborative relationships between various modules are as shown in Figure 4.

5. CONCLUSION AND FUTURE WORK

Grid is a new generation of Internet technologies. Through researches and analysis, several advantages are found in the construction of a multi-source remote sensing data platform combining with Grid technology. First, Grid is conducive to the integration of a wide range of different remote sensing data resources and can be used to improve the level of information technology application in forestry; Second, Grid can realize the multi-level distribution model of data at all levels. It also makes the dynamic collaborations between different forestry information systems available in spite of regional and level; Third, OGSA framework as a model is conducive for the development of forestry information systems, as well as the application of remote sensing data.

It is obvious that by constructing a Grid-based multi-source remote sensing data sharing platform the sharing and processing of computing resources, data resources and technology resources can be maximized. This platform enables a large amount of various distributed remote sensing data to be shared, and play their due role.

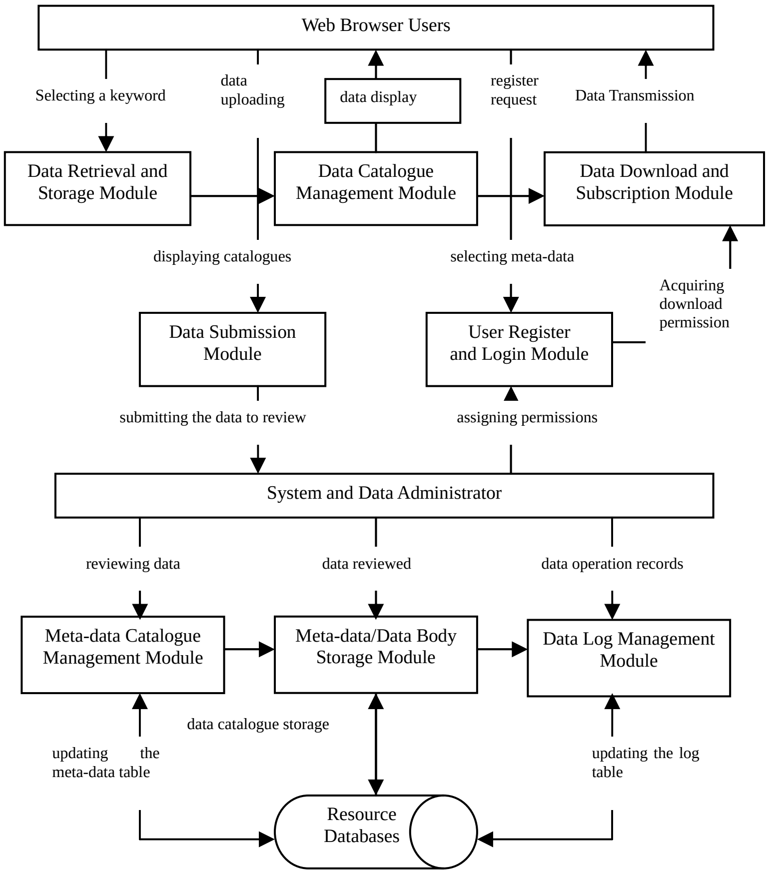


Figure 4. Relationships and Organizations of Modules in the Platform

REFERENCES

[1] Matthew Bloch, Mark Harrower. MapShaper.org: A Map Generalization Web Services, Proceedings of AUTOCARTO 2006.
[2] The Globus Project. http://www.globus.org.
[3] Forster, I. , Kesselman , C.,. The Grid 2[M] Beijing,Electronic Industry Press, 2005.124-220.
[4] UDDI.org. http://www.uddi.org.
[5] WANG Yi-jie, XIAO Nong, REN Hao. "Research on key technology in data grid. journal of computer research and development, 2003 40(12) 1749-1755.
[6] NASA. http: / /www. nas. nasa. gov/ , 2008. 12.
[7] W3C. http://www.w3c.org,2003-06.
[8] ZDNet China．http://www.cn-java.com/target/news.php news_id=1648,2002-05．
[9] WANG Juan-Le, ZHU Yun-qiang, XIE Chuan-jie, "Network platform design and development for Earth system science data sharing", Earth Science Frontiers, 2006,4～59.
[10] HUANG Ding-cheng, The overall framework of science data sharing project. China Basic Science, 2003,63～68.
[11] LIANG Hu-qing, MA Rong-hua, "Research and design of metadata for Geo-spatial data in general provincial situation GIS", Journal of Remote Sensing, 2002,72～278.
[12] WU Jin-hua, The discussion on metadata of geo-spatial data. Journal of Xi'an Engineering University, 2002,9～61.
[13] WEN Yong-ning, ZHANG Hong, "Distributed spatial-data share service model with Web Servicess", Computer Engineering, Vol.1, No.6, 2000, pp.5~26,62.
[14] YANG Lin, DU Xiao-ping. "Data exchange strategy of distributed geological spatial database", Earth Science-Journal of China University of Geosciences, Vol.1, No.5, 2006, pp.59~662.
[15] LI Fan, ZHANG Xu, "Research and development of Web Servicess design standard in forestry science data sharing system", Journal of Northwest Sci-Tech University of Agriculture and Forestry, 2007,35(extra edition)12～16.
[16] LI Fan, ZHANG Xu, "Research and Implementation of Web Servicess management and mechanics in Digital Forestry platform", Forestry Science, 2006,42(extra edition),11～114.
[17] LI Fan, ZHANG Xu, "Research on a distributed MODIS data sharing platform based on Web Servicess", Geography and Geo-Information Science, 25(extra edition),17~18.

Fan Li is an Assistant Researcher at the Forest Resources Information Techniques Institute, The Chinese Academy of Forestry. He graduated from Tsinghua University in 2003 with a software engineering major. He has published 15 Journal papers. His research interests are in distributed parallel processing, grid computing and network security.

EFFECTIVE SELF-ADAPTIVE SERVICE COMPOSITION *

Yu Dai[1], Lei Yang[2], Zhiliang Zhu[1], Bin Zhang[2]
[1]College of Software, Northeastern University
[2]College of Information Science and Technology, Northeastern University
Shenyang, Liaoning, 110004, China
Email:neu_daiyu@126.com

ABSTRACT

Web services run in a highly dynamic environment, as a result of which the QoS will change relatively frequently. In order to make the composite service adapt to such dynamic property of web services, we propose an effective and scalable self-adaptive approach for service composition. In order to make the composite service adjust itself as quickly as possible, a way of performance prediction is proposed in this paper. In order to avoid useless re-selections and improve the efficiency of the re-selections, the method for extracting the local constraint is presented. On the basis of these, the self-adaptive approach is presented including the framework, and the core methods. Experiments show that the proposed solutions are effective in supporting self-adaptive web service composition.

Keywords: Web service, QoS, Service composition, Re-selection, Performance prediction.

1. INTRODUCTION

With the popularity of web services, composing available services to form a new value-added one has become an important way for constructing the application [1]. Web services operate autonomously in a highly dynamic environment (the Web). Accordingly, the QoS of the web services may change with a relatively high frequency [2] and such dynamic nature of the services will influence the runtime performance of the composite service.

Currently, may people focus their study on how to select services according to their QoS [3, 4] to make the composite service meet users' global constraints and exhibit better runtime quality. However, most of current works assumes the accuracy of initial estimated QoS data. Such assumption could not hold true in the real application sometimes. The QoS is changed sometimes due to the dynamics of the environment services operating in. As a result, the QoS may be a deviation from the initial estimated one which will influences the runtime performance of composite service.

The easiest solution to the problem is to re-select the services every time a change occurs. However, it is not feasible due to the high time complexity of the reselection with global optimization [5], which will interrupt the execution of the composite service and influence its performance. To minimize the extra delay caused by this reselection, approaches [6, 7] offering the solution is proposed, whose core is to back up a composite service for each component service. Then, when a component service incurs a failure, the composite service can easily switch to a replacement one and the self-adaptive process will not affect the execution performance of the composite service. Since these approaches find the replacement composite service in the selection, the replacement may not be available as a result of QoS change of component service in the replacement one. This will lead to another reselection which may in turn lead to an extra delay and affect the performance of the composite service.

The goal of this paper is to set the basis to overcome the limits of the previous approaches to adaptive web service composition. The aim is to replace the service whose QoS is a large deviation from the initial estimated one, while respecting the global QoS constraints and preferences. We propose an effective and scalable approach for self-adaptive service composition based on the following main contributions. The core of the approach is to extract the local constraint from the global one. The local constraint reflects the relation between QoS degrading degree of the component service and the fulfillment of the global constraint. Then, if the services can fulfill the local constraint, it can fulfill the global one. In this way, the re-selection with global optimization can be turned to the re-selection with local optimization. The re-selection with local optimization has low time complexity and thus, the extra delay can be minimized, while the re-selection can also insure the global optimization. Besides this, to replace the service as quickly as possible in order to reduce the extra delay, a performance prediction is used in this approach. As will be discussed in the remainder of the paper, our approach is effective in particular for the runtime-aware processes when the QoS is changed dynamically.

2. OVERVIEW OF THE APPROACH

2.1. The Aim of Adaptive Service Composition

In this section, we will introduce the concept of QoS and selection with global optimization to analyze the necessity of self-adaptive service composition.

Definition 1. QoS of Atomic Service. For an atomic service s (which only contains one operation), the QoS of s can be defined as: QoS(s)=<$Q^t(s)$, $Q^p(s)$>, where:

- $Q^t(s)$ is the response time of s which is the interval of time elapsed from the invocation to the completion of service s. $Q^t(s)$ can be defined as: $Q^t(s)$=<$Q^{pt}(s)$, $Q^{dt}(s)$>, where
  - $Q^{pt}(s)$ is the processing time for the request, which is provided by service providers;
  - $Q^{dt}(s)$ is the sum of the transmission time of request and response between the execution engine and the service. The transmission time can be estimated based on past statistics and then it can be either an average of past executions of the service, or the transmission time with a given probability, e.g. 95%;
  - $Q^t(s)=Q^{pt}(s)+Q^{dt}(s)$ means that the response time is a sum of request processing time and data transmission time.
- $Q^p(s)$ is the price of invoking s.

The work by Cardoso [8] proposes a mathematical model for QoS computation of workflow, which involves aggregation functions for sequence, choice, parallel and iteration structure

* This work was supported by National Natural Science Foundation (50874123), and Hi-Tech Research and Development Program of China(2006AA06Z105)

of the workflow. In this paper, we use the aggregation functions as Ref. [8]. In order to provide composite services which satisfies global constraints and preferences defined by users, QoS-driven selection with global optimization is proposed by Ref.[2]. The problem can be formulated as formula (1). The aim of the selection is to maximize the fitness function of the QoS; and meet the constraints defined by the user.

$$\max \sum_{i=1}^{N} \sum_{j \in S_i} F_{ij} x_{ij}$$
$$s.t \quad \sum_{i=1}^{N} \sum_{j \in S_i} Q_{ij}^{t} x_{ij} \le Q_c^{t} \tag{1}$$
$$\sum_{j \in S_i} x_{ij} = 1,\ x_{ij} \in \{0,1\}\ i = 1,\ldots,N, j \in S_i$$

Where x_{ij} is set to 1 if atomic service j is selected for service class S_i in the workflow and 0 otherwise. Q_{ij} is the QoS value of service j in class S_i. F_{ij} is a fitness function which can be computed as (2); Q_c^z is the expected QoS of the composite service.

$$F_{ij} = w_t * \left(\frac{Q_{ij}^{t} - u_t}{\sigma_t} \right) + w_p * \left(\frac{Q_{ij}^{p} - u_p}{\sigma_p} \right) \tag{2}$$

Where w_t and w_p are the weights ($0 \le w_t$, $w_p \le 1$, $w_t + w_p = 1$). σ and μ are respectively the standard deviation and average of the oS values for all candidate services in a service class.

From formula (1), QoS-driven selection with global optimization is a NP hard problem [3]. Normally, it takes a long time to solve this problem. After the selection, the composite service will be executed. Due to the highly dynamic environment of web service, QoS of a component service in the execution may be deviated from the estimated one. In order to satisfy the QoS requirement, the composite service will adjust itself by the re-selection.

2.2. Framework of the Approach

In this paper, we propose an efficient and scalable approach for the self-adaptive service composition to adapt to the dynamics of web services. Fig.1 presents the framework of the approach.

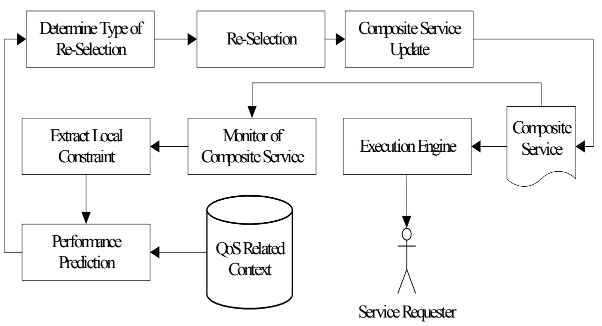


Figure 1. Framework of the Approach

When a composite service is to execute, the monitor will run to monitor the quality of the composite service. According to the monitored quality, the local constraint for each component service will be extracted. When the newly updated local constraint is a large deviation from the older one, or the prediction period is coming, the performance prediction will run to predict whether the QoS of the component service will be a large deviation from the initial estimated one. If so, the type of re-selection will be determined according to the number of services fulfill the local constraint. The re-selection will re-select services either with global optimization, or with local optimization according to the determined type. And then, the composite service will be updated and the execution engine will invoke the new part of it. Finally, the result will be sent to the service requester.

Compared with other adaptive approach, we add the components of performance prediction and extraction of local constraint to the framework. These two components can make the adaptation more efficient. Firstly, by extraction of local constraint, the re-selection with global optimization can be turned to the re-selection with local optimization. This will minimize the extra delay caused by the re-selection. Secondly, by performance prediction, even in the case of the re-selection with global optimization, since the re-selection will be triggered as early as possible, the extra delay can be also minimized. The useless re-selection is a kind of re-selections when the change of quality of re-selection does not affect the fulfillment of the global constraint. Since the re-selection brings about extra delay, the useless re-selection will also bring extra delay. In the real application, these re-selections should be removed out. Thirdly, by extraction of local constraint, the paper considers the relation between the quality of component service and the fulfillment of the global constraint. The useless re-selections can be avoided.

In this paper, we aim at the part of performance prediction and extraction of local constraint.

3. CORE COMPONENTS OF THE APPROACH

3.1 Extracting Local Constraint from Global One

As the time complexity of the re-selection with the global optimization is high, the re-selection will cause extra delay of the composite service. This will affect the runtime performance of the composite service. To cope with this limitation, we divide the re-selection with global optimization into 2 sub-problems that can be solved more efficiently in two subsequent phases. In the first phase, the local constraints on the component services are extracted from the global constraint. The local constraint reflects the relation between the QoS degrading degree of the component service and the fulfillment of the global constraint. Then, if the services can fulfill the local constraint, it can fulfill the global one. In this way, the re-selection with global optimization can be turned to the re-selection with local optimization. In the second phase, the re-selection with local optimization is performed to find the best component service that satisfy the local constraint and has the highest fitness. The re-selection with local optimization has been studied in Ref.[]. Thus, this paper will not discuss this phase in detail. In the real application, the local constraint may not set appropriately. Then, few of services can meet the constraint. In this situation, the re-selection with global optimization will perform. However, if there are services which can meet the constraint, the re-selection with local constraint will have better performance. In the following, we will describe the method of extracting the local constraint from the global one.

A composite process [8] can be divided into sub-processes. These sub-processes have parallel, sequential and conditional relation with each other. Such a composite process can be expressed as a tree. The leaf, middle and root nodes of the tree signify the service, the relation (e.g. sequence, condition, and parallel) and the whole composite service respectively. Then, formally a composite service can be defined as follows: $T=<V, E, f, ep>$, where $V=LV \cup NLV$, LV is the set of leaf nodes signifying services, NLV is the set of middle nodes signifying relations and E is used to signify the connecting relation between nodes, f: NLV→{Seq, Par, Cho} is the mapping function, ep: V→[0, 1] is the weight of the nodes signifying the probability of the branches in the condition. T can also be simplified as T=<V, E>.

Imagine a node v_p in the tree, and the node is a Seq. If the child nodes of v_p from the left to the right are v_{c1}, v_{c2}, ...,v_{cm}, it means that the child nodes will execute according to the sequence as v_{c1}, v_{c2}, ...,v_{cm}. Imagine a node v in the tree, and the node is a Par or Cho. Then, the child node of v signifies its branch. For a node v in the tree, if the depth of the node is d, then the depth of its child node is d+1. The depth of the composite service is the maximum of the depth of the leaf node.

When the composite service is (re)formed, the selected component services can meet the global constraint. Then, for each component service, the initial QoS can be served as a local constraint. Consider a branch in the parallel, the expected execution time of which is not the longest one in the parallel. Then, the execution time quality of the component services in this branch should be relaxed. That is to say, the initial local constraint should be determined with the consideration of the structure of the composite service. When the composite service runs, if the quality of selected component service is degrading, the composite service will be updated to avoid the affect caused by this degradation. If the quality of the selected component service is not degrading or evenly improving, the local constraint of the preceding component services should be relaxed. In the following, we will give the methods to determine the local constraint in the situation of when the composite service is (re)formed and runs respectively.

- Extracting the local constraint when the composite service is (re)formed.

Imagine a composite service <V, E>, and v is a node. The parent node of v is v_p, v_p's and v's expected execution time is t_p and t_v respectively. The local constraint on the execution time of v can be computed as (3). In the following, we will give the computing method of maximum degrading degree (DQ) of the execution time according to the structure of the composite service.

$$\mathrm{LC}^{\mathrm{t}}(v,CS)=t_v+\mathrm{DQ}^{\mathrm{t}}(v,CS) \tag{3}$$

If v_p is a Seq or Cho, then according to the ratio of the execution time of the node v to the one of v_p, the maximum degrading degree of the execution time of v can be computed as the following.

$$\mathrm{DQ}^{\mathrm{t}}(v,CS)=\frac{\mathrm{DQ}^{\mathrm{t}}(v_p,CS)*t_v}{t_p} \tag{4}$$

If v_p is a Par, then according to whether v is the child with the longest execution time of v_p, the maximum degrading degree of the execution time of v can be computed as the following.

$$\mathrm{DQ}^{\mathrm{t}}(v,CS)=\mathrm{DQ}^{\mathrm{t}}(v_p,CS)+(t_p-t_v) \tag{5}$$

If v is the root, then the maximum degrading degree of v is 0.

- Extracting the local constraint when the composite service runs.

Imagine a composite service <V, E>, and v is a node which has just executed. The actual execution time of it is t_v and the initial local constraint on it is t_v'. For the unexecuted nodes, the maximum degrading degree of the execution time can be computed as follows. For the composite service, if the node has executed, it will be removed out from the tree.

If v_s is an ascending node of v, v_s is a Par, v_s is the farthest ascending Par node of v and the ascending node v_n of v is the child node of v_s, then the maximum degrading degree of execution time of v_n can be computed as $\mathrm{DQ}^{\mathrm{t}}(v_n, CS)$=$\mathrm{DQ}^{\mathrm{t}}(v_n, CS)$+($t_v$-$t_v$'). For the preceding node v_d of v_n, if the parent node v_{dp} of v_d is Seq and Cho, the maximum degrading degree of execution time of v_d can be computed as (6). If the parent node v_{dp} of v_d is Par, the maximum degrading degree of the execution time of v_{dp} can be computed as (7).

$$\mathrm{DQ}^{\mathrm{t}}(v_d,CS)=\mathrm{DQ}^{\mathrm{t}}(v_d,CS)+\frac{\mathrm{DQ}^{\mathrm{t}}(v_{dp},CS)*t_d}{t_{dp}} \tag{6}$$

Where, t_{dp} and t_d are the expected execution time of v_{dp} and v_d respectively.

$$\mathrm{DQ}^{\mathrm{t}}(v_d,CS)=\mathrm{DQ}^{\mathrm{t}}(v_d,CS)+\mathrm{DQ}^{\mathrm{t}}(v_{dp},CS)+(t_{dp}-t_d) \tag{7}$$

If for each ascending node v_s of v, v_s is not Par, the maximum degrading degree of the execution time of the root node v_r can be computed as: $\mathrm{DQ}^{\mathrm{t}}(v_r, CS)$=$\mathrm{DQ}^{\mathrm{t}}(v_r, CS)$+($t_v$-$t_v$'). For the preceding node v_d of v_r, if the parent node v_{dp} of v_d is Seq or Cho, the maximum degrading degree of v_d can be computed as (6). If v_{dp} is Par, the maximum degrading degree of v_d can be computed as (7).

Then, services which can fulfill the local constraint and with the highest fitness value (seen in (2)) can be selected. Such re-selection with local optimization can find the feasible solution quickly. It will minimize the extra delay caused by the re-selections. However, as the local constraint may be tight, sometimes it is difficult to find a feasible solution. In this situation, re-selection with the global optimization should be used. In order to minimize the extra delay, this paper uses a performance prediction for finding the QoS degrading as early as possible. In the next section, we will discuss it in detail.

3.2 Performance Prediction

We will uses semi-markov model [9] to predict the performance of component service.

There are 3 states of data transmission speed: if $V(t)$>=th_V_Q, then $ST(t)$ =qualified state; if $0<V(t)<th\_V_Q$, then $ST(t)$=soft damage state; if $V(t)$=0, then $ST(t)$=hard damage state. In this paper, th_V_Q is the data transmission speed with a given probability of 70% of the past executions.

For a component service s in the composite one CS, the expected execution time of s in CS is t_f. As $Q^{dt}(s)$=a/V where a is the amount of data and V is the data transmission speed, the maximum degrading speed of s can be computed as a/($\mathrm{LC}^{\mathrm{t}}(s, CS)$-$Q^{pt}(s)$). The aim of prediction can be described as: if the current state is i, current time is t and the holding time in current state is d, we need to predict the probability of the data transmission speed V_f at future time t_f above the expected speed V_e. Let j be the state V_e belonging to. Here, V_e is the data transmission speed with which the execution time is the local constraint. To solve this problem, we will consider the following two situations:

- State j is same to i

In this situation, the probability can be a sum of the probabilities under the situation with no transition from t to t_f and the situation with at least one transition. Then, the probability can be computed as:

$$P((V>V_e)\wedge(D_i>d))=(1-F_i(V_e))*\frac{1-H_i(t_f-t+d)}{1-H_i(d)} \\ +(1-F_i(V_e))*P_i*\frac{H_i(t_f-t+d)-H_i(d)}{1-H_i(d)} \tag{8}$$

Where, $F_i(v)$ is the distribution of data transmission speed in state i; $H_i(d)$ is the distribution of duration spent in state i.

- State j is different from i.

If state j is different from i, it means that there exist at least one transition during the duration from t to t_f. Then, the probability can be computed as:

$$P\left(\left(V > V_e\right) \wedge \left(Z_{t_f} = j\right) \wedge \left(d < D_i < d + t_f - t \middle| D_i > d\right)\right) = \left(1 - F\left(V_e\right)\right) * P_j * \frac{H_i\left(t_f - t + d\right) - H_i(d)}{1 - H_i(d)} \quad (9)$$

The probability, can be either computed as (8) or (9),according to the current state of service. If the probability is low enough, a re-selection will be triggered. The detailed description of the approach is in [13].

4. EXPERIMENTS

This part will verify the effectiveness of the proposed approach.

Randomly generate 10 composite services, and compute the local constraint of execution time for each component service in each composite service. For each composite service, randomly select an un-executed component service. Change the QoS of this component service and do the re-selection when the QoS cannot fulfill the local constraint. Compare the QoS gains between the maximum degrading degree computing method and the one [10] which uses a general degrading degree (For example, the execution time is longer than the expected one to the extent of 5% and the extent of 5% is a degrading degree.). Table 1 shows the result where – means no re-selection is triggered. QoS gains=(the execution time of the composite after the QoS is changed-the execution time before the QoS is changed)/ the execution time before the QoS is changed.

Table 1. Comparison of QoS Gains

Composite Service	The degrading degree computed in this paper	The degrading degree is an extent of 5%	The degrading degree is an extent of 10%	The degrading degree is an extent of 20%
1	0.2	0.2	0.2	0.2
2	0.15	0.15	0.15	-
3	0.03	-	-	-
4	-	-0.05	-	-
5	0.26	0.26	0.26	0.26
6	0.1	0.1	0.1	-
7	-	-0.07	-	-
8	0.04	-	-	-
9	0.19	0.19	0.19	-
10	-	-0.06	-	-

Table 1 shows that using the degrading degree computed in this paper, the QoS gains are above 0. It means that the local constraint extraction method can avoid useless re-selections effectively.

Simulate test set of data transmission speed according to the Gaussian distribution $N(10, 1)$. The threshold of reliability is 0.9. The size of test set is 100000. Compare the relations among the predicted results, predicted interval and the observed interval between two neighboring contexts. Table.2 gives the result (O_Q is the observed interval between two neighboring QoS-related contexts in the test set; N is the number of predictions; R is the average accurate rate of the predictions which can be computed as the ratio of the number of predictions that is right to the number of predictions; I is the predicted interval and the unit of I is second).

Table 2. Semi-Markov Based Predicted Result

O_Q	0.5s			0.1s			0.05s		
I	10	60	180	10	60	180	10	60	180
N	300	300	300	200	200	200	150	150	150
R %	95	86	80	97	93	83	98	95	92

Table.2 shows that if the observed interval between two neighboring contexts is smaller than 0.1s, although the predicting interval is relatively bigger (e.g. 60s), the predicted result (accuracy is 93%) is acceptable. If the observed interval is smaller (e.g. 0.05s), although the predicted interval is bigger (e.g. 180s), the accuracy of the predicted result (e.g. 92%) is also acceptable. Thus, through minimizing the observation interval, the accuracy of prediction result can be improved.

Randomly generate 10 scenarios with O_Q=0.1s, I=60s. A comparison of the extra delay by the following 2 approaches is presented in Fig.2.

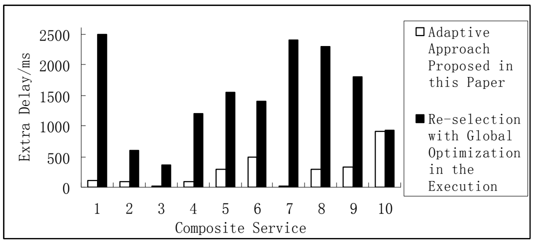


Figure 2. Comparison of Extra Delay

Fig.2 shows that the extra delay caused by the proposed approach is always the shorter than the re-selection in the execution. The reason is that on the one hand, the re-selection with local optimization can always minimize the extra delay while respecting the global constraint; on the other hand, by performance prediction, component services which will incur a QoS violation can be predicted, the reselection process aiming at finding a replacement will be more likely to finish before the invocation of the component service and thus the composite service can adjust itself from the failure as quickly as possible. In this way, the extra delay can be minimized.

The above experiments show that the proposed approach is effective in adaptive service composition.

5. RELATED WORKS

In this section, we will review works done in self-adaptive service composition.

The self-adaptive ability means that composite service can adjust itself if any execution problems occur, in order to successfully complete its execution, while respecting QoS agreements. The approaches can be divided into two categories: one is aiming at local QoS optimization, and another is for global QoS optimization.

The former one adopts multi-object policy to score candidate services and based on such score, re-selects the optimal candidate service for the service class to which the failed service belongs. For example, Ref. [11] proposes a service replication approach, in order to substitute the original component service when it is not available due to the heavy load of Network. Based on the idea of replication, Ref. [12] proposes a service composition approach based on redundant mechanism. The key to this approach is to establish a set of redundant services for each component service. Then, if one component service fails, the service can be replaced with an alternative member of the same redundancy group. In this paper, in order to achieve end-to-end constraint, we will attempt to research on issues of self-adaptive approach with global QoS optimization.

The self-adaptive web service composition with global QoS optimization adopts a global optimization policy to re-select the services in order to satisfy the end-to-end constraint. In Ref. [5], the reselection will be triggered as soon the actual QoS deviates from the initial estimates. When the failure is found, the execution of the composite service will be stopped until the reselection completes. The reselection will cost an extra delay of composite service. Thus, this approach can only be used for runtime-unaware application.

This paper tries to solve the problem of adaptive service composition with respecting for global constraint. Compared with the above works, this paper tries to minimize the extra delay caused by the re-selection by means of a performance prediction and a local constraint extraction.

6. CONCLUSIONS

We propose an effective and scalable approach for self-adaptive service composition. This approach uses local constraint and performance prediction. Then, the extra delay caused by the re-selection can be minimized and the useless re-selections can be avoided. In the future work, we will study on how to adapt the approach to deal with more complicated and realistic scenarios, where data transmission speed, failure rate and the cost of requesting a service are not assumed to be constant.

REFERENCES

[1] Milanovic N, Malek M. Current Solutions for Web Service Composition. IEEE Internet Computing, 2004, 8(6):51-59.

[2] Zeng L Z, Benatallah B. QoS-Aware Middleware for Web Services Composition. IEEE Transactions on Software Engineering, 2004, 30(5): 311-327.

[3] Yu T, Zhang Y, Lin K J. Efficient Algorithms for Web Services Selection with End-to-End QoS Constraints. ACM Transactions on the Web, 2007, 1(1): Article 1.

[4] Bonatti P A, Festa P. On Optimal Service Selection. In Proc. International World Wide Web Conference, Chiba, Japan, May 2005, pp.530-538.

[5] Canfora G, Penta M D, Esposito R et al. QoS-Aware Replanning of Composite Web Services. In Proc. International Conference on Web Services, Orlando, USA, July 2005, pp.121-129

[6] Yu T, Lin K J. Adaptive algorithms for Finding Replacement Services in Autonomic Distributed Business Processes. In Proc. International Symposium on Autonomous Decentralized Systems, Chengdu, China, April 2005, pp.427-434.

[7] Girish C, Koustuv D, Arun K et al. Adaptation in Web Service Composition and Execution. In Proc. International Conference on Web Services, Chicago, USA, September 2006, pp.549-557.

[8] Cardoso J, Sheth A P, Miller J A et al. Quality of service for workflows and web service processes. Journal of Web Semantics, 2004, 1(3): 281-308.

[9] Altinok Y, Kolcak D. An Application of the Semi-Markov Model for Earthquake Occurrences in North Anatolia, Turkey. Journal of the Balkan Geophysical Society, 1999, 2(4):90-99.

[10] Danilo A, Barbara P. Adaptive service composition in flexible processes[J]. IEEE Transaction on Software Engineering, 2007, 33(6): 369-384

[11] Jorge S, Francisco P S, Marta P M et al. WS-replication: A Framework for Highly Available Web Services. In: Proc. World Wide Web Conference, Edinburgh, Scotland, May 2006, pp.357-366

[12] Guo H P, Huai J P, Li H et al. ANGEL: Optimal Configuration for High Available Service Composition. In Proc. International Conference on Web Services, Salt Lake City, USA, July 2007, pp.280-287

[13] Dai Yu, Yang Lei, Zhang Bin. QoS-Aware Self-Healing Web Service Composition Based on Performance Prediction. Journal of Computer Science and Technology, 2009, 24(2)):250-261

Yu Dai is a lecture in Northeastern University. She receives her Ph.D degree from the college of compute science and technology at Northeastern University, Shenyang, China in 2008. Recent years, she has published more than 30 papers in “Journal of Computer Science and Technology”, “Chinese Journal of Computers”, “Journal of Information Science and Engineering”, “SEKE2008”, “DCABES2008” and other conferences as well as journals. Most of her papers are indexed by EI and SCI. Her paper in DCABES2008 has been selected as the recommended paper. Her main research interests include service composition and service oriented computing.

RESEARCH ON DYNAMIC SEMANTIC WEB SERVICES COMPOSITION BASED ONTOLOGY

Li Fan
Wuhan University of Science and Engineering
Wuhan Hubei 430073, China
Email: eslf@163.com

ABSTRACT

This paper discusses a novel method of automatic Web service composition based on ontology. A general dynamic composition service framework of ontology-based services composition system is proposed. A algorithm (HWSC) is put forward, which can fully use the semantic information of the inputs and outputs of Web services provided by the domain ontology and its inference ability. The preliminary experimental results show that it can satisfy the requirements of quality and efficiency of composition, thus generating composite Web services according to customer requirements automatically and efficiently.

Keywords: SSOA, Web Service Composition, Semantic, Ontology, OWL

1. INTRODUCTION

The semantic service-oriented architecture (SSOA) fusing semantic Web and Web service is a new-style software architecture. It is convenient for computers either to understand services or implement high level operations, e.g. intelligent reasoning. In order to really achieve this goal, SSOA need to resolve dynamic service composition in massive application services. Using Web service, complicated business processes can be created, but, only using it, dynamic automatic service composition is irresolvable. Accordingly, Web service composition enhanced by semantic technologies is one of the most hyped and addressed issue in SSOA.

Starting from an initial set of Web services, Web service composition aims at selecting and inter-connecting Web service provided by different partners according to a goal. Web services are typically described in terms of atomic and composite Web services, using languages like OWL-S [1] or BPEL[2] which provide mechanisms for Web service compositions. However, the processes of Web service compositions tend to be static in the sense that these processes are normally generated off-line.[3] Any changes to the part of a process may result in the reconfiguration of the whole process. This is especially critical for real-time enterprises since the systems cannot afford stopping, reconfiguring, and restarting.

There has been a great deal of research in the area of Web service in the past few years. They may often be the case that a Web service does not provide a requested service on its own, but delegates parts of the execution to other Web services and receives the results from them to perform the whole service. However, as the existing specifications and protocols can only describe services on the grammar level but not the semantic level, most composition methods are still not automatic.

In this paper, we propose an approach to realize automatic Web service composition based on domain ontology. The novelty of this method is to combine the domain ontology and the AI planning based algorithm to cover the factors such as service semantics, composition quality and composition efficiency.

2. SYSTEM ARCHITECTURE

2.1. OWL Ontology Model

OWL-S is an OWL ontology which supplies the service provider a core set of markup language constructs for describing the properties and capabilities of service. In OWL-S, a service is defined by three concepts: a Profile, a Model and Grounding. Generally speaking, the Profile provides the information needed for an agent to discover a service. Taken together, the Model and Grounding objects associated with a service provide enough information for an agent to make use of a service. For the service composition, we will focus on the Profile to match the input and output descriptions.

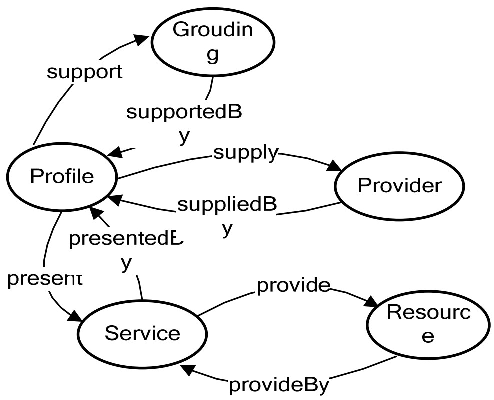


Figure1. OWL Ontology Model

A suite of ontologies are created to provide the service classification, a shared vocabulary for expressing service descriptions as well as reasoning processes to both manage the coherency of the classifications and the descriptions when they are created, and the service discovery, matchmaking and composition when they are deployed.[5] Each service description refers to a number of ontologies that define the domains of parameters like input, output, precondition, effect and a number of constraints between the parameters.

The model of our system's OWL ontology is shown in Figure 1. Concept Service will be used to classify services. It provides service on some resources. Each service is presented by many profiles, which are supplied by providers. The profile supports a grounding that specifies how the service should be invoked.

In this paper, our system uses OWL definition language to define ontology data set，concept relation set and domain rules set to form a complete Ontology Library. In this way, we can add semantic information in the process of service registry and implement service matching based on semantic information in the service composition. The system architecture is shown in Figure. 2

The UDDI Registry is a service registry container. It is in charge of storing the basic information of the service such as service name, service classification and service description etc.

Rule library define Web service's attributes that could be used in service composition. The rules consider the syntactic and semantic properties of Web services. Syntactic rules include the rules for operation modes and the rules for binding protocols of interacting services. Semantic rules mainly include the following subset: (1) Operation semantic composability defines the compatibility between the domains, categories and purposes of two services; (2) Qualitative composability defines the requester's preferences regarding the quality of operations for the composite service.

The Matchmaking Engine can find the services which meet the service consumer's requirement by using the domain ontology knowledge, domain rules and ontology-based service matching algorithm. Ontology Library is used to store the pre-defined domain concepts, the domain rules and the relationship among the concepts, helping Matchmaking Engine complete domain matching and service matching. When receiving requirement, Web Service Application Interface translate the requirement to Service Profile, which be send to the Matching Engine. Matchmaking Engine use Inference Machine which based on OWL-S Ontology and Web Service Description Libraty to match the function description, then output a series of composited services. It is not simple matching which based on key word, but based on semantics restriction.

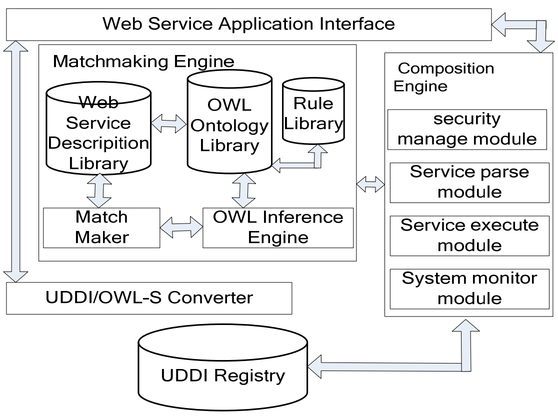


Figure2. Architecture of Service Composition

Service composition engine is in charge of running control and running environment. It is consist of four modules: security manage module, Service parse module, Service parse module and System monitor module.

3. WEB SERVICES COMPOSITION

3.1. Basic Definition

Domain ontology is the specification of the concepts and the relationship among them is specific domain.

Definition 1(domain ontology) Domain ontology can be described by a five tuple:

$$O = (C, R, H^C, rela, A^O)$$

C denotes a set of conceptions; R denotes a set of relation; H^C is the level of concepts; rela denotes the relation among concepts; A^O denotes the ontology axiom.

We adopt the OWL-S ontology to facilitate the definition of Web service. According to its model above-mentioned, each service has three service composition logic: Profile, Process and Grounding. Profile describes the semantic properties and capabilities of a service; Process represents the actual composition logic; and Grounding provides the physical binding information for runtime invocation access information. This paper mainly uses Profile, which includes the name, description, inputs, outputs, preconditions and effects of a Web service. Based on OWL-S, the following definitions are given.

Definition 2 C_i , C_j are two concepts in domain ontology. If C_i and C_j have the same semantic information, then $C_i \equiv C_j$; if the semantic information of concept C_i includes that of C_j, then $C_i \supseteq C_j$.

AS treating the Web service composition as a planning problem, it can be represented as a model of a four tuple $\Pi = (P, W, r^i, r^o)$, where P is a set of parameters; W is a set of Web services; $r^i \subseteq P$ is the input parameters; $r^o \subseteq P$ is the desired output parameters.This model defines a state space $\Psi = \langle S, s_0, S_G, \Omega, f, c \rangle$, where the states $s \in S$ are a collection of parameters in P; the initial state $s_0 \in S$ is such that $s_0 = r^i$; the goal state $S_G \in S$ is such that $r^o \subseteq S_G$; $\Omega(s)$ is the set of Web services $w \in W$ such that $w^i \subseteq s$. The function $f(w,s) = s'$ can map a state s into another state s' such that $s' = s \cup w^o$ for $w \in \Omega(s)$; c(w) is the invocation cost of w. A solution of the state model is a finite sequence of Web services: $w_1, w_2, \ldots, w_n$.

Definition 3 (Web service composition) Suppose that a request r has initial input parameters r^i and desired output parameters r^o, the Web service composition problem is to find a finite sequence of Web services, $w_1, w_2, \cdots, w_n$ such that (1) w_i can be invoked sequentially from 1 to n; (2) $r^o \subseteq (r^i \cup w_1^0 \cup \cdots \cup w_n^0)$; (3) The total cost $\sum_{i=1}^{n} c(w_i)$ is minimized.

Definition 4 (full matching, patially matching) Suppose that a state $s \in S$ is given, Web service $w_i \in \Omega(s)$. If for $w_2 \in W$, $w_1^0 \supseteq w_2^i$, then w_1 fully match w_2. on the contrary, if w_1 cannot fully match w_2 but $w_1^0 \cap w_2^i \neq \varnothing$, then w_1 partially match w_2.

3.2. Composition Algorithm

To add the semantic support for service composition, the most important step is to build the ontology library so as to standardize the service and achieve effective matching and intelligent reasoning based on semantic information. Many study have propounded some Web service composition approach based on interfaces, which seeks matches between input and output parameters to realize the dynamic composition. However, parameters with different spellings may have the same semantic meaning. So, only relying on the syntactic similarity of parameters to determine whether some Web services can be composed together is insufficient. Therefore, it is vital to find semantic relationships among parameters for interface matching.

The core of the method is to consider the Web service composition problem as a searching problem of the state space.

The first step of the method is to use domain ontology to describe the semantic meaning of the parameters of Web services and to find all relevant entailments such as the class inheritance relation between two classes that may not be directly encoded in the subclass relationships using OWL inferencing rules. According to definition 2, there are equal and subsumption relationships between ontology concepts. The match between the services whose output type is a subclass of the other services input type is called a generic match. According to the domain ontology and inferencing rules, the inference engine also orders the generic matches such that the priority of the matches are lowered when the distance between the two types in the ontology tree increases. The next step is to consider the Web service component problem as a search problem in state space, which finally satisfies the three requirements of definition 3.

We adopt a heuristic based greedy algorithm, which attempts to avoid a partial matching case by reducing the size of partial matching Web services as much as possible. The algorithm is divided into two parts: the forward search and regression search. First, we calculate the cost of achieving individual parameters starting from r^i using regression search; secondly, based on the first step, it approximates the optimal sequence of Web services that connects r^i to ro using regression search. we define $M_{r^i}(p)$ as the cost of achieving $p \in P$ from a state r^i. This cost can be characterized by the solution of the recursive equation as follows:

$$M_{r^i}(p) = \min_{w \in O_w(p)} \{c(w) + \max_{p \in w^i} M_{r^i}(p)\}$$

In the regression search, we adopt a heuristic-based greedy algorithm. The sub-goal in the algorithm is denoted as subGoal. We define W as a set of Web services$\in$W and $w_i \in PD_{ws}(p)$, $p \in$ subGoal. In each step of the regression search, we adopt the heuristic algorithm to select Web services from W. The heuristic based on a hypothesis is as follows:

Hypothesis: It may help reaching the initial state faster to choose a Web service with a greater contribution to match the sub-goal earlier in the search .

$$h_{sg}(w) = \left| w^0 \cap subGoal \right|$$

The algorithm is listed here:
Input : r^i,r^o
Output : w_1, w_2 ,…, w_n
For each $p \in P$
  If $p \in r^i$ then $M_{r^i}(p)$ =0
  Else $M_{r^i}(p) = \infty$
  s = ($r^i \setminus r^o$), C = $\varnothing$, t =1;
End For
While $\neg(s \supseteq r^o)$
  $\Lambda = \{w \mid w \in \Omega(s)\}$;
  For each $p \in w^o$ ($w \in \Lambda$)
    If $M_{r^i}(p) = \infty$
      $M_{r^i}(p)$ =t , $PD_{ws}(p)$=w , s=s$\cup${p};
  End For
    C= C$\cup$ Λ , t ++;
End While
s= ($r^i \setminus r^o$), subGoal = s
While $\neg(\text{subGoal} = \varnothing)$
  W= $\bigcup_{p \in subGoal} PD_{ws}(p)$
    x= $\max_{w \in W} h_{sg}(w)$
    s=s$\cup$(x \\r^i)
    sola=sola$\cup${x}
    subGoal= {subGoal$\cup x^i$}\s
End While
s=r^i
While $\neg(sola = \varnothing)$
  If $(w \in \Omega(s)) \wedge (w \in sola)$
    Print w ; s= s$\cup w^o$; sola=sola \ {w}
End while

4. EXPERIMENT EVALUATION AND CONCLUTION

To evaluate the efficiency of the method introduced in this paper (HWSC), we made a comparison with traditional Web services composition algorithm—BF* and a semantic-based best-result service composition method (BSC). All experiments were performed on the same platform; all the algorithms were implemented in Java. We adopted random generating Web services and random generating user goals as test sets. The experiment selects several data sets—the total number of services are 300,600,900,1200,1500,1 800, 2 100, 2 400,2700 plus 100 service composition requests to test the service composition and calculate the average consuming time. The experimental results are shown in Fig.3

From Fig.3, we can see that BSC has the lowest efficiency, since this method simply considers the service quality and at the same time sacrifices the efficiency. The performance of the BF* algorithm has greater fluctuation with the changing of the number of services, which is related to its inferencing rules.

HWSC has the best performance, since this method attempts to avoid a partial matching case by reducing the size of partial matching Web services as much as possible. The core of the method is to consider the Web service composition problem as a searching problem of the state space.

This paper discusses the common situation of service composition, current implementations assume that all services are atomic service request and the service invocation can be as simple as HTTP and internal functional call, or as complex as SOAP invocation. Our future work is to expand the algorithm proposed by this paper to more service composition situations, For example, several service outputs combined to semantically contain the input of the subsequent services.

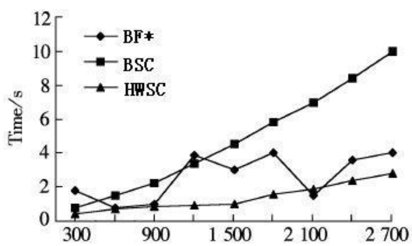


Figure 3. Composition Efficiency Comparison

REFERENCES

[1] Claude Moulin,Marco Sbodio. Using Ontological

Concepts for Web Service Composition[C]. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. 2005

[2] Semantic Web. http://www.w3.org/2001/sw. 2007.

[3] Web services discovery and composition as graph search problem [C] //Proc of the 2005 IEEE Intl Conf on e-Technology, e-Commerce and e-Service. Las Vegas, Nevada, USA, 2005: 784-786.

[4] PaolucciM, Kawm ura T, Payne T, et al. Semantic matching of Web services capabilities [C] // Proceedings of the First International Semantic Web Conference (ISWC). Sardinia,2002: 333-347.

[5] Yang Yanping, Tan Qingping, Yu Jinshan, et al. A new approach to development of composite Web services[J].Wuhan University Journal of Natural Sciences, 2006, 11(I):211-216.

[6] Massimo Paolucci, Takahiro Kawamura, Terry R Payne, et al. Semantic Matching of Web Services Capabilities[A]. Proceedings of the First International Semantic Web Conference[C].Sardinia,Italy:Springer-Verlag,2002.333-347.

[7] Ruoyan Zhang,I Budak Arpinar,Boanerges Ale-man-Meza. Automatic composition of semantic Web serv-ices[C] .Proceedings of International Conference on Web Services. Las Vegas,USA. 2003 :38～41

Li Fan, male, born in 1973.He received his Masters from Wuhan University of Technology. He is currently a Lecturer in school of computer science in Wuhan University of Science and Technology His research interests include E-business and integration of information system.

A GRID RESOURCE DISCOVERY MECHANISM BASED ON P2P *

Zuhua Guo, Zhiyong Du, Dan Liu, Shixia Ma
Computer Science and Technology Department,Henan Mechanical and Electrical Engineering College
Xinxiang,Henan 453002, China
Email: gzhgg@163.com; zhydu@126.com

ABSTRACT

In order to better achieve resource sharing of distributed system, the paper discusses resource discovery problems of grid, using the organic integration method of P2P and grid, which improves the dynamic and scalability of grid. This paper builds a grid resources architecture model based on P2P, and proposes a resource discovery algorithm based on DHT and ant colony algorithm. The theoretical analysis and simulation show that the algorithm can effectively improve resource discovery performance under the P2P Grid environment.

Keywords: Grid, P2P, Resource discovery, DHT, Ant colony.

1. INTRODUCTION

With the continuous improvement of computer performance as well as the rapid development of the Internet, working together and sharing resources among distributed systems has become an urgent demand of a wide range of network applications, and has been developing toward the destination of high-performance, large-scale, diversified, multi-purpose. Therefore, there needs to connect the various distributed, heterogeneous computing resources, storage resources, data resources, information resources, knowledge resources, expert resources with other special resources through high-speed networks to achieve the sharing of resources of high performance resources. Grid and P2P are all hot areas of sharing research on large-scale resources currently.

The essence of Grid is interconnecting and integrating the distributed geographically, heterogeneous high-performance computers, data servers, and large-scale retrieval storage systems etc., through high-speed coadjacent network, to form a logical whole so as to solve problems commonly. The virtual organizations can ultimately achieve sharing resources and working together in this virtual environment to completely eliminate information islands and most fully realize the sharing of resources [1].

P2P (peer-to-peer) is a computer peer-to-peer network, the basic idea of which is equal authority and openness, sharing of computer resources and services through direct information exchange. All nodes in these networks are peer to peer, each node has both resource providers as well as resources for consumers, which can add and withdraw randomly [2].

Grid uses the relatively centralized distributed resource management, tries to achieve accessing to all of the shared network resources, which is to share special heterogeneous resources in the trusted environment, and users in which have a high degree of security, confidentiality and reliability, but it does not resolve the issues that the node can join randomly, and the resource-sharing can be dynamic; P2P has flexible topology, nodes in which have a high degree of autonomy.
While the frequent join and withdrawal of nodes from the P2P networks has led to a high degree of dynamic, but there aren't QoS quality guarantee between nodes and the interaction isn't safe. Integrating P2P technology with the grid can improve the dynamics and scalability of grid, enhance the security of P2P, and solve the sharing of large-scale distributed, heterogeneous, dynamic resources.

2. THE GRID RESOURCE MODEL BASED ON P2P

The integration of P2P and grid is a hot area of research currently and there are many types of the common integrated programs. In literature [3], Nazareno Andrade etc. proposed a network resource management model known as the OurGrid community; In literature [4], Talia etc. proposed a network resource management architecture integrating Grid and P2P and complying with a standardized OGSA; In literature [5], Mastroianni etc. proposed a network resource management information services model of integrating Grid and P2P; In literature [6,7] in, Iamnitch, CaoJiannong etc. proposed a model, the bottom layer of which uses P2P mode and the upper layer of which uses of Grid mode.

The P2PGrid (P2P-based Grid Resource Architecture) model built by the paper is as shown in Figure 1, which uses a distributed multi-layer resource management model, provides an effective balance in the efficiency of the centralized management and the robustness of distributed resource discovery. The bottom layer is the grid resource layer, including various types of grid resources and services resources, in accordance with the functions and types of resources to form different virtual organizations (VO), each VO corresponds to a super node (SuperPeer); the intermediate layer is the network security layer, including the security information signing the credibility of resources; the top layer is the P2PGrid layer, which form the multiple SuperPeer into a large-scale P2PGrid in accordance with the ring topology.

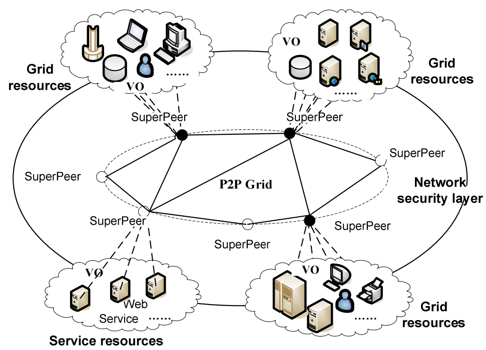


Figure 1. P2P Grid Model

The grid resource layer contains all resources which can be shared through grid, including computing resources, storage resources, communication resources, software resources, information resources, knowledge resources, services resources, external equipments etc., such as supercomputers, mass storage, equipment, application software, databases, a

* The Natural Science Research Project of Henan Province Education Department (2007520008).

variety of services and so on. Grid resources are geographically distributed, the functions of which are relatively centralized and which can work together in heterogeneous platform. A series of grid resources having specific functions constitute a virtual organization VO, which is controlled by SuperPeer as a super node, SuperPeer also can be regarded as the server of the Client / Server model in VO.
The multiple SuperPeers can be formed into a large-scale P2PGrid in accordance with P2P mode, which uses the ring topology. The SuperPeer node in P2PGrid has a high degree of autonomy and the P2PGrid can achieve dynamically joining and exiting of SuperPeer node. The SuperPeer in each layer of P2PGrid has dual role as clients and servers. On the one hand, as provider of resources, it allows its resources to carry out other super-node resources; on the other hand, as consumer of resource, it uses of local or idle resources of other super-nodes to perform its task.

Because each node in P2P relies on its own organizations, non-QoS quality assurance, in order to enhance the network security of P2PGrid, the model adds network security layer. On the one hand, the layer determines the reliability of resources by increasing the credibility value (past experience feedback between nodes) for SuperPeer to establish the inter-node's trust and security mechanisms; on the other hand, each node in grid resources has the identification, authorization , firewall protection, so as to have reliable service quality assurance mechanism.

When the user needs to find resources, first of all, the model will find the matching resources in the VO managed by SuperPeer by grid resource discovery mechanism, if not find, and then through SuperPeer, by P2P way to forward to other SuperPeer, the resources will be queried. Because the P2PGrid model changes the traditional grid resource discovery way, there needs to new algorithm to adapt to such P2PGrid resource discovery mechanism.

3. THE P2P-BASED GRID RESOURCE DISCOVERY ALGORITHM

Resource discovery mechanism is the key issue which is related to the efficiency of sharing resources and working together in the wide-area distributed environment. When the user requires the use of a resource, the user request, the corresponding resources request is submitted to the resource discovery mechanism; resource discovery mechanism, through some discovery strategy, finds a collection of resources in line with the request in the system, which is the target resources meeting the request conditions. At the same time, the resource information obtained by resource discovery will be returned.

The above-mentioned P2PGrid model combines with the grid technology and P2P technology, which needs a distributed, dynamic resource adaptation, easy expansion, and high-performance grid resource discovery mechanism.

Distributed hash table (DHT) algorithm is a matching way of structure resource based on keywords; grid resources are described and stored in the grid information base by keywords and attribute values. DHT, through performing the keyword-based matching in the grid information base, achieves searching and positioning grid resources, and eventually schedules the execution of grid tasks. This technology has some advantages such as certainty of searching, simplicity and distribution, and uses the structure topological relations between nodes to achieve efficient searching and routing.

The principle of the resource discovery algorithm based on ant colony algorithm is that, the information is transmitted through pheromone between the individual ants and the ants will choose to move to the path of high strength material, at the same time leave the exohormones secreted by their own. Using of the theory in searching the appropriate resources, the path load, delay and other factors can be considered to find the lowest integrated cost path and more efficiently use of network resources.

Make full use of the advantages of these two algorithms, the DHT algorithm and ant colony algorithm can be combined to achieve the distributed resources discovery with some advantages such as high credibility value, low integrated cost, accuracy and reliability in the P2PGrid model.

3.1 Prerequisite

Condition 1:P2PGrid = $\{SuperPeer_1, SuperPeer_2, \ldots, SuperPeer_n\}$, $SuperPeer_n$ of which is a super-node;
Condition 2: each $SuperPeer_n$ manages a VO_n;
Condition 3: $VO_n=\{Node_1, Node_2, \ldots, Node_m\}$, $Node_m$ is a grid resource node, each node maintains a routing table, routing table contains the identification and IP address of the node which has established a connection with the local node. When locating the resources, performing selective forwarding through the routing table, which can position any resources in certain Hop;
Condition 4: $Node_m = \{Key_m, Value_m, ID_{VO}\}$, Key_m is the keyword to identify resources, ID_{VO} is the identification of VO where resource is in, $Value_m$ is the value of resource, which is used to build local HashTable;
Condition 5: $SuperPeer_j = (SKey_j, RE_{ij})$, $SKey_j$ is the keyword identifying $SuperPeer_j$, RE_{ij} is the credibility value of $SuperPeer_j$, which is used to build Super-HashTable.
Condition 6: The topology between SuperPeer can be expressed by connected undirected graph G = (V, E), and G has no multiple edges;
Condition 7: The topology between SuperPeer is random and unpredictable;
Condition 8: SuperPeer can randomly and dynamically join and exit in P2PGrid;
Condition 9: Assume that the content of the pheromone in the links of P2PGrid layer is same initially.

3.2 Algorithm Design

Achieve integrating the DHT algorithm with the ant colony algorithm under the P2PGrid environment. Firstly, find the resources needed in VO managed by SuperPeer using DHT algorithm, if not found, in the layer P2PGrid, use ant colony algorithm to find the best and credible SuperPeer and then use the DHT algorithm to find resources needed in VO managed by the SuperPeer.
1) The user request identification is X, and the request will be sent to the $Node_m$ in current virtual node VO_i.
2) $Node_m$ will search local HashTable, if there exists the node providing X resources, then the resource discovery find is end; If there is no X, then the request information will be moved to the nearest node $Node_n$ in the routing table.
3) Repeatly search the routing table of $Node_n$ by step 2), if the node providing X resource in the current VO_i is found after O(logN) hops are completed, then the find is end; if there is no X, then go to step 6).
4) The current VO_i corresponds to the $SuperPeer_i$ node; judge the credibility of $SuperPeer_i$ to $SuperPeer_j$.

$$RE_{ij} = \sum_{m \in \mathrm{Re}\, G} R_{mj} CR_{im} \Big/ \sum_{m \in \mathrm{Re}\, G} CR_{im} \tag{1}$$

RE_{ij} is referred to the credibility value of the $SuperPeer_j$, ReG is the recommended set, R_{mj} is the local trust degree of the recommened node $SuperPeer_m$ to the node $SuperPeer_j$, CR_{im} is the credibility of the node $SuperPeer_i$ to the recommended node $SuperPeer_m$.

5) After calculate the credibility of the node $SuperPeer_j$ in the node $SuperPeer_i$, update RE_{ij} of the Super-HashTable.

$$\varepsilon = \left| RE_{ij} - R_{mj} \right| \Big/ S_j \tag{2}$$

S_j is the standard deviation of local trust degree of all recommended nodes to the node $SuperPeer_j$.

$$CR_{im}^{k+1} = \begin{cases} CR_{im}^{k} + \delta\left(1 - CR_{im}^{k}\right)(1-\varepsilon), 0 \le \varepsilon \le 1, k > 0 \\ CR_{im}^{k} - \gamma CR_{im}^{k}\left(1 - \dfrac{1}{\varepsilon}\right), \varepsilon > 1, k > 0 \\ 0.5, k = 0 \end{cases} \tag{3}$$

6) Use of ant colony algorithm, synthetically consider the bandwidth, delay jitter, CPU resources, trust degree etc., find a path P of resource request q to meet the most optimized and the most trusted conditions.

$$\begin{cases} f_1(e) = k_1 cost(e) + k_2 bandwith(e) + \mathrm{CPU(v)} \\ f_2(e) = k_3(delay(e) + delayv(v)) \\ f(e) = f_1(e) + f_2(e) + k_4(1/RE_{ij}) \end{cases} \tag{4}$$

Formula (4) is the generalized integrated cost function, k_1, k_2, k_3, k_4 is the weights of each factor. The topology of P2PGrid is connected undirected graph G = (V, E), e∈E. cost(e) is the cost function, bandwith(e) is the link bandwidth function, delay(e) is the link delay function; for any node v∈V, CPU(v) is a function of available CPU resources, delayv(v) is the node delay function. RE_{ij} represents the credibility value of $SuperPeer_j$, which can be get from step 4). Formula (4) should meet the constraint condition from (5) to (9).

$$\sum_{e \in P} \cos t(e) \le C \tag{5}$$

$$bandwith(e) \ge B, \forall e \in E \tag{6}$$

$$\sum_{e \in P} delay(e) + \sum_{v \in P} delayv(v) \le D \tag{7}$$

$$CPU(V_{dest}) \ge S \tag{8}$$

$$RE_{ij} \ge W \tag{9}$$

V_{dest} is the target node, C, B, D, S, W respectively represents meeting the requirements of bandwidth, link delay, integrated delay, CPU surplus utilization, the credibility value. Initialize each node's CPU remaining utilization, delay size, initialize each edge's cost size, bandwidth size, and delay size, and initialize each edge's pheromone. In accordance with the constraint condition from (5) to (9), select the next hop to ensure that f (e) is smallest.

7) Repeat step 6) to find the path optimal solution of resource $SuperPeer_j$.

8) Repeat step 2), 3) to find the resources needed in the VO managed by $SuperPeer_j$.

4. TEST RESULTS AND PERFORMANCE ANALYSIS

In the system based on the flooding, the overlay network is formed between nodes through the TCP connection, which rely on the information broadcast between the neighborhood nodes to find the necessary resources, that is in addition to the neighborhood node which will introduce the request, it is obvious that the node receiving message will develop transmitting (broadcasting) to all adjacent nodes. The P2P search system based on flooding system is inefficient and has also the problems on the way of the scalability because flooding uses up a lot of network bandwidth.

In order to verify the performance of DHT ant colony resource discovery algorithm of the P2PGrid model in this paper, we analyze a P2P topology simulator similar to the Gnutella's. Assume that in P2PGrid topology, there respectively contains the N = 100, N = 200, N = 300, N = 400, N = 500 SuperPeers. SuperPeers are all managed by the structured topology. The neighbor number of each SuperPeer is in line with the normal distribution, an average of which is N/20, and the standard deviation is 1.5. The neighbor node of each SuperPeer is randomly selected and connected through the routing table of SuperPeer。The resources can be indicated by adding an identifier to the routing table of SuperPeer, each identifier represents the name of a resource. The distribution density of resources is D, the unit is 1/100. Assume that the resource density of $SuperPeer_k$ D has two situations: D1 = 3, D2 = 5. SuperPeer can dynamically enter and exit the P2PGrid environment.

Each group experiment conducted 10 experiments to get average value in accordance with the nodes' number of SuperPeer in topology. We compared the flood method and the DHT ant colony algorithm. The basic standards of the comparison are:

Average Hops: It is the average hops' number of the resources being found under the condition of the same nodes' number, as well as the same resources distribution;

Average Success rate: It is the average success rate of the resources being found under the condition of the same nodes' number, as well as the same resources distribution. Figures 2 and 3 respectively shows the average hops' number and the average success rate of the resources being found under the same conditions in the flooding and the DHT ant colony algorithm.

Figure 2 shows the performance of the DHT ant colony algorithm is excellent than the flooding in the average hops' number of the resources being found under the same conditions, particularly under the condition of the increasing number of SuperPeer. In addition, with the reducing of the resource density, the average hops' number of the resources being found of DHT ant colony algorithm resources is also to reduce. In accordance with the average search success rate, Figure 3 also shows the performance of the DHT ant colony algorithm is higher than the flooding's. With the reducing of the resource density, the search success rate of DHT ant colony algorithm also decline。Therefore, the success rate of finding the resources is also related to density.

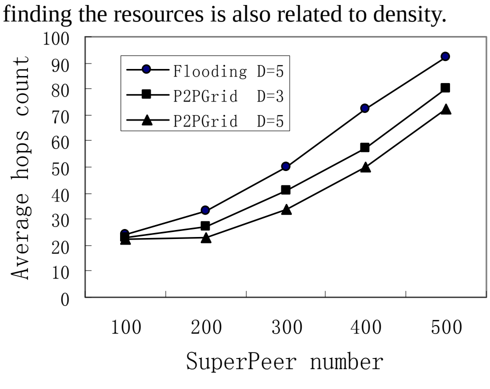


Figure 2. The Average Hops' Number of The Resources Being Found Under The Same Condition

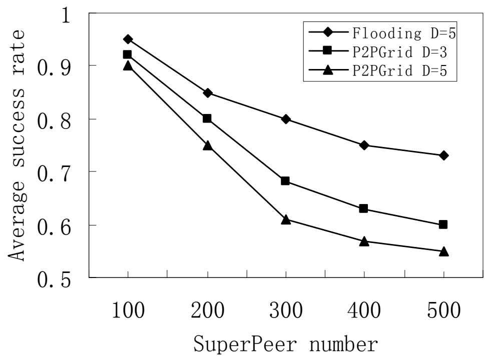


Figure 3. The Average Success Rate of The Resources Being Found Under The Same Condition

5. CONCLUSIONS

The paper builds the P2PGrid (Grid Resource Model based on P2P), the model introduces the credibility value of the node to enhance the security of P2P, and proposes the DHT ant colony resource discovery algorithm, which can accurately and reliably develop distributed resource discovery with high credibility value, low integrated cost. The simulation verifies the advantages of DHT ant colony algorithm

REFERENCES

[1] David De Roure,Mark A Baker,Nicholas R Jennings,et al.The evolution of the grid[J].Grid Computing making the Global Infrastructure a Reality,2003,12(3):65-100.

[2] Paolo Trunfio,Domenico Talia,Paraskevi Fragopoulou, et al. Peer-to-peer models for resource discovery on grids[C].Proc of the 2nd CoreGRID Workshop on Grid and Peer to Peer Systems Architecture.Paris,2006:1-23.

[3] Nazareno Andrade,Lauro Costa,Guilherme G,et al.Peer–to-peer grid computing with the ourgrid community[C].Proc of the SBRC 2005.Brazil,2005:1-8.

[4] Talia D,Trunfio P. Peer-to-peer protocols and grid services for resource discovery on grids[J].Grid Computing:The New Frontier of High Performance Computing,Advances in Parallel Computing,2005, 14(2):1232-1243.

[5] Mastroianni Carlo,Talia Domenico,Verta Oreste.A super-peer model for resource discovery services in large-scale grids[J].Future Generation Computer Systems,2005,21(10):1235-1248.

[6] Iamnitchi A,Foster I.A peer-to-peer approach to resource location in grid environments[C].Grid Resource Management.Kluwer,2003:413-429.

[7] Cao Jian-nong,Liu Fred B.P2P grid:Integrating P2P networks into the grid environment [C]. GCC2005. Beijing,2005:871-883.

Guo Zuhua, is a lectorate of Computer Science and Technology Department of Henan Mechanical and Electrical Engineering College. She graduated from Zhengzhou University with specialty of computer sodtware in 1999; from Xi'an Jiaotong University in 2002 with specialty of electronic communication engineering. Her research interests are in application of Internet and network security.

QOS-AWARE WEB SERVICE SELECTION ALGORITHMS BASED ON SMALL-WORLDS NETWORK

Xiaoming XU, Nan JIANG , Qiuling DING
College of Information Science and Technology,
Nanjing University of Aeronautice and Astronautiec,Nanjing,210016,China
Email: Roger2001xu@163.com

ABSTRACT

In this paper, a theory based on the small world of a QoS-aware Web Service Selection Algorithm was proposed. This algorithm use small-world network theory and ant colony algorithm, not only provide better optimization path rapidly, but also optimize load balancing on the service path. It solves the shortcoming that only considering the choice of the optimal path the common algorithm, losing sight of making the optimal path becoming the hottest path because of the frequent choice of the optimal path, on the contrary, that making optimal path turn into the bottleneck of the system's. Simulation and experiment showed that the algorithm has efficient services selection performance.

Keywords: Web Service Selection, QoS-aware, Small-world Network, Ant Colony Algorithm

1. INTRODUCTION

Web service is composed by the business services and business components, adapt to the needs of B2B,and the development strategies of the enterprise's information technology of next generation. It uses scalable language XML (eXtensible Markup Language) definition of web services protocol stack, through SOAP, WSDL, UDDI, WSFL, BPEL4WS and other open protocols and standards of services across different platforms and organizational structure [1,2,3], to provide unified service registration, discovery, binding and integrated call mechanism for Internet applications. In essence, WEB service is a self-describing , modular, new model of distributed computing. Open standards make Web services have good interoperability, self-describing, modular features allow developing loosely coupled application components in any platform by any programming language, and through appropriate mix of services in support of integration and development for application. But how to dynamic select, bind, and call the Web service of most appropriate user needs is one of the issues that need to be resolved urgently. At the same time because service based on QoS for web services in the successful application of business is very critical, so that selecting, binding and invoking dynamically the most appropriate services based on user for the target service demand for QoS become a hot research.

In 1998, Watts and his tutor named Strogatz revealed small-world network which has the characteristics of high clustering coefficient and low average path length[4], that is, small-world network has the characteristics of "six degrees separation". The method of its construction: re-connect of each edge randomly with the probability of p in a ring of N vertices coupling network, so that there realize evolution from rule network to stochastic network by adjusting the p value. At the same time, Frutos brothers discovered that node or web page for the connectivity is power law relationship in the process of observation of Internet network, and this relationship exists in a variety of network models[5, 6].SOAP protocol does not handle display routing in the Web service, WS-Routing and WS-Referral prescribe message path and their dynamic configuration [7]. QoS is the key to affection on its routing path selection in Web services, Web services routing path selection based on the QoS-aware is the key to manifest on-demand, dynamic service selection .

In this paper, we absorbed the ideals and approach of small-world network theory proposed in the literature [8], QoS-aware Web service selection algorithm was proposed in the base on literature [8,9,10], and by improving ant colony algorithm [11,12]. The rest of the paper is organized as follows. In section 2, we describe Small-world web Service network construction; In section 3,we introduces QoS-based Constraints Web services selection model; In section 4,we studied applying small-world network theory to QoS-aware Web services selection and its detailed design; In section 5,we describe the result of the experiment and simulation; Finally, a summary of this paper and the idea of further work.

2. SMALL-WORLD WEB SERVICE NETWORK CONSTRUCTION

2.1 Characteristics of Small-world Network Structure

Small-world network can expressed using the characteristics network path length (L) and similar system coefficient γ .

Definition 1 The characteristics path length L of a graph is the media average of the shortest path length that each vertex $v \in V(G)$ is connected to the indexing other vertex.

Definition 2 Γ_v similar Coefficient γ_v for any vertex v , degree of rapprochement is $\gamma_v = |E(\Gamma_v)| / (\frac{k_v}{2})$ between its adjacent vertex. $|E(\Gamma_v)|$ is the number of edges in region of vertex v , $(\frac{k_v}{2})$ is the total number of the edges that Γ_v may exist. k_v is the numbers of Vertices that given graph Γ_v . The similar coefficient γ of the graph G is the average of similar coefficient all the vertices.

Most of the node connectivity is smaller in the small-world network, but connectivity of a small number of nodes is highly. We have a small number of nodes that connectivity is highly as the central node (or called Hub node), which identifies the role of cluster, at the same time it takes advantage of some node that connects the different regions (known as a shortcut),it makes its range of network connection wider and available resources richer and querying or searching more efficient.

2.2 Build the Small-world Web Services Networks

The basic principles that build the small-world networks of web Service: Aggregating the similar services based on the

Partial ordering sets of the key attributes of Web services, and forming the aggregation sets. so Web services can be expressed similar node cluster, choosing the service node in the cluster as a Hub node, connecting all nodes of the service cluster, and then connect the Central Hub into ring, and then selecting node random in the ring to connect as a shortcut, so that Web services network with a small-world characteristics is constructed.

2.2.1 Reselection

That selecting the Hub node is based on the size of vertex degree. selecting the vertex which degree is the largest as a Hub node. The approach is : each vertex select randomly vertex to send message in the service sets, calculating the amounts of the each vertex received, sorting in descending according to the amounts of the message, the vertex of the largest message amount as hub node, the sort of sequence as a standby hub node. Then selected hub node connect all the vertices, so that forming the equal-type connections. while the current node fails, standby hub node make a key node, which controls access loop and manages host service set.

2.2.2 A shortcut to The Connection

The function of weak link ensures the connection between the various clusters made up of the hub nodes in small-world, it is also the key of the smallest characteristics length in carrying out small world network, and also the key to web service path. Web services have comparability to the nature of the attachment, i.e. similar service node may be gathered in a hub nodes , and the new service will be dependent in a similar hub nodes. The weak link connects the cluster by the hub node gathered, forming small-world property network.

To determine the connection shortcut by the following steps to.

Step 1：each hub node generated by all services cluster uses label T logo each hub node to form a loop;

Step 2: determine the short-cut numbers $s=[N*a]$, N is the number of hub node, a is the coefficient with values of 0.1 ~ 0.3;

Step 3: selected s vertex randomly, connecting to the label T_j ($T_j=(T_i+[i/2]) \bmod N, i\in[1,s]$) vertex to create a shortcut. T_i is the i vertex label.

After all of the upsides services gathered, hub node selected, a shortcut connected, the network structure have the structure as shown in Figure 1. In the structure, each service cluster, its similar coefficient $\gamma=\dfrac{2(N-2)}{N\bullet(N-1)}$, when N>10, γ <0.2, the characteristics path length L ≤ 6, which have a small-world network characteristics.

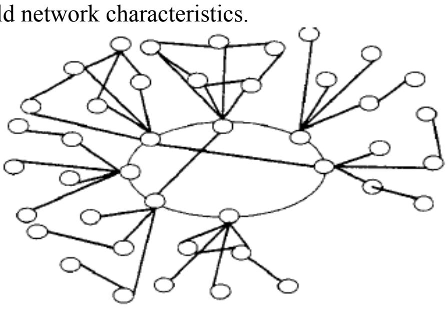


Figure1. Web Service Small-world Network Model

2.2.3 Web Services Network Maintenance

The connection of Service node have characteristics of attaching to preferences , the network size increases with adding new nodes. New nodes tend to connecting to the existent node that have the greater connectivity, that is to say, the new service select the service cluster and connect to the hub node of this cluster in accordance with the attributes of service.

When some hub node of service cluster fails, the standby node centralized in the hub service will capture on the connective vertex information around the hub node, then become the current hub node and take over the all valid node in the current service cluster. When the standby nodes are more than valid node one-half in the service cluster, re-select the hub node; Otherwise, select node as hub node randomly.

Using token ring management mechanism manage the hub node .using following method determine the status of hub node in the loop: assign randomly the vertex to manage the token transceiver, if sending and receiving tokens fails, which shows hub node failure, poll and locate the hub node fault. when the fault location come forth new and effective hub node, the token continue to sending. If the hub node of the managing the token fails, then randomly assigned node to manage token, and poll the token, in order to ensure the loop reliability and stability.

3. QOS-BASED CONSTRAINTS WEB SERVICES SELECTION MODEL

In the small-world networks, Web services once determine the source of target node and destination node, the next work is how to find an optimal path to meet the QoS constrained requirements efficiently and to obtain services. Combination of services can be expressed as a directed graph G = <V,E> according with a small world network, V is the collection of combinatorial service nodes, E is collection of all edges in the graph, edges denote the combinatorial path of the adjacent service node and they are arrived. Literature [13] gives out the three-dimensional model of QoS of Web service, which is more scientific, reasonable than the model proposed in the other literature [14,15,16], but There is no effective solution to the comprehensive QoS performance. In this paper, the concept of utility function based on the above-mentioned is proposed to solve the deficiencies of the past QoS model.

Definition 3 Service utility refers to that running the services can reach the level of the overall expected . Service utility is Constrained by function, Qos indicators, using method and other factors. Select n item key indicators from QoS service model, using measure matrix QS to measure and evaluate the QoS indicators for each service in S = {s1, s2, ..., sm}.

$$Q_s = \begin{pmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & \ddots & \vdots \\ q_{m1} & \cdots & q_{mn} \end{pmatrix} \tag{1}$$

Because different types of QoS indicators, there is inconsistency in its dimension, in order to facilitate the calculation and evaluation of service utility, there should standardize the indicators so as to solve the consistency between different indicators and to facilitate comparison and synthesis computing. Literature [17] focus on various quantitative classification, Aiming at benefit type and index type, QoS indicators will be standardized in this paper.

Benefit type QoS Standardized:

$$q_{ij}^{'} = \begin{cases} q_{ij} / \sqrt{\sum_{i=1}^{n} q_{ij}^2}, & if \sqrt{\sum_{i=1}^{n} q_{ij}^2} \neq o \\ 0, & if \sqrt{\sum_{i=1}^{n} q_{ij}^2} = o \end{cases} \quad (2)$$

Cost type QoS indicators Standardized:

$$q_{ij}^{'} = \begin{cases} \sqrt{\sum_{i=1}^{n} q_{ij}^2} / q_{ij}, & if \sqrt{\sum_{i=1}^{n} q_{ij}^2} \neq 0 \\ 0 & , if \sqrt{\sum_{i=1}^{n} q_{ij}^2} = 0 \end{cases} \quad (3)$$

Based on formula (2) and (3), after treatment of the matrix get the matrix as follows:

$$Q_s^{'} = \begin{pmatrix} q_{11}^{'} & \cdots & q_{1n}^{'} \\ \vdots & \ddots & \vdots \\ q_{m1}^{'} & \cdots & q_{mn}^{'} \end{pmatrix} \quad (4)$$

The definition of utility function:

$$F_E(s_i) = \begin{cases} \sum_{j=1}^{n} w_i * q_{ij}^{'} + k, & if \forall j, 1 \le j \le n, r_j(qij, u_j) \ \ is\ true \\ 0 \quad , & if \exists j, 1 \le j \le n, r_j(qij, u_j) \ \ is\ \ false \end{cases} \quad (5)$$

W vector $(w_1, w_2, \ldots, w_m)$ denotes the weight of each QoS parameters for the effectiveness, which reflects the relative importance of the n item QoS indicator, as well as a quantitative for user preferences, U vector is setted QoS benchmark value for a user or system; k is the quantification that other factors have an impact on the utility in addition to the QoS of service, only service with the same functions and usage patterns, its k value is the same. So only when according with the n-QoS indicators basic requirements, the utility function play a role. There establish relationship to the user QoS requirements and service capabilities of the service providers through the utility function, and be specific quantitative.

4. ALGORITHM DESCRIPTION

4.1 Mathematical Model

In the literature [18], the authors propose an improved ant colony algorithm for any cast routing algorithm, routing and Any cast Path Selection are very similar. In this paper, the ideas for any cast routing use to choose the path of Web services for small-world network characteristics.

（1）Transition probability

Node i to node j of the transition probability $p_{i,j}(t)$ as follow:

$$p_{i,j}(t) = \begin{cases} \dfrac{\tau_{ij}(t)^{\alpha} \eta_{ij}(t)^{\beta} \Gamma_{ij}(t)^{\gamma} \xi_{ij}(t)^{\sigma}}{\sum_{s \in A_k} \tau_{is}(t)^{\alpha} \eta_{is}(t)^{\beta} \Gamma_{is}(t)^{\gamma} \xi_{is}(t)^{\sigma}} & if\ j \in A_k \\ 0 & if\ j \notin A_k \end{cases} \quad (6)$$

Where α is heuristic factor for the information, which expresses the relative importance of the services path, and which reflects that information accumulated in the process of ants movement has effect on the ants campaigns. $\tau_{ij}(t)$ is the pheromone on the path ij (the path between nodes i and j); β is expected heuristic factor, reflects that heuristic information of the process of ants in the movement attaches importance to how important for ants select the path, $\eta_{ij}(t)$ is visibility on the path ij, it is a heuristic function for selecting path, expression using the formula (7); $\Gamma_{ij}(t)$ is the remaining service capacity of the node j, γ is the heuristic factor; σ is the heuristic factor of load information ,and $\xi_{ij}(t)$ is food odor between node i and node j, Which food odor refers to the goal service node broadcast the capacity of the remaining information to all service nodes periodically.

The purpose to the introduction of $\Gamma_{ij}(t)$ is to ensure choice of services to achieve load balancing, to avoid optimal path becoming the hottest path, and reduce system operating efficiency; the introduction of $\xi_{ij}(t)$ is to ensure that ants can find the optimal target service.

$$\eta_{ij}(t) = \frac{1}{D(i,j)} = \frac{1}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}} \quad (7)$$

$D(i,j)$ is the Euclidean distance function between nodes i and node j.

Taboo table $tabu_k$ used to record ant k has gone through a collection of nodes, A_k is node i's next hop nodes allowed to choose a collection, and

$$A_k = \{N_i - tabu_k\} \quad (8)$$

（2）Global update rule for pheromone

When the end of a cycle, according to the formula (9) to carry out the adjustment of pheromone:

$$\tau_{ij}(t + \Delta t) = (1 - \rho_{\tau})\tau_{ij}(t) + \rho_{\tau}\Delta\tau_{ij}, \quad \rho_{\tau} \in [0,1), \quad (9)$$

$$\rho_{\tau} \in [0,1),$$

Where ρ_{τ} is pheromone update coefficient between $t + \Delta t$, $\Delta\tau_{ij}$ is incremental pheromone in time Δt ,

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \quad (10)$$

$\Delta\tau_{ij}^{k}$ is the ant k left the amount of information between the inner path ij and between time t and $t + \Delta t$,

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{\min I(k)}{\Omega(k)}, if\ i\ is\ a\ general\ node, \\ \dfrac{\min \Pi(k)}{\Omega(k)}, if\ i\ is\ a\ hub\ node; \end{cases} \quad (11)$$

$\min I(k)$ is the cost to choose to an ordinary node, $\min \Pi(k)$ is the cost to choose the central node, $\Omega(k)$ is the total service cost to choose service path.

（3） Local update rule for pheromone

local adjustment and Simulation for the pheromone evaporation process. For any path, with the time goes, the past left pheromone will volatilize gradually, after time ΔT , the pheromone in the path ij update in accordance with (12):

$$\tau_{ij}(t + \Delta T) = (1 - \delta_{\tau})\tau_{ij}(t) + \delta_{\tau}\tau_0, \quad \delta_{\tau} \in (0,1) \quad (12)$$

δ_τ is the evaporation coefficient of pheromone; τ_0 is the minimum pheromone. Using formula (13) Computing.

$$\tau_0 = \begin{cases} \dfrac{1}{\sqrt{\dfrac{2\ln N}{\pi N}}}, if\ i\ is\ a\ general\ node, \\ \dfrac{1}{\ln N\sqrt{\dfrac{2\ln N}{\pi N}}}, if\ i\ is\ a\ hub\ node; \end{cases} \quad (13)$$

（4）Global update rule for load information

Global update rule load information is that the network will disseminate the web server's available resources of information to all nodes, odor in the dissemination process is influenced by the distance and other factors. update rule is formula(14):

$$\xi_{ij}(t+\Delta T_f) = \rho_\xi * \xi_{ij}(t) + \Delta\xi_{ij} \quad (14)$$

Where ρ_ξ is odor update coefficient in period of the time t to $t+\Delta T_f$, $\Delta\xi_{ij}$ is the incremental information in period of time ΔT_f ,as formula(15):

$$\Delta\xi_{ij} = \sum_{k=1}^{n}\Delta\xi_{ij}{}^{k} \quad (15)$$

Where n is the number of odor which reach the target server, $\sum_{k=1}^{n}\Delta\xi_{ij}{}^{k}$ is odor of the k th nodes service left in the path ij, and as formula(16):

$$\sum_{k=1}^{n}\Delta\xi_{ij}{}^{k} = \begin{cases} \dfrac{E_k\sqrt{\dfrac{2\ln N}{\pi N}}}{E_i H(k,i)}, if\ i\ is\ a\ general\ node, \\ \dfrac{E_k \ln N\sqrt{\dfrac{2\ln N}{\pi N}}}{E_i H(k,i)}, if\ i\ is\ a\ hub\ node; \end{cases} \quad (16)$$

Ek is the surplus energy of the kth nodes of the services, Ei is the total energy of the ith nodes, H (k, i) is the number of which transit nodes from starting node reaching the node i, the formula (16) which ensure that the food will be in accordance with the concentration of odor from distance and different, ants will choose the path of the service node that odor is dense.

（5）Local update rule for load information

Local odor update Simulate the odor of food volatile cyclical process,update rule is formula(17):

$$\xi_{ij}(t+\Delta T_f) = (1-\delta_\xi)\xi_{ij}(t) + \delta_\xi\xi_0 \quad (17)$$

Where δ_ξ is the volatile Coefficient, $1-\delta_\xi$ is the Residual odor factor, ΔT_f is odor update cycle, ξ_0 is the minimum of update.

$$\xi_0 = \begin{cases} \dfrac{E_{\min}(S)\sqrt{\dfrac{2\ln N}{\pi N}}}{E_i}, if\ i\ is\ a\ general\ node, \\ \dfrac{E_{\min}(S)\ln N\sqrt{\dfrac{2\ln N}{\pi N}}}{E_i}, if\ i\ is\ a\ hub\ node; \end{cases} \quad (18)$$

$E_{\min}(S)$ is to the critical energy value of the service node.

4.2 Algorithm Implementation

(1) Initialize parameters

(2) Confirm critical transmission range and services nodes.

(3)Wait command.

(4)Trigger service path selection.

(5)Ei ≥Ec.

(6) Put m ants to service nodes.

(7) k← k + 1 (k is the index of tabu list) .

(8) m ants select next node following transition probability pi, j (t) (j∈Ak) .

(9) Modify the index of tabu list, and move a new node to the list of ants.

(10) If k <m, goto (7); else, goto (11).

(11) Record the optimal path at this iteration.

(12) Update pheromone following Eq. (9).

(13) Timer Δ t.

(14) Update pheromone following Eq. (12) .

(15) Timer Δ Tf.

(16) Update food odor following Eqs. (14)and (17) .

(17) If Ei < Ec , end loop and output the results; else empty all tabu lists and goto (4).

(18) Goto (3).

5. EXPERIMENTS

We test the network structure in Figure2, each vertex V = (S, D, M), the S, D, M, denotes the node delay, node node loss rate and delay variation respectively. Edge E = (C, B, L), the C, B, L, denotes the cost of bandwidth and link delay respectively.

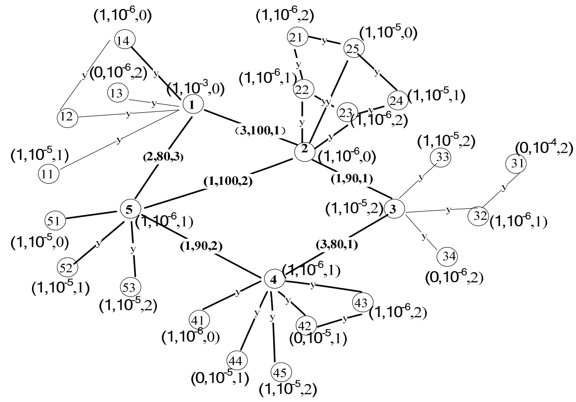


Figure 2. Network topology and parameters

QoS constraints Effectiveness $F_E(s_i)$ use (B_w, D_w, L_w, J_w) expression, its value is (80,12,10-5,3), the initial value set of the algorithm is: α=1,β=2,ρτ=0.2,δτ=0.2,MAX=100,ΔT=10,ρ=0.2,δξ=0.2,m=20, the outside of the loop is initialized with a random function, the results of the path of choice as shown in Table 1.

Table 1 results of service path selection based on small-world QoS aware

Request path	Selection path	Cost	delay
(43,21)	43→4→5→2→25→21	7	9
(34,52)	34→3→2→5→52	5	8
(14,23)	14→1→2→23	5	8

In this paper, the Comparison of the proposed algorithm and conventional ant colony algorithm as shown in Table 2.

Table 2. Simulation experiment results

algorithm	The average number of iterations	The average running time (s)	The average accuracy rate (%)
QoS based on the conventional ant colony algorithm	327	100.3	78
Proposed algorithm	158	67.4	82.4

The experimental results from the simulation can be seen, compared to conventional services based on ant colony routing algorithm ,the algorithm proposed in this paper is to choose the path of service, has obvious advantages.

6. CONCLUSIONS

In this paper, we have small-world network theory applied to Web services, in the establishment of Web selection model based on QoS constraints. we proposed based on the theory of small world-based QoS-aware Web service selection algorithm, ant colony algorithm in the basis of general , the use of small-world theory, an increase of food smells at the same time to disseminate the process, making all nodes and links can be server-to-date information, choose the path of an effective solution to service the dynamic load balancing problem, simplifying the complexity of multi-target multi-path option. Future work will focus on increasing self-adaptive algorithm based on machine learning algorithms, to better serve the Web services selection based on QoS-aware.

REFERENCES

[1] Burstein M H,Hobbs J R,Lassila O,et a1.DAM S:Web services description for the semantic Webl C 1.In:Horrocks,ed. Proc of the Int’l Semantic W eb Conf,Sardinia:Springer-Verlag,2002.348～ 363

[2] Bell D，Bussler C，Yang Jian．The Semantic Web and Web Services[J]．Information Systems，2006，31(4—5)：229~231

[3] Clark n Next—generation Web servicesl J 1．IEEE Internet Com—puting，2002，6(2)：12～ l4

[4] Watts D J,Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393 (6684):440-442.

[5] Kleinberg J．The Smal1、IVbr1d phenomenon：An algorithmic per—spective[D]．In：ACM Symp．on Theory of Computing，2000

[6] Wang Xiaofan，Chen Guanrong． Co mplex networks small-worldscale-free and beyond[J]．IEEE Circuits and Systems Magaxine，2003，3(1)：6～ 20

[7] Ballinger K ． The architecture of web services[M]．Beijing: China Electric Power Press，2004．180～191.

[8] Helmy A. Small worlds in wireless network,IEEE Communications Letters,2003,7(10):490-492.

[9] XU Xin-Wei,XU Xiao-Ming,ZHOU Liang,Ding Qiu-Lin.An Algorithm of Web Routing in Small-worlds Network,Computer Science，2007，34（9）：48-51.

[10] Jiang Nan, Yang Shuqun, Li Lingzhi and Ding Qiulin. Power-aware Anycast Routing in Wireless Sensor Networks Exploiting Small World effects, Journal of Southwest Jiaotong University，vol. 16, no. 4, pp. 366-371, 2008.

[11] DORIGO M,GAMBARDELLA L M. Ant colony system:a cooperative learning approach to the traveling salesman problem[J]. IEEE Transactions on Evolutionary Computation, 1997,1(1):53-66.

[12] DORIGO M,MANIEZZO V,COLORNI A.The ant system:optimization by a colony of cooperating agents[J].IEEE Transactions on Systems, Man, and Cybernetics, 1996,26(1):1-13.

[13] GUO De-Ke,REN Yan,CHEN Hong-Hui,LUO Xue-Shan.A QoS-Guaranteed and Distributed Model for Web Service Discovery, Journal of Software,Vol.17,No.11,November 2006,pp.2324-2334.

[14] Liu YT,Ngu AHH,Zeng LZ.QoS computation and policing in dynamic Web service selection.In:Feldman SI,uretsky M,NajorkM,Wills CE,eds.Proc.of the 13th Int’l Conf.on World Wide Web(WWW 2004).New York:ACM Press,2004.66-73.

[15] 15.Tian M,Gramm A,Ritter H,Schiller J.Efficient selection and monitoring of QoS-aware Web services with the WS-QoS framework.In:Liu J,Cercone N,eds.Proc.of the IEEE Int’l Conf.on Web Intelligence(WI 2004).New York:IEEE Press,2004.152-158.

[16] Yu T,Lin KJ.Service selection algorithms for Web services with end-to-end QoS constraints.In:Hsu MC,Zhang LJ,eds.Proc.of the IEEE Int’l Conf.on E-Commerce Technology(CEC 2004).New York:IEEE Press,2004.129-136.

[17] L IU Shu2lin , QIU Wan2hua. Studies on the basictheories for MADM[J] . Systems Engineering - Theory& Practice , 1998 , 18 (1) :38 - 43.

[18] Li LingZhi,Zheng Hongyuan,Ding Qiuling. An Anycast Routing Based on Improved Ant Colony Algorithm,Journal of Electronics & Information Technology,2007,2007,29(2):340-344

Xiaoming XU ,born in 1972, Ph.D. candidate. His research interests are in the areas of Web services, Web intelligent computing, and Information System Integration.

A SCHEME OF GLOBAL SERVER LOAD BALANCE IN CDN SYSTEM

Hu XiaoFeng [1,2] , Wang FangSong [3], Meng XiangYe [2], Wan Jing [4]
[1] Wuhan University of Technology
[2] National Cultural Information Resources Sharing Project
[3] Beijing Kong Yue Technology Co., Ltd.
[4] National Library of China
Email scorpionhxf@sohu.com

ABSTRACT

In order to provide best content service, the CDN (Content Distribution/Delivery Network) technology was invented. The Achievement of CDN depends on multiple Internet technologies; the important one is balanced Load technology. First, Global Server Load Balance and Balanced Load technology are introduced which are based on user-side detection, then, for attaining better effect, The two kinds of technologies are combined which can make up each other.

Keywords: Content Distribution; Balanced Load; CDN

1. INTRODUCTION

CDN (Content Distribution Network) is a holistic system that has been tactically deployed. By judgment of accessibility and server load, CDN make sure the Content be delivered to users' requirement with an efficient way. Generally speaking, the perfect goal is that the selected edge server is single hope far away from the user. Load Balance system is the core of CDN, so the accuracy and efficiency of Load Balance directly determine the efficiency and function of CDN. It decides and judges by the request of users and status of edge server, then chooses the best edge server.

At present, the CDN scheme Of Global Load Balance put forward, mostly acquires the load status from every edge server actively or passively,and on that basis chooses the optimal edge server. Generally speaking, the functional requirement of load balance and visit attemper contains several aspects hereinafter: ①To understand Health and load state of every CDN nodes, mainly to Find out real time laden state of each nodes and health of each server, so as to make correct decisions and reacts when the interview requirements of users get; ②To temper the users' requirement to the node which has the right content, and according to the net currency, connect and load status of each node beside of the distance and react time, delivering the users' requirement to the nearest normal serving node.③Interior distribution in node: To provide load balance between servers on original node and the other nodes and make sure the CDN surrounding persistently provide service on every node and improve load capacity.

2. THE PRINCIPLE OF GLOBAL SERVER LOAD BALANCE

The principle of Global Server Load Balance mainly contains the proximity and service Redirection Mechanisms, the success of Global Server Load Balance is directly related to success and quality of end-users' visit.

2.1 The server-based detection program

The method Based on server-side detection is to acquire server load status of each edge from the server-side actively or passively, so as to select the optimal edge server for the user request.

At first the traditional methods to achieve load balancing are described below:

2.1.1 DNS-based Redirection

Smart DNS server (DNS Scheduler) adopted, timely detects CDN nodes ES (Edge Server) which can give the fastest response to visitors, and guides the visitor's request to the speed node and provide content services by the ES. CDN users do not need to change the original site structure, but only to modify a little DNS configuration, so as to speed up the response speed of the network, that is the source station need to shield the original domain name records, or to delete (www.sitename.com IN A *.* .*.*), to add the following records www.sitename.com IN CNAME www.sitename.ppcdn.com.

When a user visits a Web site using the CDN service, DNS domain name server redirects the request of the final domain name to the DNS scheduler through CNAME way. Through a set of pre-defined strategies (such as content type, geographic area, network load condition, etc.), DNS scheduler will be able to address cache server to users which fast responds to them, so that the users can get quick service.

At the same time, it also keeps communications with all the CDN nodes distributed in different locations, and collects the health status of each node , to ensure that the user's request not to be assigned to any one that can no longer be used on the ES. Shown in Figure 1:

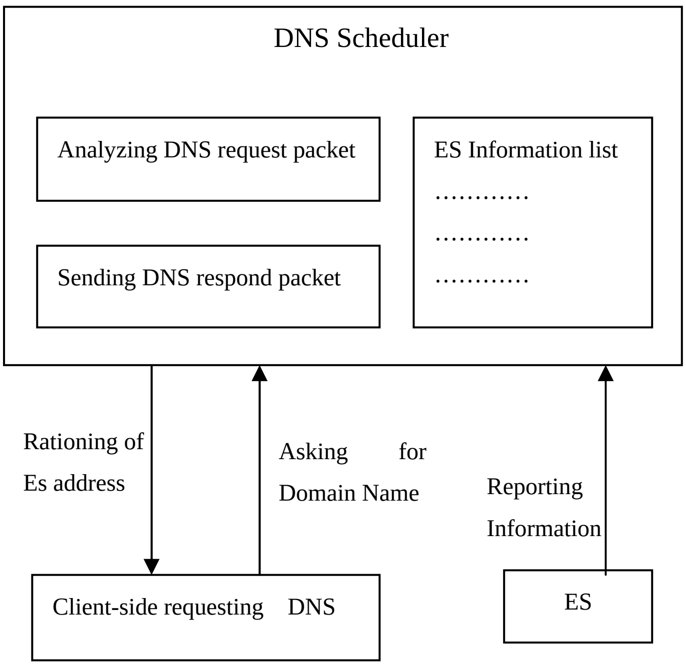


Fig.1 DNS Scheduler Principle

① When ES (Edge Server) report load information to the DNS Scheduler, Load Server will store cachesrv information into cachesrv address list, which contains the information of cachesrv load, ISP types and the type of service they can provide.
② When the client-side requesting name resolution send name resolution requests to DNS Scheduler, DNS Scheduler will analyzes the received protocol packets, and make accordingly processing.
③ According to the local DNS address of client-side which requests Domain Name, DNS scheduler acquires its own ISP type in the DNS looking-up ISP table.
④ According to the client's ISP type and type of service requested, DNS scheduler selects one or more ES addresses of the corresponding ISP type, load lighter, and providing the requested services.
⑤ DNS Scheduler will return the DNS protocol packets to the client in which the ES address has been chosen and added.

Until now, DNS load balance is still used in many web sites to ensure the site's operation and accessibility. From the point of view of its implementation and effectiveness, exists mainly following advantages and disadvantages:

The main advantages:
First, the technology is more flexible, convenient, simple, low cost, and suitable for most applications.

Second, transparent can be achieved to the user for general access, that is after adding the cache to the user client-side, no need for any setting, just direct input is enough to accelerate the domain name to access the original site.

Third, the source site server can be located anywhere on the Internet.

The main drawbacks:
First, the judgment on client location is lack of high accuracy, when the client's local DNS server is not set right, a large deviation will come about in the judgment on location of the client-side; meanwhile, the environment of china's network is very complex, network conditions are changing, sometimes this network is the fastest one, sometimes another one.

Second, it is need to update and maintenance IP database, when a DNS server is added or replaced, IP database also need to update.

Third, the network additional problems may be caused. In order to interact the DNS servers each other, to ensure the DNS data be updated in time and addresses be randomly assigned, refreshing time to DNS will be set shorter generally, however too short time will cause an additional significant network problem for increase of DNS currency.

2.2 The Client-based detection program

The server-based technology has the following shortcomings: ①The user has been assigned to an ES can not be a validly measured by other ES. ②The metric resulted from Statistical estimation procedures is calculated from the previous measurement, it can not effectively track the load on the network congestion and ES. This results it can not assign the optimal ES for users. ③Measurements make the orientation error.

But from a practical business application model, the real service organization is founded at the user-side, and the ultimate beneficiary of the request for quality improvement is the client, so such a measure departure from the server-side now inevitably has congenital defects and deficiencies.

Client-based detection [1,2] requires the collection and maintenance of Measure, and the measure reflects which edge server can provide better service for user requests for better service. User access to the global load-balancing scheduler, and the scheduler informs the client to measure the round-trip time for each ES and return them back to the global load-balancing scheduler. Then, the global load balancing scheduler maintains its database.

2.2.1 Combination of DNS and client-side-based detection technology

Only detection from the client-side will lead to more congestion throughout the network, so we uses the programs which combine DNS global load balancing with client-side detection technology.

When the clients visit CDN system at first time, CDN uses the DNS global load balancing technology to find the latest ES. At the same time it returns to the client all ES nodes within a recent broader context. The clients detect the ES node on a regular basis, write down the value of these degrees and return to the global load balancing scheduler. The global load balancing scheduler records this information in the database, when the user visits it next time, the scheduler will find the corresponding records in the database and returns to the client.

3. CONCLUSION

Above raised a load-balancing program which combines DNS global load balance with client-based detection, it can more effectively select the optimal edge server for the user. It is applied in our streaming media P2PCDN system very well.

REFERENCES

[1] Ma Yuhui: A CDN network routing optimization research program[J]. http://www.paper.edu.cn.
[2] Rami Mukhtar,Student Member,IEEE and Zvi Rosverg . A Client Side Measurement Scheme for Request Routing in Virtual Open Content Delivery Networks . 2003 IEEE
[3] BEHNEL S. SLOSL – A Modeling Language for Topologies and Routing in Overlay Networks: proceedings of the 15th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, Naples, 2007[C]. New York: IEEE, 2007.
[4] Rami Mukhtar,Student Member,IEEE and Zvi Rosverg . A Client Side Measurement Scheme for Request Routing in Virtual Open Content Delivery Networks . 2003 IEEE

RESEARCH ON IMPROVED GRID RESOURCE DISCOVERY ALGORITHM BASED ON P2P NETWORK SYSTEM

Zhaohuan Pan,Qingping Guo
Department of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, China, 430063
Email: panzhaohuan12@163.com; qpguo@mail.whut.edu.cn

ABSTRACT

At present, there are still many problems to be resolved in P2P network. For example, it is inefficient to organize and manage the grid resources, and there is no way to deal with resource updating brought by dynamic change of the grid resources, besides, there is no efficient strategy of discovering grid resources, etc. Mobile Agent, which has its own advantages such as mobility, distribution, concurrency and intelligent routing, should be effectively applied to the P2P network system. This paper firstly puts forward the network model of P2P system based on Mobile Agent, under which the resource discovery algorithm-GAMA (Genetic Algorithm Mobile Agent), which is integrated with genetic algorithms and Mobile Agent, has been put forward, then analyze the principle of integration, and show the explanations of integration process as well as more detailed description of the algorithm.

Keywords： P2P; Grid Resources; Mobile Agent; Genetic Algorithm; Resource Discovery

1. INTRODUCTION

Currently, the network computing mode has evolved to the third generation of Peer-to-Peer (P2P) computing mode, which mainly studies how to make better utilization and manage the growing global distributed information and computing resources. P2P network use distributed resource management mode, by which resources can be dynamically added and withdrew. At present, some advantages of P2P have been applied to the grid system to improve expandability and dynamics of traditional grid. Thus, P2P will play a more important role in the grid system in the future.

1.1 Problems in P2P Network

The resources in P2P network can be divided into two kinds：computing element (CE) and storage element (SE). Each CE may have the following attributes: CPU speed, the CPU number, RAM size, the average number of waiting jobs in latest seconds or minutes, and latest CE utilization, etc. And each SE may have attributes such as the total storage capacity, and the available disk space etc. Obviously, some attributes such as CPU speed, memory utilization, available disk space etc., are dynamically changing, so it is critical how to effectively organize and manage these resources.

At the same time, dynamic addition and withdrawal of P2P resource nodes, and heterogeneous hardware and software environment between nodes also challenge the load and integrating ability[1-2] of P2P network.

In addition, some factors-the load of node, resource dynamics, the time delay of resource discovery, and the trust value of node[3] and so on, should be considered in resource discovery of P2P network. Therefore, the reasonable and high-efficient resource discovery algorithm is not only crucial for successful resource share among P2P users, but also the key of development of P2P.

1.2 Background Resolvent

This paper puts forward a P2P network model based on Mobile Agent to solve the technical problems in P2P network. And a P2P network resource locating method combined with Genetic Algorithm and Mobile Agent (GAMA) is proposed, so as to construct a reasonable and stable P2P network with strong integrating ability, and research comprehensive resource discovery algorithm with high performance based on this network.

2. THE ADVANTAGES OF MOBILE AGENT

2.1 Advantages of Mobile Agent

Mobile Agent migrate among the resource nodes and communicate with them according to the route and strategy predetermined by resource information server, and it is responsible for the dynamic resource information process and collecting information assigned by resource information server. What's more, it has the following advantages [4-6]:

(1) Mobility

Mobile Agent can independently and automatically migrate in heterogeneous network and distributed computing environment, carrying information or finding proper information resources to processing information locally, accomplishing various tasks instead of user like information transfer, web query, data and knowledge discovery, and information interchange and so on.

(2) Heterogeneity and Asynchronism

Mobile Agent can support heterogeneous computer hardware and software environment, and is capable of asynchronous communication and calculation.

(3) Reduce Network Communication Costs

It is time consuming and easy to cause network congestion to transfer large amount of original information. If we put Agent in information storage place where local search and selection are done, then the selected information is sent to users by the network, the cost of connecting the remote computer network will be reduced greatly.

(4) Distribution and Concurrency

Mobile Agent offers a unique distributed computing architecture. In order to accomplish a task, users can create multiple Mobile Agents and run them in the same node or different ones. Moreover, the load of a single node can be dispersed over several nodes. Thus the mini system can cope with large scale and complicated problems.

(5) Intelligent Routing

Mobile Agent has the ability to plan the next operation dynamically according to factors like target, network communication ability and server load etc. Intelligent route can well optimize network and computational resources, balance the load, accelerate solving problem, and avoid accessing resources aimlessly.

2.2 Function of Mobile Agent in P2P network

Therefore, introducing Mobile Agent into P2P network can

solve and optimize many existing network problems, which is mainly embodied in the following aspects:
(1) In P2P network, resources inquiry can generate a huge flow of communication, but most of them are redundant, while Mobile Agent can move to each node, and reduce that waste through localization operations.
(2) Mobile Agent stores all the data it needs. Even if the machine producing is down, it still continues the search task. After the search, the result is carried to the original node, or waiting for it online again.
(3) On the nodes which accept it, Mobile Agent can get information of other nodes it has visited, so as to discover efficient resources.
(4) Mobile Agent can be allocated in different directions of the network by cloning, and carry out the parallel operation. Thereby it can find resources more quickly and improve fault-tolerance.

3. P2P NETWORK SYSTEM MODEL BASED ON MOBILE AGENT

System model is showed as Figure 3-1. The P2P model studied in this article includes supper peer(SP) layer and common peer(CP) layer. CP layer use grouping to manage each peer. Supper peer is responsible for managing common peer in regional group in CP layer. For instance, SP manage services supplied by common peer and provide inquiry for peers in the regional group. In addition to the above functions, each SP should also contain other components, which mainly are Mobile Agent Running Environment (MARE), P2P application layer, information/resource layer, among which, MARE module includes information agent, search agent and its communication mechanism, showing in Figure 3-2. This paper mainly researches Mobile Agent (MA), so other components in each SP are ignored. Suppose that all SPs support the Mobile Agent service.

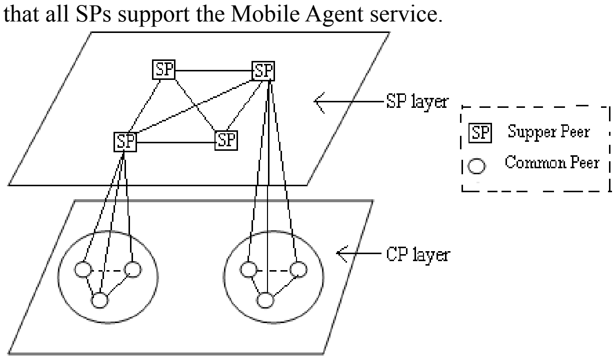


Figure 3-1. P2P System Model based on Mobile Agent

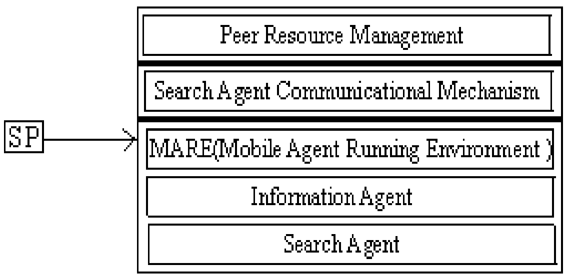


Figure 3-2. Functional Components of Supper Peer

In this system model, the process of resource location is described like this: when receives request from common peer, the responsible SP first search within its managed group, if the resource is included, it return the delivery; otherwise, SP locate the resource in SP layer using GAMA algorithm, till finding the optimal route.

4. THE BASIC PRINCIPLES OF GAMA ALGORITHM

4.1 Performance Analysis of Genetic Algorithm

J. Holland, American professor, proposed Genetic Algorithm in 1975[7]. The advantages of genetic algorithm can be sum up as this:
(1) Wide Applicability: Genetic Algorithm is a natural algorithm simulating biosphere by means of probability selection. It can handle target functions and constraints of arbitrary complexity.
(2) Global Optimization: Genetic Algorithm is global search in sense of probability because of using probabilistic search rather than path search, so it can obtain optimal solution theoretically and avoid falling into local minimum point.

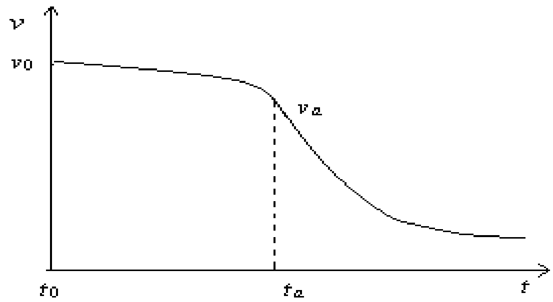


Figure 4-1. Genetic Algorithm speed-time curve

As is shown in Figure 4-1, the speed converging at the optimal solution is rapid in early stage of the search using genetic algorithm ($t_0 \sim t_a$ stage), but reduces significantly after that.

4.2 Performance Analysis of Mobile Agent in the P2P Network System Based on Mobile Agent

At present, some advantages of Mobile Agent such as distribution and parallelism, mobility, intelligent routing, have been widely applied in P2P network system. However, it is regrettable that current application of Mobile Agent is just used confined in regional network.

Moreover, in current resource location methods, mobile agents are transferred to all SPs or its adjacent SPs, then those adjacent SPs do the same thing. These methods causes waste of bandwidth in P2Pnetwork due to too much search agents migrated, and continuing transferring generates the same effect as flooding resource location in pure distributed P2P model.

Hence, while using Mobile Agent to locate resource, the key point is how many search agents should be used to migrate and which route each search agent follows. If these two critical problems are solved, not only the efficiency of resource discovery in P2P network can be improved, but also the communication load can be reduced largely, so that its scalability would be enhanced and P2P would be used more widely.

4.3 The Basic Idea Integrating Genetic Algorithm and Mobile Agent

Through the above analysis, We can come to the basic idea of the integration of Genetic algorithm and Mobile Agent: use the genetic algorithm to search resource in the early stage of resource location, that is before the optimal point t_a , and use Mobile Agent after t_a .

5. RESOURCE DISCOVERY ALGORITHM INTEGRATED GENETIC ALGRITHM AND MOBILE AGENT

5.1 The Integration Process of Genetic Algorithm and Mobile Agent

The initial information obtained from Genetic Algorithm in SP layer P2P network system decides the number of search agents and initial route. Then search agents are distributed along these initial routes.

Integration process is shown in Figure 5-1.

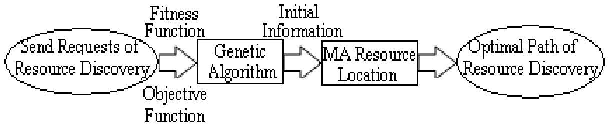


Figure 5-1. The Integration Process of Genetic Algorithms and Mobile Agent

We use undirected graph $G(V, E, \omega, \phi)$ to represent a collection of SP nodes, where V represents endpoint set, E represents edge set, ω represents weight of edge, and ϕ represents weight of endpoint. Suppose a route in $G(V, E, \omega, \phi)$ is $P_t = (v_1 v_2 \cdots v_i v_j \cdots v_k)$, $(t, i, j, k = 1 \ldots n)$, here i ≠ j ≠ k.

S_j notes the fitness value through one jump on the route, then,

$$S_j = \phi_j - k\,\omega_{ij} \qquad (1)$$

$$f_i = \sum_{1 \le i \le n} (S_i - \delta_i) \qquad (2)$$

$$\text{ResultRoute} = \max(f_i) \qquad (3)$$

Formula (1) gets total appraisal value after one jump in the special route, k represents proportionality coefficient, and the proportion of the weight of edge in fitness can be dynamically adjusted according to the network. Formula (2) indicates appraisal value of the route P_t, δ_i represents the cost of rapid adjustment of algorithm due to dynamics of node on route P_t. Formula (3) shows the final route with maximum appraisal value.

5.2 Analysis of Genetic Algorithm in Integrated Algorithm

(1) Initial Population

Random t routes P_1 to P_t is generated in the undirected graph $G(V, E, \omega, \phi)$ beginning from the query node SP1, and there is no loop in every route. The objective function is described as formula (3).

(2) Fitness Function

The greater the fitness is, the higher the individual fitness is. The fitness function is described as formula (2).

(3)Genetic Operators[8]

Selection operator: In initial population, selection probability of each individual is determined according to individual fitness and Roulette wheel selection strategy. The selection probability can be obtained from formula (4).

$$p_i = \frac{f_i}{\sum_i f_i}, \quad (i = 1,2,\ldots,n) \qquad (4)$$

Crossover operator: select two individuals (called genes) for pattern matching from each genetic population, that is to choose the same pattern, and directly add the remainder of one gene to the other's route. Then, delete the repeating part if there was. And, if there is no same pattern of two individuals, define a random point in the one with higher fitness value, to look for a better route to replace current route.

Mutation operator: The purpose of mutation is to avoid failure of finding satisfying solution due to premature convergence resulting from local optimum. In order to ensure the diversity of individuals, mutation strategy is to adopt reverse direction among random points, and delete unreachable edges.

5.3 Steps of Resource Discovery Algorithm Integrating Genetic Algorithm And Mobile Agent

(1)Initialize parameters of genetic algorithm. Suppose the number of population is n, the crossover probability is p_c, and the mutation probability is p_m.

(2)Set up terminal condition for genetic algorithm. $Gene_{max}$ and $Gene_{min}$ respectively represent the maximum and minimum numbers of iteration. Thus the terminal condition can be discribed as: in the range of $Gene_{max}$ and $Gene_{min}$, evolutionary rate is less than 5% each time in 3 consecutive iterations, then the algorithm can be terminated. The frequency of iteration and the range formulated by evolution rate is up to actual situation of the network, which can be gradually optimized in test.

(3)Set up randomly generated initial population P(t), where t represents the tth generation and is initially set to be 0. Initial population can be got by randomly generating n routes starting from initial node, noted as SP1, and there is no loop in each route.

(4) Genetic operation: when it does not meet the terminal condition, calculate each individual selection probability p_i in P(t). Then select two parents according to p_i; and hybridize them according to crossover probability p_c to get two new individuals to insert into new population P(t+1); then, insert descendants of the two hybridized parents into the new population P(t+1); t = t+1.

(5) Calculate fitness value of each individuals in P(t).

(6) Select individuals with stronger adaptability from P(t) according to fitness value, 20% for instance.

(7) The number of search agent is set as the number of selected individuals above. Then, distribute Search Agents according to corresponding route of those individuals.

(8) Each Search agent migrates to the designated SP, to interact with local Information Agent, querying the list of resources on local SP node.

(9) Each Search Agent return with query result after executing the designated route.

(10) The initial query node SP1 submits the query result to the peer node requesting resources, thus the query task is completed.

5.4 Description of GAMA Algorithm

P(t) notes the tth generation of population;
p is to hold individuals in P(t);
p1, p2, p3 are to hold individuals in P(t+1);
p_i notes the selection probability;

```
QAMA_QueryRoute
{
    t = 0; //t notes the 1st generation as 0
    define p, p1, p2, p3;
    for populationNum = 1 to n;
        p[i] = Initialization();
    //n is the number of initial population, array p notes individuals of //initial population
```

```
initialize Gene_max ; // the maximum number of iterations
initialize Gene_min ; // the minimum number of iterations
set p_c ; // probability of crossover
set p_m ; // probability of mutation
geneCount = 0; // the number of continuous iteration, and the //evolution rate is less than 5% after each iteration
do
{
    for i = 1 to m; // m notes the number of individuals in P(t)
    // calculate selection probability according to individual fitness // value and Roulette wheel selection strategy
        caculate p_i ;
    //select parent up to selection probability
    select initial population from P(t);
    p1 = CrossOver(p);
    p2= Mutation(p1);
    if(evolution rate is less than 5%)
        geneCount++;
    else
        geneCount = 0;
    t = t+1; //transmit to (t+1)th generation
} while(geneCount <= 3)
selete p3 from p2; //p3 notes individuals with strong adaptability
agentNum = the number of selection;
SearchAgentRoute[i] = p3[i];    // i = 1 to agentNum
for i = 1 to agentNum
    MA_QueryRoute(SearchAgentRoute[i]);
//discover resources with resource locating method on Search Agent
return the optimal route;
}
```

6. CONCLUSIONS

This paper proposes the P2P network System Model based on Mobile Agent to solve some problems in current P2P network, making full use of the advantages of Mobile Agent. Meanwhile, based on this model, the paper also shows a resource discovery algorithm combined Genetic algorithm and Mobile Agent. That algorithm develops the strengths of Genetic algorithm and Mobile Agent, taking into account of the pros and cons of the implementation of each other's advantages, to be more usefully applied in practical situation. The future work is to simulating analyze ideas proposed in this paper, collect data, and optimize parameters of the algorithm, to enhance the reliability and utility of the performance of the algorithm.

REFERENCES

[1] Androutsellis-theotokis S,Spinellis D. A survey of Peer-to-Peer content distribution technoligies[J]. ACM Computing Survers,2004,36(4):300-410.

[2] Menasce D,Kanchanapalli L. Probalistic scalable P2P resource location services[J]. ACM SIGMETRICSL Performance Evaluation Rev,2002,30(2):45-60.

[3] HAMADA T,CHUJOK,CHUJO T,et al. Peer-to-Peer traffic metro network: analysis, modeling,and policies[A]. IEEE International Conference on NOMS[C]. Seoul,2004. 425-438.

[4] Harrison C et al. Mobile Agent:are they a good idea?, Technical Report,RC 19887,IBM,Mar.1995.

[5] Riecken D. Intelligent agent. Communications of the ACK,1994;37(7):18-21.

[6] Cao JunWei,Spooner D P,Jarvis A,et al. Grid Load Balancing Using Intelligent Agent[[J]. Futher Generation Computer Systems,2005,21(1):135-149.

[7] Mitechell M. An introduction to genetic algorithms[M]. MIT Press,Cambrige,MA,1996.

[8] Neubauer A. The Circular Schema Theorem for Genetic Algorithms in Engineering Systems: Innovations and Applications[C]. London:IEE Press,1997.209-214.

Zhaohuan Pan is a master candidate of School of Computer Science and Technology, Wuhan University of Technology, focusing on parallel computation and multimedia.

RESEARCH ON THE METHODS TO IMPROVE THE PERFORMANCE OF VOICEXML-BASED VOICE SYSTEM

Liu Wenyun, Hu qun
Institute of Scientific and Technological Information, Shandong University ofTechnology
Zibo, Shandong, China
Email: qun1002@126.com

ABSTRACT

This paper analyzes the major factors which impact the performance of VoiceXML-based voice system, and then researches on three methods to improve the performance of the system. First, we discusses the importance of cache and its application in VoiceXML-based voice system and give experimental data to show its importance. Second, we find it's very important to optimize the acquisition of VXML documents to enhance system performance. At last, we give an improved method of DOM cache.

Keywords: VoiceXML, Cache, Document Acquisition, DOM

1. INTRODUCTION

VoiceXML(Voice Xtensible Markup Language) is a markup language based on XML, which is used to develop voice applications. VoiceXML is not only to achieve the functions of traditional CTI, but also to identify the user's voice input, and also combine the public telephone network with the Internet, making the two networks can communicate with each other. As a system can generate and realize new business with the help of the current communication network fast, easily, flexibly, economically and effectively, voice service platform based on VoiceXML is the best choice for meeting the growing needs of customers and bringing huge profits to operators. While making VoiceXML speech platform integrate with the WWW closely, VoiceXML also has some drawbacks, such as speech delay and so on. In this paper, we maily offered a method that can improve the system performance based on VoiceXML, which can improve the availability of the system by reducing voice latency.

2. THE MAIN FACTORS IMPACTING THE PERFORMANCE OF VOICE SYSTEM BASED ON VOICEXML

Both the VoiceXML implement platform and file server must send request and get response through network, and this has given rise to network delay. Figure 1 is a voice platform architecture based on VoiceXML. For example, when a user want to access to the voice platform by telephone, the platform will start the VoiceXML interpreter (parser), and it will visit the Web server for VoiceXML document, then after the interpreter's explaining, the platform will enforcement procedures and download voice resources to communicate with the user by voice, and again it will access the server according to the communication results for obtaining next document. Clearly, waiting time of the user from he calling to hearing the first response is the interval time of the voice platform to start and obtain documents, in other words, it is the whole time of Web servers obtaining VoiceXML document and voice resources and explaining the implementation of this document. Among all these time, the time of Web servers obtaining VoiceXML document and voice resources is the longest but the least controllable.

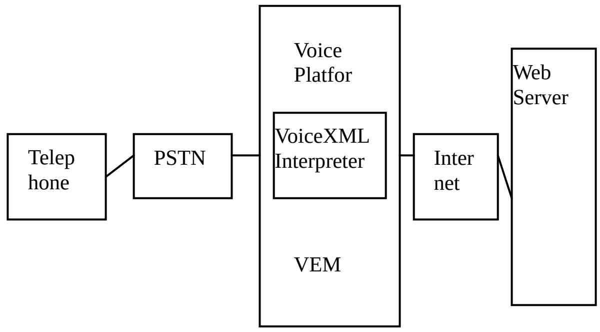


Figure 1. Voice Platform System Architecture Based on VoiceXML

In general,high quality voice requires the delay maintain within 200 milliseconds[1]. If accessed by voice, the user's patience is limited and the response time that he expected is far less than the expected time that accesses through the web browser. Consequently, the customers are more likely lost as the result of the unsatisfaction with the Qos(quality of service). This is the real reason that almost all telecommunication equipment have strict requirement with the response time[2].

To address this issue, VoiceXML Forum recommends:
⑴A VoiceXML document should contain as many as possible of dialogues, and by reducing the times of interaction with the server and the network delay to improve the network response speed.
⑵Setting cache server between VoiceXML interpreter and the server, it can reduce network load and the server pressure, last reduce the response time.
The following sections will discuss several methods for improving the performance of the voice system based on VoiceXML.

3. USING THE CACHE TO REDUCE NETWORK DELAY

3.1 The Use of Cache

Cache technology is considered as one of the effective ways to reduce server load, the network congestion and network delay[3][4].
It can bring many benefits such as follows:
⑴ It can lighten network congestion by reducing network traffic.
⑵ Reduce users visit delay, and the main reasons are: ① Customers can obtain the knowledage directly from the agent rather than from the remote server, so that it can reduce transmission delay②Contents that not stored in cache can be easily obtained by users as the result of the reduction of the network congestion and server load.

⑶ As a part requests of users can be obtained from the agent, it obviouly reduces remote server load.
⑷ If the remote server cannot respond customer request due to a remote server failure or network fault, customers can obtain cache copies from the agent, making the quality of service more strengthened.

3.2 The Application of Cache in VoiceXML Speech Platform

Figure 2 describe the application model of cache in VoiceXML speech platform[5]. It can store the resources that VoiceXML requires on a agent server or a closer server, so that the system will not need to obtain resources from the original server and users' requests can be responsed from the near server, consequently,it reduces the network transmission time and address analytical time.

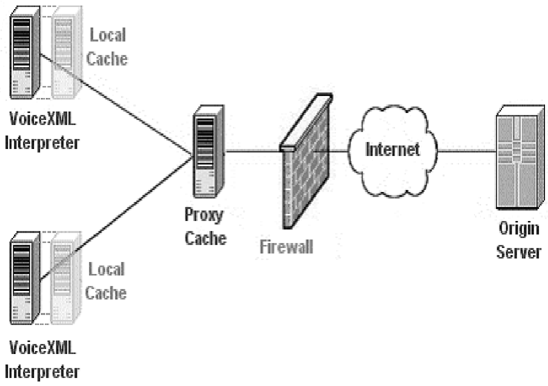


Figure 2. Cache Application model in VoiceXML Speech Platform

Voice content business belongs to telecommunication application and needs a high level availability. So VoiceXML system uses multi-cluster method (two VoiceXML execution platforms shows in the above picture) to provide uninterrupted sevice of 24 hours a day and seven days a week. In this kind of structure, business enforced in two platforms has their own Cache system.At the same time, cache frontend machine is introducted to local area network of the system to constitute a double-deck cache system.

The fronted cache system preserves all the available resources of the executed platform that are cached down. In order to ensure the effective resources of cache system was not replaced due to capacity problem, it's best to make the capacity of the fronted cache system totled all of the capacity of the platform. Surely if the capacity of the fronted cache system is not very big, it still can play its role in operation.

The hierarchical structure can shorten request response time through the graded cache mechanism closing the distance between users and resources. Buffer cache can be classified into more levels, also it can be located any corners of the network. In addition, each level of buffer (agency) server still can use different working ways.

3.3 Case Study

Using Nuance Café as the speech server and a personal computer in the library of Shandong University of Techology as a remote server, we tested a 100Kb speech file. Test data are shown in table 1.

From the above data, we can see that without using the cache, system's average speech delay is 5529.6 millisecond, on this condition users may probably lose patience and hang up. After using the cache, system's average speech delay is only 811.6 millisecond. Performance of the system improved significantly. Therefore we can prove that cache plays an important influence in improving the performance of the voice platform.

Table 1. The Experimental Data of the Influence about Speech Delay Using Cache

	speech delay of not cache files (ms)	speech delay of cache files (ms)
1	5687	812
2	5879	782
3	5662	812
4	5365	828
5	5323	797
6	5642	813
7	5201	797
8	5332	854
9	5421	822
10	5784	799
Average	5529.6	811.6

4. OPTIMIZE THE ACCESS TO VOICEXML DOCUMENTS

4.1 One Understanding of Optimizing the Access to VoiceXML Documents in Improving the Performance of Voice Platform

Some researchers have thought that the time of Web servers obtaining VoiceXML document and voice resources is the longest but the least controllable[6]. So if we can move the VoiceXML document that once stored in the web server to the local disks, the system performance may be improved. And the article[7] did some researches about it.

Article[7] did some tests using the access time of voiceXML document as example. According to the definition, VoiceXML document obtaining time is the total time from sending out a document request to obtaining the required document. From the qualitative perspective, the time to obtain a VoiceXML document is more longer through computer networks than through a local disk.Besides,the obtaing time is different from remote server and local disks of a same voice platform and the VoiceXML document obtaining time just reflects this diffierence rightly. If the VoiceXML documents that are stored on the local disks have more shorter time and can make the voice platform's response speed more faster, it can be proved the optimization to the performance indicators is helpful to improve the overall performance.

Using a HPDL380 servers as the speech platform, a personal computer in the Internet as a remote server and the size of all the VoiceXML documents are 100KB.

We did ten times single call serial simulation test about the two differnt kinds documents position respectively. The test results are shown in table 2 and table 3.

Table 2. VoiceXML Document Obtained from the Remote Server Document

	VoiceXMLdocument access time (ms)	Speech platform response delay (ms)
1	84.763	1269.03
2	82.136	1264.415
3	83.869	1264.069
4	89.039	1263.033
5	76.625	1261.656
6	80.855	1262.375

7	85.321	1259.053
8	88.584	1259.265
9	87.216	1259.233
10	82.518	1260.361
average	84.093	1262.249

Table 3. VoiceXML Document Obtained from Local Speech Platform Server

	VoiceXMLdocument access time (ms)	Speech platform response delay (ms)
1	55.865	1189.952
2	62.186	1220.057
3	53.826	1220.069
4	49.099	1210.033
5	56.716	1210.086
6	50.056	1210.053
7	48.928	1210.053
8	51.574	1210.065
9	51684	1220.062
10	58.531	1220.068
average	53.847	1212.050

The conclusion of article[7]: optimizating the VoiceXML document obtaining time could improve the performance of the speech platform , and the index of VoiceXML document obtaining time is effective. From another point, we can also find that optimizating the VoiceXML document obtaining time does not significantly improve the executive efficiency of speech platform, which proves the bottleneck of the speech platform is not in document obtaining.And it need optimizating other performance indexes to improve the fuctions of the whole speech platform better.

4.2 Views of Our Reseach

We think that the point offerd in article [7] is one-sided and our study did relevent tests.In the tests, we use Nuance Caféa as the speech platform server and a personal computer in the library of Shandong University of Techology as a remote server, the size of the VoiceXML documents are 100KB. The experimental results are presented in table 4 and table 5.

Table 4. VoiceXML Document Obtained from the Remote Server

	VoiceXMLdocument access time (ms)	Speech platform response delay (ms)
1	4922	5766
2	4516	5672
3	4468	5296
4	4469	5344
5	4500	5359
6	4594	5500
7	4500	5359
8	4622	5328
9	4486	5268
10	4576	5366
average	4565.3	5425.8

Table 5. VoiceXML Document Obtained from Local Speech Platform Server

	VoiceXMLdocument access time (ms)	Speech platform response delay (ms)
1	62	984
2	62	907
3	62	875
4	62	969
5	94	921
6	62	906
7	78	891
8	62	1093
9	62	875
10	62	922
average	66.8	934.3

From the above test data,we can see thatthe performance of the system has a significant improvement after optimizating, at the same time, it also proves that the viewpoint stated of article [7] is partial. And this results are due to their selection of speech server and the remote Web server without universal meaning. Probably its PC that was as the "remote" Web server and the speech server are very close, even perhaps they are in the same laboratory. In such an environment, the difference between a VoiceXML document cache speech server and a Web server is not obvious at all.

5. REDUCE THE ANALYSIS TIME OF VOICEXML SPEECH PLATFORM

5.1 The Executive Process of VoiceXML Document in the Local Speech Platform

After VoiceXML document was added to the speech platform, The executive process in the local speech platform is as follows: fristly, we can transforms the VoiceXML document to a tree in EMS memory with the help of the VoiceXML explanation module, and the tree was named as a Document Object Model Tree (DOM，Document Object Model). Then, VoiceXML implementation module uses form interpretation algorithm to explain the executable files and control speech resources and call control resources in order to achieve the business functions [8].

5.2 A Method to Accelerate the Analysis Speed of VoiceXML Document

In fact, DOM is an abstract specification that adopts the approach of tree object collection to access to given document contents [9]. Tree structure model of DOM just reflects the level relationship between the various elements of XML documents, which can also facilitate the various elements of the content to add, delete and modify.However, it needs to read the entire document into memory when DOM analized a XML document, and the DOM tree structure also used more memory, so the performance of DOM decline rapidly when dealing with large-scale document, and also requiring a longer period of time [10]. In addition as we must do an analysis when a VoiceXML document was loaded from a remote Web server to the VoiceXML speech platform, and then it is generated to a tree object model, which also led to the delay. Article [11] pointed out that we can store the document object model tree for the next analysis of the same document.And by this way we can avoid the system to repeat the same analysis when dealt with one document, and finally we can reduce the system delay.

5.3 Improving the Methods of Caching DOM Tree

A problem of Caching DOM tree is that if all DOM tree of VoiceXML documents are cached , then they will take up more system space. So, can we do some improvements? Fristly, we should analyze the structure of DOM tree.

Such as following XMLdocument:

```
<?xml version=”1.0” encoding=”gb_2312”>
 <addressbook>
  <person sex=”male”>
   <name>Liu Xianfeng</name>
    <mentor>tutor</mentor>
  </person>
  <person sex=”male”>
   <name>Cao Buwen</name>
```

        <graduate>graduate student</graduate>
    </person>
</addressbook>

When this XML document is loaded to a memory, it can shape to a DOM tree, which is shown in Figure 3 [12].

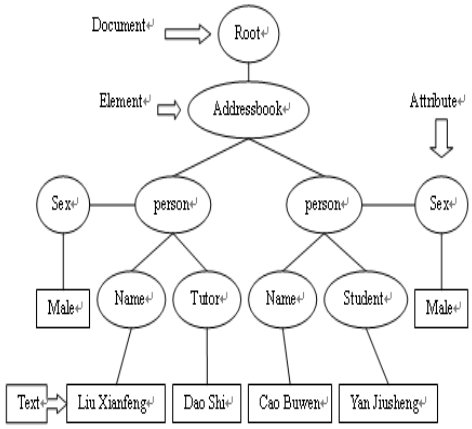


Figure 3. The DOM Tree Generated by XML Document in Memory

We can see from the picture, except the differences in test between the XML documents whose elements and hierarchical structure are same, the rest of theirDOM tree are still in the same condition. If the two modules of the VoiceXML document structure are the same, the DOM trees they are transformed are still same only except for the text. Therefore, this paper presents a supplementary to article[11], that is by sharing DOM tree cache. The main ideas of sharing DOM tree cache are as below:

①DOM tree cache module will be divided into two main modules: a general DOM tree cache module and a shared DOM tree cache module. And the shared DOM tree module is divided into a DOM tree structure module and a DOM tree content module.

②After accessing to all the VoiceXML documents the first time, the system transforms documents to the DOM tree and adds it to the general DOM tree cache module.

③After transforming all the documents, the parser will extract the contents in the general DOM tree cache for analyzing to find the document with the same structure. Then it will analyze them to the DOM tree structure and content and put them into the DOM tree structure modules and the DOM tree content modules respectively.

④When accessing to the document next time,the system can select the corresponding DOM tree to run it according to the different URI of different documents. This method can avoid analyzing VoiceXML documents repeatedly, and reduce system latency, at the same time, it can lower the buffer space occupied by the DOM tree to avoid the system performance declining. But it also has disadvantages. The limitation of the method is that it usually plays a vital role in a system that has many VoiceXML documents with the same structure. And if a system has very few documents with the same structure and the method will not apply.

6. CONCLUSIONS

The paper verified the effects of using cache and optimizing document obtaining to improve the performance of VoiceXML system by doing related experiments. In addition, we proposed a opposite point about the relationship between optimizing document obtaining and improving the VoiceXML system performance. Finally, we offered a method to improve caching DOM tree, but duing to limitations of capabilities and object conditions, we failed to obtain the relevant experimental data.And this will be our next striving direction.

REFERENCES

[1] Chen Xiangguo, Jia Dongli，Han Li. real-time voice communication based on DirectSound and Winsock[J]. Fu Jian Computer, 2006(8):102-103.

[2] Wang Wenlin, Liao Jianxin, Zhu Xiaomin. VoiceXML speech platform caching Technology review[J]. Journal of communication, 2007(2):101-108.

[3] CACERES R,DOUGLIS FELDMANN A,et a1. Web proxy caching:the devil is in the details[J]. ACM Performanec Evaluation Review,1998,26(3):11-15.

[4] DUSKA B M,MARWOOD D,FEELAY M J. The measured access characterisitcs of world wide Web client proxy caches[C]. Proceedings of the First USENIX Symposium on Internet Technologies and Systems. Monterey,l997.23-35.

[5] Dave Burke.Enhancing VoiceXML Application Performance By Caching[EB/OL]. http://www.voicexmlreview.org/Dec2002/features/vxml_app_performance2.html, retrieval date:2007-12-26.

[6] Wang Wenlin, Liao Jianxin, Zhu Xiaomin. Speech platform cache consistency control algorithm based on VoiceXML[J]. Journal of electronics, 2007(4):640-646.

[7] Pang Tingting, Liao Jianxin, Zhu Xiaomin. VoiceXML Speech platform performance index research [J]. Computer system application, 2007(8):19-22.

[8] Guo Zhiwei, Wu Jiachun, Liao Jianxin. Research of the mechanism of buffer applied in VoiceXML system [J].journal of information network, 2005(11):61-63.

[9] Lv Wenfeng. VoiceXML business test design and implementation of the system of [D]:Beijing: Beijing university of posts and telecommunications, 2006.

[10] Liu Yong. The research and design of XML document analysis model[J]. Computer simulation, 2005(2): 253-254.

[11] Dirk Schnelle,Fernando Lyardet. Voice User Interface Design Patterns[EB/OL]. http://elara.tk.informatik.tu-darmstadt.de/publications/2006/EuroPLoP2006.pdf,retrieval date:2009-3-15.

[12] XML ’s DOM research and application [J]. Computer era,2008(2):1-3.

Liu Wenyun was born in 1964, Shandong province. He is a full professor and a superisor for Master Candidates in Science and Technology Information Research Institute of Shandong University of Technology. In addition, He has edited 2 textbooks and has published more than 80 essays as well. His research interests are in information resources management and information analysis research.

Hu Qun was born in 1983, in Shandong province. She is a student for master degree of Information Science in Science and Technology Information Research Institute of Shandong University of Technology. She has published two papers on core journals and participated in many research projects during her studies. Her research interests are in information resources management and information analysis research.

A SUPPORTING FRAMEWORK FOR WEB SERVICE DISCOVERY BASED ON OWL-S*

Lou Yuan-sheng[1], Yang ning[1], Xu Hong-tao[2]
[1]College of Computer & Information Engineering, Hohai University, Nanjing 210098
[2]Data management Center of Zheng zhou Labour and Social Security,Zheng zhou 450007
[1]Email:Wise.lou@163.com

ABSTRACT

It has become to a research hotspot to incarnate the advantages of Web services on how to achieve high efficiency and automation of Web service discovery in a specifically environment. Service discovery methods in specific environment are always considered to increase semantic information in the service description, and given their semantic matching algorithm. In this paper, a Semantic Web Service Discovery framework which combine OWL-S and UDDI is delivered.The framework is composed by the function modules of service publish, services search, UDDI adapter, ontological parser and the field of ontology library as well as the UDDI registry center. Finally, the paper give out the main modules of the framework and its work mechanism

Keywords: Semantics Web Service, OWL-S, Service Discovery, Service Matching

1. INTRODUCTION

The Application of the semantic to service discovery will help to improve the precision ratio and recall ratio,and some works has been done in this related academic research. Mark Klein from MIT and Abraham Bernstein from New York University put forward a service discovery methods based on the process of ontology [1].METERO-S [2] [3] is a research projects about Semantic Web/workflow and Web services technology which is led by the LSDIS laboratory of the University of Georgia. METERO-S describe the Web service functions and behavior by the use of OWL-S [4] with no ambiguity and computer understandable method,and by the application of P2P technology connect multi register to a distributed network to share the service description information. The project OWL-S/UDDI Matchmaker[5] by Carnegie Mellon University (CMU) expand a OWL-S/UDDI matchmaker in the basis of WSDL and UDDI, it add semantic information in UDDI by the way of the combination of OWL-S's and UDDI's tModel.

Service discovery methods in specific environment are always considered to increase semantic information in the service description, and given their semantic matching algorithm.Therefore, in specific environment the use of semantic web and web services technology to realize semantic matching based on web services is an important research fields of web service discovery, but also an important way to solve the problem of service discovery performance.

OWL-S (Ontology Web Language for Services) is a specification language based on the ontology for the semantic Web service described, it is a markup language used to describe the properties and functions of Web Service with the character of non-explicit semantic,ambiguity,machine understandable[6].By the use of ontology and Semantic inference to the study of Web services discover and match, this paper gives out a web service discovery and matching framework based on OWL-S.

2. WEB SERVICE DISCOVERY FRAMEWORK

2.1 The Architecture of the Framework

In this paper, we designed a Semantic Web services (SWS) discover and match model framework with the combination of OWL-S and UDDI, the framework built on the UDDI registry by the use of public or private UDDI registry as physical storage for the information of preservation services and service providers.The core processing module of the framework is OWL-S matching engine which used to analyze the degree of match between service requester and publisher.The framework connect multi UDDI registry center together in the network, and classify the Registry by the use of Semantic with the Combination of OWL-S and UDDI. Architecture of the framework as Figure 1.

This framework is composed of the service publish, service search, UDDI adapter, Ontology Parser , Ontology Library, and the UDDI registry center. The service query module of the framework can realize mapping between OWL-S and WSDL.After mapping, the corresponding information which used to support query request based on OWL-S or WSDL description will be saved into the UDDI information library and service description information.

2.2 The Composition of the Framework

2.2.1 UDDI Adapter

The module is mainly responsible for the interaction between the outside and the UDDI registration center, the compositation of this module as follow:

(1) In accordance with mapping mechanism of OWL-S to UDDI ,it release to the UDDI registry owlSpec types of Model, and create businessService information structure in UDDI, and then stored this information in the UDDI service registry by calling the save_service() interface of UDDI.

(2) It find all types Model belong to owlSpec from the UDDI registry by calling find_Model () interface of the UDDI,and find the corresponding OWL-S service description, achieving service information for OWL-S matcher to use through the OWL-S parser.

(3) It get the corresponding services list by the way of finding the businessService in UDDI which using these model according to the Model List of matching results based on semantic similarity.

* Supported By the National High-Tech Research and Development (863) Plan of China under Grant No.2007AA01Z178.

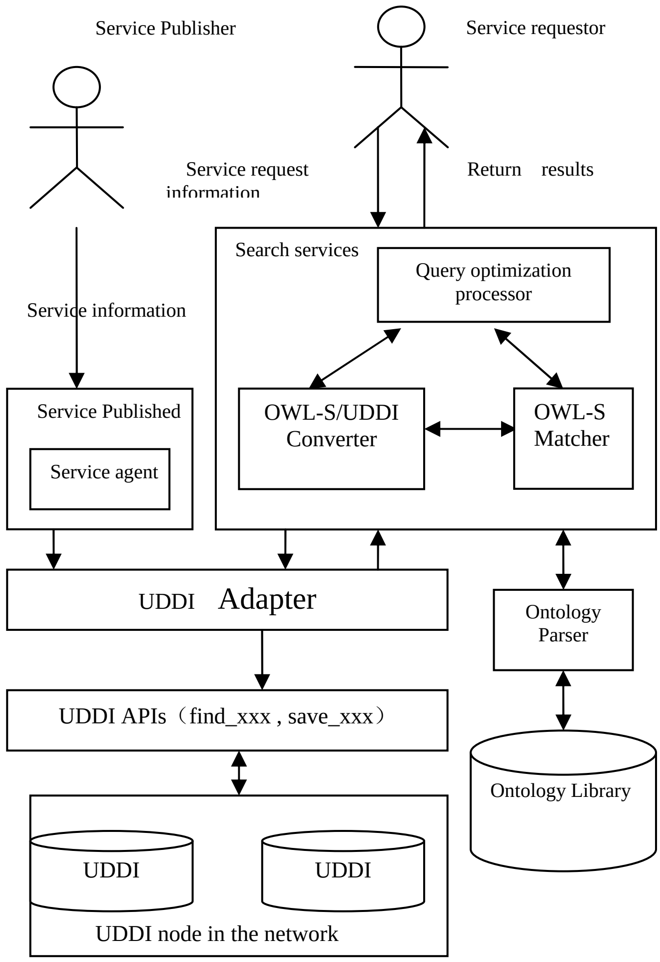


Figure 1. The Framework of the Service Discovery and Match

2.2.2 OWL-S/UDDI Converter

The functions of the OWL-S/UDDI converter designed in this paper as follow:

(1) Establish the conversion method of OWL-S/UDDI, corresponding the function description information of the service by adding a new tModel type, preserving semantic information of the services by reference tModel, storing those information in UDDI by call save_business () and save_service () interface in the API of UDDI.

(2) Obtain particular user's business entities and service information from the UDDI registry by the call of get_registeredInfo() interface.

(3) Search all business services which reference to tModel expand type from the UDDI registry and access to its URL of service ontology described by the call of the UDDI's find_service () interface.

2.2.3 OWL-S Matching Engine

The target of Web service Match is a Web service and a user's request must meet the following two conditions: the output of the services meet to the output of user's request and the input of user's request meet to the input of the service at the same time.
Semantic matching is to determine the degree of similarity by analyzing the two groups semantic description of the service and requester in order to achieve the purpose of Web services matching. The matching engine of the framework as Figure 2.

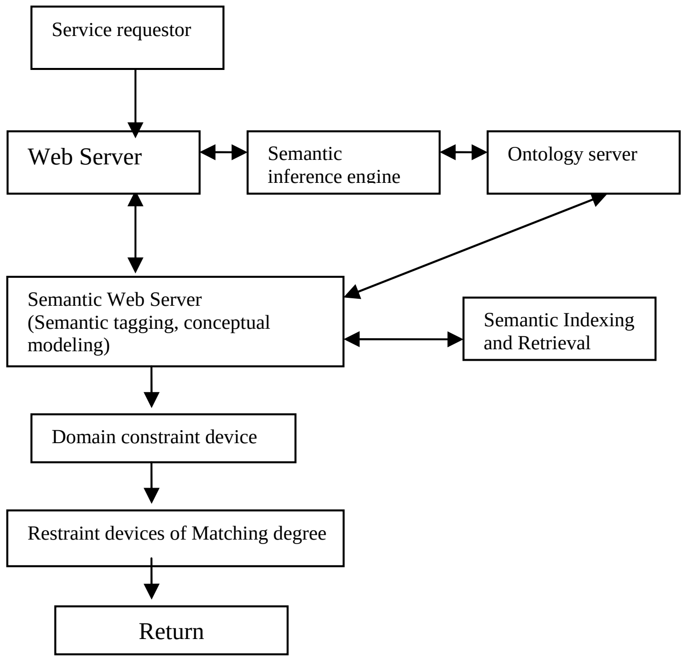


Figure 2. The Structure of Matching Engine

The working principle of the matching engine is as follows: When the matching engine receives a service request, Web server accept the request, and calls the Semantic Web server, Semantic Web server obtain the semantic sign which according to the content from the ontology server based on the understanding to the requests, then the semantics will be assigned to the request in order to complete the semantic tagging to the service request. After the end of the semantic tagging, combined with the domain of rule base and inference engine for logic reasoning semantic-based queries is realized.

When the service requester searching a service, it always hoping to find a service fully meet the requirements which is called exact match, but it is difficult to find an exact match in general, more is more search engine users search for the needs of a large number of similar services. In order to make the engine more flexible implement the matching, this matching engine adopt the multi-layer filter mechanism of Domain constraints and matching degree constraints to improve the matching efficiency.

2.2.4 Ontology Parser

This module provide the function of parse, search, reasoning interface to the domain ontology in the domain ontology library. Parser is implemented by Jena provided API, through which parse the domain ontology defined by OWL and process reasoning. The parser has the following functions:

(1) It determine if the type A and type B are from the same ontology through determine if there name space accordant.

(2) Get the attribute set of the class.

(3)To determine the relationship of the classes. The relationship of the classes include: equivalence relations, sub-class relations, and inclusive relationship.

3. WORKING PRINCIPLE OF THE FRAMEWORK

3.1 Service Published

Service providers publish the Web services through the service proxy who achieve to the services published by the interaction between UDDI adapter and service providers. OWL-S service generator is one of the main components, which generate service's OWL-S service description according with the OWL-S service specification and the information of the service provided by the provider such as service name, service description, service input / output, as well as the results of pre-conditions and other information.

3.2 Services Registration

Service providers publish services through a proxy, proxy identify register nodes by searching the XML document which store the semantic classified results. UDDI adapter enter the service query module, OWL-S/UDDI converter convert the service described by OWL-S to UDDI format through the combination of OWL-S and UDDI's tModel, and then deliver the service to the UDDI registry. The ontology in the framework provide describing information for matching model used to determine whether the services provide the user needed function.

3.3 Services Find

This module is mainly responsible for the interaction between UDDI adapter and service requestor, and get the realization of services find. The process of the service find as follows:

(1) Service providers publish services through the Service proxy, the proxy find the service register center by searching the XML document which store the semantic classified results. Service requestor sent out the request of service discovery to query optimization processor which will get the services list of "the candidate service".

(2) In order to get higher availability service to meet the personalized needs from the "candidate service" list, query optimization processor sue dynamic service requests to these "candidate service" through UDDI adapter, as well as to get the dynamic information correlative to these "candidate service".

(3) Each service providers of the "candidate services" return the dynamic information of these services, and optimize processing module will confirm "candidate services" service level (the weight) by the static information, dynamic information, and others of the "candidate service". The optimization processing module sort these "candidate service" in order of the weight, some lower level candidate service maybe throw away. "The location" list of the remaining one or more services to meet the user’s need will send to OWL-S matcher, OWL-S matcher find the most appropriate services and returned to the service requester through the ontology parser match ontology library.

3.4 Matching Strategy

This paper put forward a step-by-step service match strategy which achieve the whole service match process through the different step-by-step matching algorithm. The step-by-step match is divided into two phases: the first phase search all candidate services advertising set, calculate the service similarity for each "candidate service" between the services requester. If the service similarity greater than the user-specified liminal, this service will be added to the list of results. The second phase optimize the "candidate service" in the results list, add the weight to the calculation of semantic similarity according to optimize the results, and sort to the services in the results list according the value of the weight, lower value weight of the service will be abandoned. At last, if the user give out the specified similarity liminal it will return the corresponding matching results.

In this paper, the Web services classification information are described by the use of ServiceClassification in OWL-S specification, tagging service information using ontology concept of OWL (domain ontology).By calculating the similarity value between the requested service and published service, it returns the services list which the similarity is greater than initialization as a candidate for the next matching. The step-by-step matching algorithm based on Service optimized represent that the time complexity reduced significantly, time-saving services matching and improve the efficiency of matching engine.

4. CONCLUSION

As a developing technology,Semantic Web services has many problem worth to study.The Web Service discovery of Semantic Web based on OWL-S has become a very active study area in recent years. This paper put forward a semantic Web service discovery and matching framework by the combination of OWL-S and UDDI , and designed a service optimization Web services matching method under consideration of how to improve the precision rate and recall rate.

REFERENCES

[1] Banaei-Kashani F,chen C,shahabi C. WSPDS:Web services peer-to-peer discovery service. In:intl. Symposium on Web Service and Applications. Las Vegas, USA, 2004

[2] Kunal Verma,Karthik Gomadam,Amit P. Sheth, John A. Miller,Zixin Wu. The METEOR-S Approach for Configuring and Executing Dynamic Web Processes. http://lsdis.cs.uga.edu/projects/meteor-s/techRep6-24-05.pdf.

[3] Ma Bing xian，Wu Zhe hui, Xie Neng fu，Modeling Agent-based Semantic Web Service with Petri Nets. Journal of System Simulation[J]. 2005, 1(17):120~125.

[4] Verma K,Sivashamugam K,Sheth A,et al. METEOR-S WSDL:A Scalable Infrastructure of Registries for Semantic Publication and Discovery of Web Services. Journal of Information Technology and Management,2005,6(1):17~39.

[5] Massimo P, Takahiro K,Payne T R,et al. importing the semantic Web in UDDI[A].In:Proceedings of Web Services,E-business and Semantic Web Workshop(CAiSE-Workshop)[C].Toronto,Canada:[s.n.],2002:225~236.

[6] W3C:OWL Web Ontology Language for Services, Http://www.w3c.org/submission/2004/07/.

WEB-BASED E-BUSINESS APPLICATIONS

Zhang Lei,Wang Wei
Control engineering department
Jiangxi Blue-Sky University, Nanchang, China, 330000
Email: Wangwei__2000_0@163.com

ABSTRACT

The World Wide Web has contributed significantly to the creation of a global marketplace. The Web offers business users a wide range of opportunities for advertising and marketing their goods and services as well as for conducting e-commerce over the Internet. Some of the more commonly used capabilities of the Web include the use of Web sites, business-to-customer e-business applications, and business-to-business e-business applications. In addition, as advances in information technology continue, businesses find new and innovative ways to use the Web to reach customers and partners alike.

Keywords: E-business Strategy On-line

1. OVERVIEW

The Internet in general and e-business in particular have provided business opportunities for both small and large firms. Small businesses have been utilizing the Internet more than their larger counterparts. This is believed to be due to fewer obstacles associated with systems integration and more flexibility to implement change. However, large companies are finding e-business to be a competitive advantage and a powerful management tool for their complex supply chains. E-business implementation allows businesses to share information with suppliers, buyers and partners, and to better plan and manage supply and demand. Companies can use e-business to conduct business at anytime, day or night and to reach potential Web connected buyers regardless of their location. In addition, companies can provide better customer service and reduce the costs of production and distribution of products and services.

Computer networks are like the nervous system of a business. Networks can mobilize a matrix of business activities creating high levels of efficiency by facilitating rapid communications and integrating disparate functions into an end-to-end system of production or performance. Computer networks can support internal administrative functions, preproduction and post production activities, and relationship management activities.

The World Wide Web offers business users a wide range of opportunities for advertising and marketing their goods and services as well as for conducting e-commerce over the Internet. Some of the more commonly used capabilities of the Web include the use of Web sites, business-to-customer e-business applications, and business-to-business e-business applications. In addition, as advances in information technology continue, businesses find new and innovative ways to use the Web to reach customers and partners alike.

2. ON-LINE RETAILING AND ELECTRONIC STOREFRONTS

Most people today are familiar with the ability to purchase products over the Internet even if they have not done so themselves. One of the most common applications of e-commerce is on-line retailing and electronic storefronts. In this approach to e-commerce, customers are able to visit a business's web site on the Internet and examine product pictures and information, compare different products, fill an electronic shopping cart, and checkout and pay for their purchases. In this scenario, the organization's home page is the electronic equivalent of a brick and mortar storefront, and the various web pages being the electronic equivalent of the aisles of a traditional store. Customers visit the company's web site and look for the items that they need in much the same way that they would visit a brick and mortar store. Typically, e-commerce web sites allow customers to "stroll the aisles" by having links to various categories of products or to search for specific items by various characteristics such as key words, title, product name, item number, or model number so that they can go directly to a specific product using a search engine. For example, TaoBao allows customers to go directly to apparel, baby, electronics, entertainment, home, jewelry, pharmacy, photo, sports, and toys and then search for items within each of those categories or do a search of the site using specific key words or phrases. In a close correspondence to the way that people grocery shop in the real world, Peapod, an on-line grocery store, allows customers various shopping options including browsing the aisles (e.g., going directly to the baking "aisle" and browsing the items there), express shop (which allows the customer to input a grocery list that the application software uses as key words to search the database so that it can present a list of options from which the customer can choose), and browse the current specials.

3. ELECTRONIC MARKETS

Another facet of e-commerce is the electronic market. These are collections of individual shops that can be accessed through a single location on the Internet that are the virtual equivalent of shopping malls. Electronic markets are also referred to as cybermalls or electronic malls. These services can be particularly attractive to customers who need a product or service but do not know where to find it. For example, alibaba.com enables shoppers to find products and services in a wide variety of areas ranging alibaba from arts and entertainment, beauty and fashion, and travel through home and family, professional services, and real estate. Within each of these general categories are hyperlinks to more specific categories or specific e-commerce web sites for individual businesses. For example, under the alibaba.com heading of "Travel Center" are listed agencies, cruise packages, discounts, general, tours, and vacation resorts. To find a travel agency, for example, one would click on the "Agencies" hyperlink, which would then bring up a list of travel agencies participating in the e-market along with a brief description and a hyperlink to their home page.

4. ON-LINE AUCTIONS

E-commerce is not only the electronic equivalent of a physical retail store or mall, however. The third popular approach to e-commerce is through on-line auctions. This approach to e-commerce allows customers to determine the price of products rather than paying a fixed, predetermined amount. There are three on-line approaches to auctions. In the forward auction, shoppers bid on an item and the seller takes the highest offer. In the reverse auction, bidders list their product or service needs and the maximum they are willing to pay for the product or service. Sellers then bid against each other in an attempt to offer the product or service at the lowest price. The third approach to on-line auctions is the Dutch auction. In this approach, an item is offered on-line at a high opening price. At a predetermined interval, if no one has purchased the item, the price is lowered. This continues until someone is willing to purchase the item at the price offered.

5. DEVELOPMENT OF A STRATEGIC PLAN

First, one must consider the place of e-commerce within the greater strategic plan for the business and the strategy of the electronic channel specifically. Success in business – whether it is through convention or electronic means – must be based on a well-developed strategy. This strategy is a plan of action to help the organization reach its goals and objectives. As for any other business, organizations that desire to add an e-commerce channel to their efforts need to first determine whether or not this approach is appropriate and possible. Failure to do so can lead to failure of the enterprise. Strategic planning is the process that helps the organization determine what goals to set and how to reach them. Through strategic planning, the organization determines and articulates its long-term goals and develops a plan to use its resources – including materials, equipment and technology, and personnel – in reaching these goals. The resultant business plan summarizes the operational and financial objectives of the organization and is supported by detailed plans and budgets to show how these objectives will be achieved. The business plan also analyzes the risk involved in reaching these goals. In business terms, risk can be defined as the quantifiable probability that a financial investment's actual return will be lower than expected. Higher risks mean both a greater probability of loss as well as the possibility of a concomitant greater return on investment. In addition to the objectives or end goals that the organization is trying to reach, strategic planning should also consider what resources are necessary to accomplish these. This should include the financial resources of the organization such as capital structure, new issues of common stock, cash flow, working capital needed, dividend payments, and collection periods. In addition, the strategic planning process also needs to give consideration to the resources necessary to accomplish the other objectives including costs of a web presence, physical infrastructure, employees needed, and so forth.

6. WEB SITE CONTENT AND DESIGN

Once a strategic plan has been developed, the next step is for the organization to flesh out its concept and design a web site. Part of this process is the choice of a domain name. This is a unique, easily understood identifier for a set of addresses on a network. On the Internet, domain names comprise two parts: The first part of the name typically relates to the owner of the name and the second part roughly designates what kind of business it is. Having a domain name that is easily remembered and associated with the company's name can help customers remember the company and find its web site in the future for additional transactions. Domain names that are not closely associated with the organization and its products, however, are less likely to be remembered and revisited. The design of the web site must also take into consideration the image that the business wishes to project. Branding is just as important for e-commerce as it is for conventional marketing efforts. For example, the design of a web site for a high-end consulting firm would undoubtedly be different from the design of a web site for a firm that is selling toys; not only in content, but in layout and other design features. The consulting firm, for example, might want to include a video clip from its president, use a subdued color palette that conveys the serious nature of the services offered, and include a significant amount of white space or quotes from satisfied customers. The web site for the toy company, however, might include animation, bright colors that appeal to children and convey the notion of fun, and use a lot of color pictures of toys. An organization needs to work closely with its web designer to develop a concept that will help put across what kind of organization it is. No matter the type of product or service offered by the web site, consideration needs to be given to the inclusion of appropriate customer service features. Web sites can collect data for use in customer relationship management, including organizing, tracking, and analyzing customer data and buying patterns. As mentioned above, e-commerce web sites also need to be user friendly, allowing customers to easily navigate the site and find the information that they need. Although this often means the inclusion of search engine capabilities, it also means providing a way for customers or prospective customers to contact the business to get the information that they need. This can be done by providing contact information for phone and mail, having a pop-up e-mail message form or an on-line form to fill out to submit questions, or on-line chat.

To do business over the Internet, an organization must first have a presence on the World Wide Web. To be an effective tool for conducting e-commerce, a web site must be well-designed. This means that it is attractive and otherwise memorable so that customers will think of that site and the organization when they again need a similar product or service. In addition, the web site must be user friendly, that is, easy for the customer to use. Most consumers who use the Internet for e-commerce have experienced web sites that are difficult to navigate, do not provide all the necessary information needed to make a transaction, or that attempt to force the consumer to provide personal information that is not necessary to the transaction. When such qualities become egregious enough, consumers are unlikely to return to the site in the future. Without the loyalty of repeat customers, the e-commerce venture is unlikely to be a success. There are a number of considerations to be taken into account when developing a web site for e-commerce.

7. ADVANTAGES OF E-COMMERCE

Traditionally, shopping meant physically going to a store, searching aisles for products of interest, comparing product features, and purchasing. The traditional business model for business-to-business operations involves a procurement staff that negotiates with various suppliers. In the e-commerce business model, a procurement staff shops online for supplies and other items necessary to the business. Just as it does for the consumer in the business-to-consumer business model, the Internet allows businesses to comparison shop online in order to find the most appropriate product at the best price. This reduces many of the front-end costs for finding goods and products that are incurred in the traditional model.

One of the most common applications of business-to-consumer e-business applications is online retailing and electronic storefronts. In online retailing, customers are able to visit a business' web site on the Internet and examine product pictures and information, compare different products, fill an electronic shopping cart, and checkout and pay for their purchases. The e-business equivalent of the storefront is the home page of the company's web site, with the various web pages being the electronic equivalent of the aisles of a traditional store. Rather than going to a brick and mortar store, customers can visit the company's web site and look for the items that they need. Most businesses engaging in e-business allow customers to "stroll the aisles" by having links to various categories of products.

8. CONCLUSIONS

The World Wide Web offers business users a wide range of opportunities for advertising and marketing their goods and services as well as for conducting e-commerce over the Internet. Some of the more commonly used capabilities of the Web include the use of Web sites, business-to-customer e-business applications, and business-to-business e-business applications. In addition, as advances in information technology continue, businesses find new and innovative ways to use the Web to reach customers and partners alike.

REFERENCES

[1] C.Zhou, "The new weapon of network marketing", E-commerce World, No.7, July 2006, pp.58-59.
[2] Xiaotong Li, John D. Johnson. Evaluate IT Investment Opportunities using Real Options Theory [J].Information Resources Management Journal, 2002, (3):32-47.
[3] Gunjan Samtani, Dimple Sadhwani.J2EE-based Application Servers, Web Services and EAI.Web Services Journal, may 2002.
[4] Geert Van de Putte,et al.Using Web Services for Business Integration.IBM Red books.2004.4.
[5] Hongbing Wang,et al.Web services:problems and future directions.

SERVER BASED SOFTWARE BEHAVIOR COMPREHENSION *

Kong Weiguang[1], Wang Yuan[2]
[1]College of Computer Science, Wuhan University of Science and Engineering, Wuhan
[2]Library, Zhongnan University of Economics and Law
Wuhan, 430074, China
Email: wgkong@public.wh.hb.cn

ABSTRACT

Client-Server mode is widely used in distributed computing system. Behaviors from software of client means users' intention usually, software behavior controlling technology meets the demand of forward defense in information systems. Software behaviors from different environment have different base elements such as API callings, menu selections or requests, and different technologies are required to monitor and understand them. In server, request behavior as the research object of this paper is mainly composed of requests from client users. Request behavior is better than request to be the base element of log and access control. Request is the component of one or more request behaviors. This paper, based on theory of software behavior, puts forward the definition of request behavior and behavior set, and studies the expression of request behavior, and builds request behavior comprehension model based on fuzzy matching algorithm. It can get the finished request behaviors and partly finished request behaviors from user's request list, and the degree of partly finished request behaviors which may be useful in user's behavior forecast.

Keywords: Client-Server mode, Behavior comprehension, Behavior set, Fuzzy matching, Request behavior.

1. INTRODUCTION

Client-Server mode is widely used in distributed computing systems. We concern not only what client software and subjects can do, but also whether it has done anything forbidden or not[1].Client behaviors are performed by software's requests submitted to server. It is necessary to monitor and control client behaviors to hold the order of network system.

Forward defense technology becomes the development trend of security with the faster producing speed of threat in information systems. As one of the main way of forward defense, behavior controlling technology will monitor and control software behaviors, and prevent any behavior figured dangerous[2]. It runs in real system and consults with security manager while underlying vicious behavior is occurring, and it can judge software behaviors in more intelligent way by analyzing their main features.

Up to now, the main implement of behavior controlling is watching the probably dangerous operations which are key operations of vicious behaviors such as system registration modification, system process injection, keyboard recording and program hiding. There are limitations to regard an isolated operation as abnormal software behavior because normal software may perform these operations too. The better approach is monitoring and comprehending the software behavior so as to filter target software behaviors accurately.
More and more business has been migrated to the Internet with the popularization of Internet applications, and the security problem of applications has become serious. There doesn't only exist vicious operate intention from invalid users, but also exist requests exceed one's authority initiated by valid users. Servers must face this kind of threat timely. Behavior controlling technology may be one appropriate settlement[3].

In Client-Server mode, client submits request to server, and server executes the request and sends back the result. Through this style, one server can serve more clients.

Access controlling technology can partly defeat this kind of threat. It can identify one client and bind it to its authorities so that the server can check according to its authorities. The bounded authority is usually client requests. Server may deny invalid clients and requests not authorized. In access controlling model, client is usually bound with its authorities statically, and authorities do not change with time or space.
Log may record series of requests occurred in the system, and its basic record element is requests. Administrators can resume the finished process according to log records, but the comprehension of log record is now manual work.
In server, log and access control is commonly based on request operations, but the most appropriate basic element of log and access control is request behavior because it is clients' behavior that implies user's intention. One operation is isolated, and it is the component part of one or more request behavior. Client's request behavior is abstract; one request behavior may be composed by a series of requests.
It is possible to enhance the security control ability of server by controlling clients' request behaviors in real time if log and access control found on request behavior. Server will refuse request behaviors violate the security rules.

The author of reference [1] advanced the definition of general software behavior and pointed out that there exists several levels of software behaviors, and he studied the work style and target of software behavior controlling, monitoring, authentication and confrontation. Ivan Porres researched modeling and analyzing of software behavior in UML[4]. Reference [5] and [6] studied automatic classification of software behavior. In reference[7], the author researched the pattern recognition and analysis of software behavior, the model can establish a state of health of the target software by comparing real-time measurements with the baseline indicating the target software performance and activity. The baseline is automatically derived from a subset of measurements by a third-part tool. The tool includes sensors embedded into the target software to measure specific code segments and examiner which receives measurements from the sensors. The baseline made up of feedback measurements represents the target software's behavior.

The primary work of this paper is researching the method to build request behaviors from client request list. The steps are monitoring client request as basic elements of request behavior, forming request sequence in real time, and matching the running request sequence with each request sequence of request behaviors in request behavior set defined by software designer. The matching result includes information of finished and partly finished request behaviors, and supports request behavior controlling and forecast.

* This work is partially supported by Hubei Education Foundation Grant #200717005 to Mr. Hu Ming.

2. SOFTWARE BEHAVIOR COMPREHENSION

2.1 Definitions of Request Behavior

Client behavior act on server will be carried out by a series of requests. Request behavior may be defined as following:

Request Behavior::= Behavior_Id
×Behavior_Type
×Behavior_ClientId
×Behavior_ServerId
×Behavior_Time
×< Request_Sequence >

Behavior_Id is the exclusive Id of this behavior assigned by server for the convenience of managing. Behavior_Type is the type of this behavior classified by server. Behavior_ClientId and Behavior_ServerId represent the subject and object of this behavior, Behavior_Time indicates the appropriate time domain of the behavior.

Request_Sequence means the necessary requires in order for one behavior. There are a great amount of different requests, only specified requests are collected into Request_Sequence, and there must be one or more signal request in one Reqest_Sequence.

In one specified system, all possible request behaviors are included in its behavior set:

Behavior Set::={ Request Behaviors }

Subset of one behavior set can be made from behavior set by classifying its request behaviors, such as forbidden request behavior set, permitted request behavior set and dangerous behavior set. The server will deny behaviors in forbidden behavior set, perform behaviors in permitted behavior set and alarm behaviors in dangerous behavior set.

2.2 Server Based Request Behavior Comprehension

Request behavior comprehension process will end up with a conclusion: which behaviors are finished in a specified request sequence and the contain degree (discussed below) of partly finished behaviors in the request sequence. The key task is determining whether a behavior's Request_Sequence is contained in the specified request sequence or not, and computing the contain degree of a request behavior.

There are two steps to perform request behavior comprehension:

1) Ensuring that the compounding form (subject, object, time) of a request behavior is consistent with the running process. The compounding form is commonly assured and relatively easy to get and compare.

2) Comparing the request sequence in a request behavior with the request sequence in the running process in order to decide the contain degree of the request behavior contained in the running process.

For the convenience of computing, Request_Sequence of each request behavior and the specified running request sequence must be expressed as a matrix, and the contain degree can be computed by fuzzy computing.

On the assumption that one standard request behavior named B has been set, and B contains a required sequence expressed as $(r_1,r_2,\ldots\ldots r_M)$. We can distribute one weight a_i ($0\le a_i\le 1$) to r_i, and the sum of all a_i is 1. Matrix $(R_1, R_2, R_3, \ldots\ldots, R_N)$ represents all possible requires on the server, B can be extend to the matrix below by a special rule: if R_i appears in B as r_j then replace R_i with a_j, else replace R_i with 0 implying that B has no relation with R_i .

$$B=(a_1, a_2, a_3,\ldots\ldots, a_N) \qquad \sum_{i=1}^{N} a_i = 1 \tag{1}$$

Through monitoring the requests of running process, the running request sequence b will be expressed as matrix:

$$b=(r_1, r_2, r_3,\ldots\ldots r_N) \tag{2}$$

All r_i are initialized to 0, if R_i has been detected in the running process, then r_i will be set to 1, else r_i remains 0, and r_i will be set back to 0 if R_i is retracted.

Matrix b changes with the going of process. We can match b and B with fuzzy operation "and" in real time, this operation may result with a value between 0 and 1. The result means the contain degree (d) of request behavior B contained in running request sequence b.

$$d=B\bullet b=\begin{pmatrix} \mathbf{a_1}, \mathbf{a_2}, & \ldots\ldots, \mathbf{a_N} \end{pmatrix}\bullet\begin{pmatrix} \mathbf{r_1} \\ \mathbf{r_2} \\ \ldots \\ \ldots \\ \mathbf{r_N} \end{pmatrix}=\sum_{i=1}^{N}(a_i \times r_i) \tag{3}$$

If the contain degree d is 1, it means that request sequence b contain request behavior B; if the value of d is 0, it means none of requests in behavior B is included in request sequence b; if d is between 0 and 1, it means part of requests in behavior B has been found in request sequence b. Contain degree d represents the weight of occurred requests in request behavior b, and it can help information system forecast user's next operation.

Because there are more than one request behavior in a request behavior set commonly, each request behavior must be fuzzy matched with running request sequence b to get its contain degree, and contain degrees of all request behaviors may form a behavior contain report.

2.3 Server Based Request Behavior Control

The log system will analyze all request behaviors finished or partly finished in recorded data, and the behavior control mechanism may predict the probability of the running request process contains a specified request behavior and answer the predict suitably.

The request behavior control flowchart may have four important functions to work together (Fig. 1); they are request monitor, request behavior comprehension, behavior control strategy database and request behavior control mechanism.

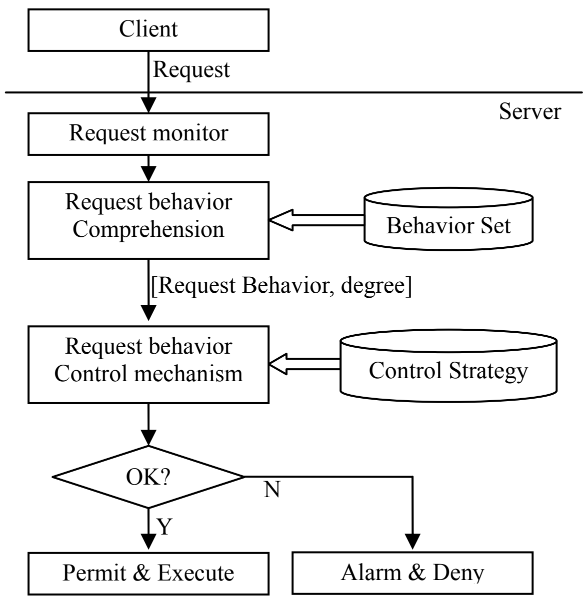


Figure 1. Flowchart of Request Behavior Control

In a server application, there might be several clients' link at the same time. Request monitor will pick up each request from one specified user continuously, obtain the basic elements information including subject, object and time, generate and update the request sequence matrix b of current process, and pass the half-backed request sequence matrix b to request behavior comprehension function.

When a request sequence matrix b arrives, the request behavior comprehension function will fuzzy match it with each request behavior in behavior set, and output a contain degree report to request behavior control mechanism.

The request behavior control mechanism will decide the next step according to its behavior control strategy database and the contain degree report. Generally, if one running process has been judged abnormal, the decision will be turned false and the process will be interrupted.

For example, in simulate network bank system, there are requests (except Login and Logout) as following:

1) Account_Enquiry
2) Account_Transfer_Form_Request
3) Transfer_Confirm
4) Payment_Form_Request
5) Payment_Confirm

Such request behaviors can be defined:：:

Enquiry=(1, 0, 0, 0, 0)
Transfer=(0, 0.7, 0.3, 0, 0)
Payment=(0, 0, 0, 0.7, 0.3)

Now userA login this system, and the account number is 8001. The user queries his account once, and then he requests the account transfer from, but he has not submitted his form. At this time, his request behavior report can be gotten in the background of server. (Table I)

Table I Request Behavior Report

Current Request Behavior Report

Subject: User1
Object: Account 8001
Time: 2009-8-1 16:30
Status: Login

Request Behavior Name	Probability
Enquiry	1
Transfer	0.7
Payment	0

Following steps generates the result of probability:
1, the system gets the running request matrix:

b=(1, 1, 0, 0, 0)

2, the probability of each behavior is generated ("Transfer" for instance):

d(Transfer)=Transfer $\bullet$ b

$$= \left(0, 0.7, 0.3, 0, 0\right) \bullet \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = 0.7$$

The result indicates that 70% weight of request behavior "Transfer" has occurred. It tells us that this user has finished behavior "Enquiry", he may be doing behavior "Transfer" and it can be predicted that the next request would be Transfer_Confirm.

3. CONCLUSIONS

Request from client to server is relatively easy to be monitored and controlled by server. Request behavior comprehension model developed in this paper is found on fuzzy matching of request sequences, it can catch on finished request behaviors and contain degree of partly finished request behaviors. The output is propitious to be the basic element of log system and real time behavior controlling. Server will be more secure by adopting this kind of forwardly defensive technology. In the meantime, the model increases the server processing expenses while enhancing system safety.

Synthetically, along with flooding of network crimes, forward defense will take more important role in the future. The request behavior comprehension model is feasible to be the foundation of server based request behavior control.

REFERENCES

[1] Yanwen Qu, Software Behavior(in Chinese). Beijing:Publishing House of Electronics industry, 2004.10.

[2] W. Gao, "Forwardly Defensive Technology", Software World, vol.19, 2007.10.

[3] Tim Mccollum, "Applications control: Software behavior monitoring offers another layer of protection to corporate information systems", Computers & Auditing, Internal Auditor, 2002.4.

[4] Ivan Porres.(2001), "Modeling and Analyzing Software Behavior in UML", [Online] Available: http://citeseer.ist.psu.edu /porres01modeling.html

[5] James F. Bowring, "Software Behavior: Automatic Classification and its Applications", Georgia Institute of Technology, 2003.

[6] James F. Bowring, "Active Learning for Automatic Classification of Software Behavior", [Online] Available: http://www.cc.gatech.edu/aristotle/Publications /Papers/p398-bowring.pdf.

[7] Noy, "Software behavior pattern recognition and analysis", United States Patent 7269824, [Online] Available: http://www.freepatentsonline.com/7269824.html.

[8] B. Liu, "The Summarize of the Technique about Proactive Network Security Protection". Proceedings of the 11th China Symposium on Computer Application in Mordern Science & Technology, 2003.9.

[9] Fred Cohen, "Computer Viruses - Theory and Experiments", Computers & Security, vol.6, no.1, pp.22-35, 1987.

Kong Weiguang is an associate professor in college of computer science in Wuhan University of science and engineering, He graduated from Wuhan University in 1991, he has published two books, over 15 Journal papers, and his research interest is in network security.

THE RESEARCH ON COLLABORATIVE DESIGN OF SUPPLY CHAIN PRODUCTION BASED ON GRID TECHNIQUE

Hua Fang[1], Hua Jiang[2]
Department of information Shandong Science Technology Vocational College
Weifang ,Shandong province 261053,China
Email: sdfanghua@sina.com

ABSTRACT

It tries to apply the grid technique into collaborative design and puts forward the architecture of collaborative design system based on grid technique. The collaborative design based on grid technique provides users with the design results or the intermediate results of all design nodes in type of grid services. And design nodes can package their design results to network services and login them to server for other nodes to call and refer conveniently. By this way, the flexibility and scalability of collaborative design platform is increased.

Key words: Grid Service, Node, Collaborative Design

1. INTRODUCTION

Supply chain of manufacture companies contains all the business activities of the production or service from suppliers to clients. Supply chain includes not only manufacturers and parts (raw material) suppliers, but also wholesalers (distributors), outsourcing services, retailers and customers themselves. Each organization of supply chain performs different procedures and interacts with others to realize the profit of the whole supply chain. And the supply chain of manufacture companies must have rapid production innovation to adapt the changes and developments of market. The key method is to use all types of advanced design technologies, manufacturing technologies, quality assurance technology and so on that are required for product development. But it is difficult for the core company in supply chain to master and apply the relative techniques relative to new production in short time, the collaborative design and development between core companies with multiple suppliers is necessary. The research result indicates that if the suppliers join the production innovation procedure early, the techniques of the companies in supply chain can be merged and the cycle of R&D of product innovation team can be shorten greatly. Because the suppliers understand the parts design of relative production more than manufactures in some degree, so the method can decrease the repeatability of development processing of new products, shorten the cycle of development, and introduce the products into markets rapidly.

Collaborative design of supply chain production innovation mainly includes four parts: collaborative design oriented to product design and manufacture, collaborative design benefit to project management, collaborative design meeting customer needs and collaborative design meeting requirements of real-time supply. This paper will analyze the collaborative design oriented to product design. Collaborative designer supported by computers requires every node enterprise can browse, download, negotiate and communicate the CAD design data by remote operation. For the research on collaborative design, literature [1] built a group of networked collaborative design tools, literature [2] did research on it in terms of graph share, literature [3] proposed a lightweight model of product model data and deal with the data transmission issue in design environment, and literature [4] analyzed the method for grid to be applied to collaborative design in methodology view. But the platform built by collaborative design lacks the support of dynamic changes virtual organization collaboration in literature [1，3], so this paper will build the collaborative design based on the grid platform in technical view, and add the flexibility and expansion of collaborative design; Realize the package of design resources based on the WSRF standard and the collaborative design of supply chain product innovation on the grid platform.

2. THE ARCHITECTURE OF SUPPLY CHAIN PRODUCT INNOVATION COLLABORATIVE DESIGN SYSTEM BASED ON GRID TECHNIQUE

Collaborative design needs the share, negotiation and communication of design resources, so the architecture of supply chain product innovation collaborative design system based on grid technique should meet the function requirements as follows: task decomposition and distribution of specific products of supply chain core enterprise, the display of every design resource model, the package function of the design resource of node enterprise in supply chain, and so on. More information please review Figure 1.

2.1. The Resources of Collaborative Design

Every node enterprise on the supply chain supplies the design resources because they have advanced technology and production capacity on the Research and development for a certain category of products or parts, which improve the overall benefit of supply chain product development. The resources of collaborative design include CAD template library and CAD file library that every node enterprise on the supply chain forms when they design and develop parts of products. The CAD library files may be design results or the intermediate results so that the collaborative designer can refer the relative design results at any time conveniently.

2.2. Extraction Tool of Secondary Development

It realizes the extraction of feature information, geometric information and files of CAD template library and CAD file library and the conversion of format through secondary development of CAD software. The non-geometric information of CAD template which contains rich semantic information is extracted by the feature information extraction tool, such as material size, geometric tolerance, heat treatment, and surface roughness, and other works. The Geometric topology information of CAD template is extracted by the geometric information extraction tool. The CAD files are extracted and put in the grid service resource files by the tool of file extraction for convenient download. And the format conversion function is also provided to transfer the CAD file into VRML file so that clients can browse the geometric model of the parts design results on the by installing VRML plug-ins. Because of the difference of the CAD systems, the language of secondary development is different, and the extraction tool should be developed for different CAD systems.

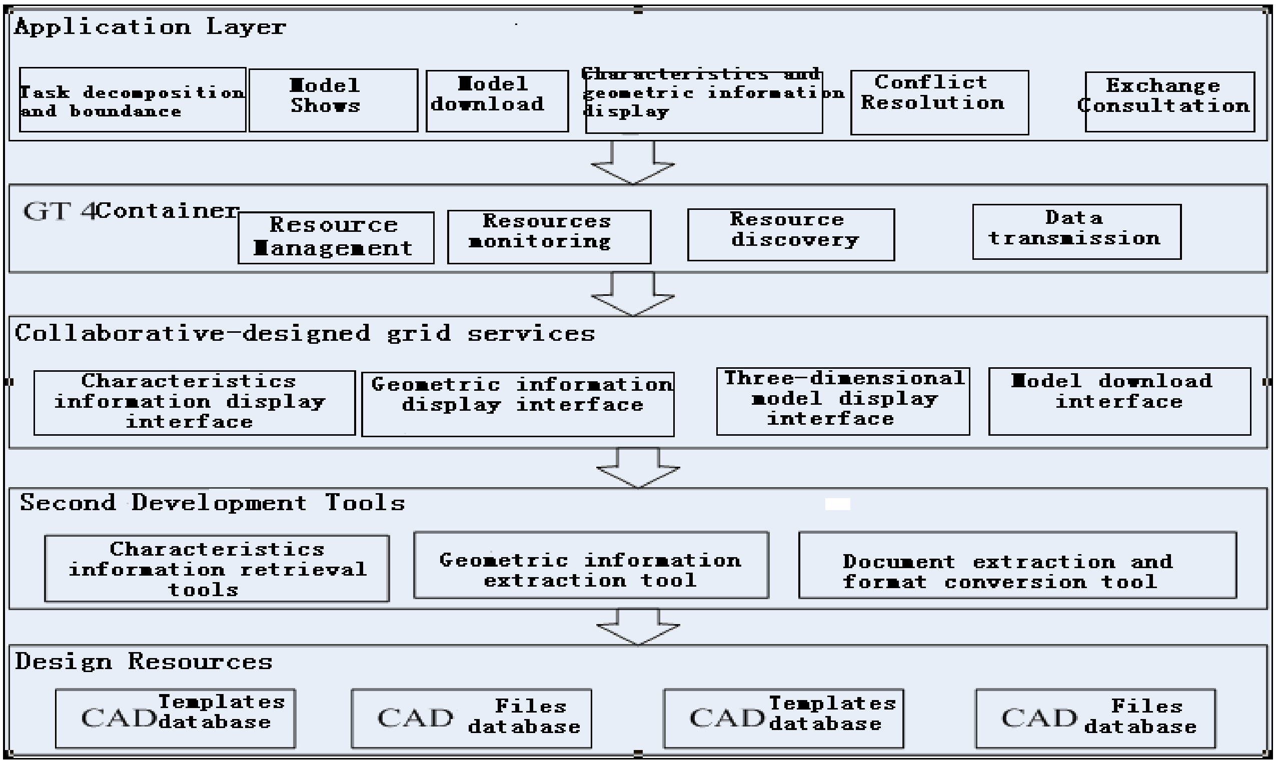


Figure 1. The Architecture of Supply Chain Product Innovation Collaborative Design System Based on Grid Technique

2.3. Grid service based on collaborative design

By packing the output of each secondary development extraction tools as grid services, it shields the difference of secondary development extraction tools and provides the Unified query and access interface for the foundation layer of grid. Grid service based on collaborative design includes four interfaces: characteristics information display interface, geometric information display interface, three-dimensional model display interface and model download interface.

2.4. Foundation Layer of grid

Foundation Layer of grid achieves the management of grid services, provides the resources management, data transfer, security and information services. It provides the Operating environment and operating platform for grid services. It could be de developed through the latest version GT4 of pop toolkit.

2.5. Application Layer

Application Layer should include the functions of task decomposition and bound, model display, model download, display of characteristics and geometric information, Conflict Resolution, the exchange and consultation, and so on.

The function of task decomposition and bound provides interface for the core enterprises of supply chain to publish tasks, so that every node enterprise on the supply chain could understand their design tasks and the rules and constraints should be followed in time.

The function of model display realizes that the designers can mutually browse the progress of the associated design model and any conflict , through the interaction with the three-dimensional model display function of grid services.

The function of model download calls model download interface of grid services for the node of collaborative design to provide the download of the design results. It facilitates of virtual assembly and simulation optimization

The display function of characteristics and geometric information calls the display interface of characteristics information and the display interface of geometric information to provide the display of various types of each parts design information during the product design, including geometry, topology, size, tolerances, processing, assemblage, material, and so on. The function can allow the parties of collaborative design to reference model data of associated parts, so the design could become more efficient and the follow-up CAM technology integration could become more convenient.

The function of conflict resolution provides a consensus function for the conflicts during the process of collaborative design. The inevitable conflicts could be appeared during the process of the collaborative design. It may be appeared that supporting conflict, associated geometry conflict, dynamic design conflict, and so on, because there is the difference between design units with the knowledge and experience. After the conflict, firstly conflict resolution should carry on in the exchanges and consultations module between the collaborative design units. If the conflict could not be resolved, the conflict could be put out in the module of conflict resolution function. Then design experts should be organized by the core business for conflict resolution, and the results would be put out in the conflict resolution function module.

The function of exchange and consultation provides the platform for the collaborative design enterprise to exchange. On the platform, all members are free to speak, and view the mentions of other members. This architecture use the Universal application technology—BBS to achieve. Since the BBS is a very mature technology, this article does not give the details.

3. REALIZES MECHANISM OF GRID SERVICES BASED ON COLLABORATIVE DESIGN

To achieve the above-mentioned collaborative design architecture of supply chain product innovation based on grid technique, the following key technologies should be included：① To achieve the secondary development programs for different CAD systems (such as UG, Pro / E, Catia, etc.), which have been detailed in many articles and books. ②To carry on state resources package for the secondary development programs. ③ To achieve Web services of manipulation resources. ④To display the client of characteristic geometry information through calling grid services. ⑤To achieve the client of model display and model downloading through calling the related interfaces of grid services ⑥To achieve the functions of task decomposition and bound, and the function of conflict resolution. These two functions can be achieved by JSP technology, which is simple and convenient. This chapter focuses on that the state resources package, to achieve Web services of collaborative design and the client of calling grid services.

3.1. Package of State Resources

It is easily to manipulate on the state resources through utilizing the WSRF to develop grid services. By packing the model and data of intermediate results of supply chain co-design and development as state resources, the grid services can be easily called. And the intermediate results of the each node enterprises designed and developed can be shared easily. The output data, which are produced during the process of the secondary development program calls the CAD system, are packaged as state resources in this chapter. Since the CAD systems are different, the secondary development tools are different, for example, the secondary development tools of AutoCAD is ObjectARX, the secondary development tools of UG is UG/Open, the secondary development tools of Pro/E is Pro/Toolkit. Although the secondary development tools are different, all of them can use the VC + + development environment, the secondary development program can be compiled to generate .dll files library, so the callings of the can actually be the unified callings of the C + + file library methods.

To package output data of the secondary development program outputs as the state resources, the following methods can be used: program JAVA resources class to achieve ResourceProperties, ResourceIdentifier and other interfaces, use Java Native Interface（JNI）in the JAVA resources classes, call the internal methods (the extraction methods of characteristics information, geometric information and format conversion)of local library which is generated by the secondary development program, access the data of design resources.

3.2. Realization of Collaborative Designs Web Services

The Web services of Collaborative designs are responsible for the manipulation and management of the design resources. Based on the characteristics of the Collaborative design, this chapter divides the Web services into four types interfaces: characteristics information display interface, geometric information display interface, model display interface, model download interface. Every interface calls related resource attributes to achieve the specific functions through programming. For example, the display interface of the characteristics information not only displays all the characteristics information, but also allows users to query the characteristics information that they are interested in.

The Web services of collaborative designs apply to all nodes enterprises. It uses WSDL to describe the related operation interface of Web services and the resources properties documents, and deploys in the Golbus container. Therefore, the resources are packaged for the grid services in the GT4.

4. CASE ANALYSIS

This chapter takes the collaborative design grid services of the small helicopter as the example to demonstrate the preliminary realization of the collaborative design grid services platform. The software package of Globus Toolkit 4.0 is used as the platform of the grid services. Eclipse is used as integrated development environment of developing grid services. Tomcat is used as the Web services container of GT4 WSRF Web application. The implementation steps are as follows:

(1) Firstly, for AutoCAD system and UG systems, develop the corresponding secondary development programs, and save by the form of library files on the collaborative design machine of the each nodes enterprise.

(2) By using the above library file to develop the resource files which are mentioned in 3.1, the characteristics information, geometric information, CAD model files and VRML files are respectively gained.

(3) Program the Web services of collaborative design, and save respectively the collaborative design Web services and the resources properties documents in each design node enterprises.

(4) Deploy the Web services of each node enterprise to the GT4 container.

(5) Start the programs of server application layer to call the grid service, display the results. Figure 2 shows the grid services that may be supplied.

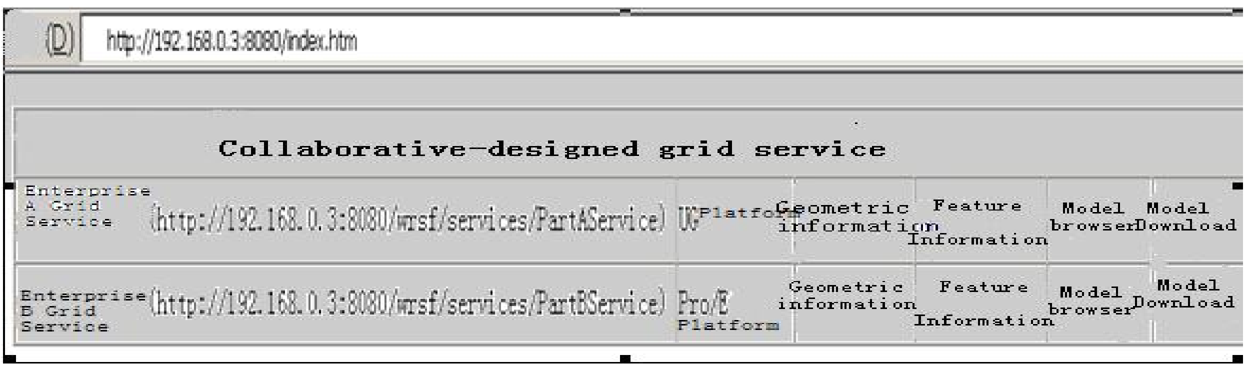


Figure 2. The Grid Service Portal of Cooperative Design

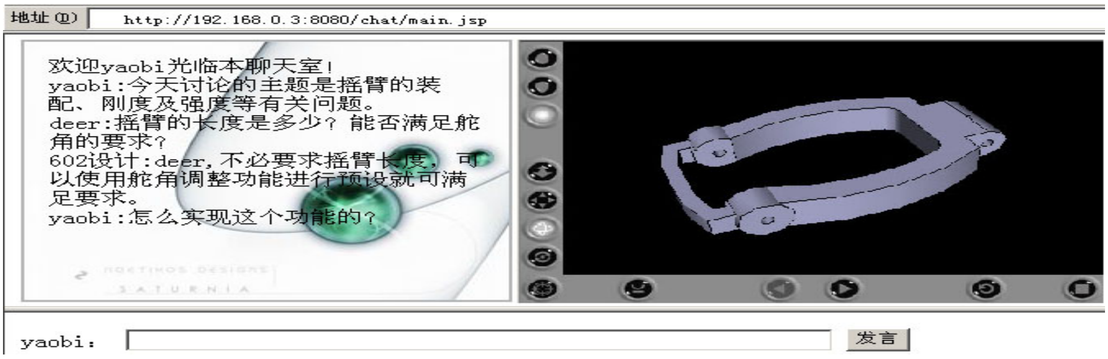


Figure 3. The Result of Enterprise A Transfer the Service of Grid

When the browsing link of grid services model is clicked, the Web server uses Servlet to receive the request, access parameters, and generate the endpoint reference of co-design grid service, then according to parameter and node refer the grid services of enterprise A in GT4 container. By calling the file transferring tool, the grid services of enterprise A should return the design results of enterprise A by the .wrl file format to the client. For client convenient discussion on some parts, the BBS technology is used. Figure 3 is the discussion interface of the small helicopter rocker partial model which is returned to the browser.

5. CONCLUSION

Collaborative design of supply chain production innovation of manufacture companies could make full use of core technology of suppliers and shorten the cycle of product innovation. By providing the design result or the intermediate result of each design node by the form of grid services to the users, the collaborative design based on the grid increases the flexibility and the extendibility of the collaborative design platform. By proposing the architecture of collaborative design system based on grid technique and integrating the secondary development technology of CAD system into the grid services calls it effectively shields the isomerism of each CAD system. This paper is only the initial attempt of applying the grid technology into the collaborative design. The next step will be to further improve the function of collaborative design platform, integrate the functions of the virtual assembly and the simulation optimization.

REFERENCES

[1] Ling Tian, Jizhong Chen. Network Collaborative Design Tools. China Mechanical Engineering, 2004, 10: 1774-1777

[2] Gang He, Bolin Li. Research on Graphics Sharing Technology for Computer Supported Collaborative Design. Mechanical Science and Technology for Aerospace Engineering, 2004, 4: 45-47

[3] Qifu Wang, Lei Wang. Model Simplification in Collaborative Design. Journal Of Computer-aided Design & Computer Graphics, 2006, 1: 108-113

[4] Zhigang Hu, Zhaokun Yan. Rearch on Modern Collaborative Design Based on Grid. Journal of Central South University, 2004, 12: 988-992

[5] Zhisheng Chen, Liping Huang. Supply Chain Management Information System Based on WSRF. COMPUTER ENGINEERING, 2007, 24: 68-70

Author synopsis: **Hua FANG** (1967--) female, master, professor. Her current research interests are network &database, distributional system theory and algorithm.
E-mail: sdfanghua@sina.com

FAST CALCULATION OF CELL VALUE IN WEB REPORT SYSTEM

Aimin Li, Jing Zhao, Qian Gao
Shandong Institute of Light Industry
Jinan, Shandong, 250353, China
Email: lam@sdili.edu.cn

ABSTRACT

In Web report system, there is complex dependence between cells because of calculation formula of each cell. It is important to arrange the formula calculation sequence reasonably for improving the computation efficiency and decreasing the computational cost. The directed acyclic graph and dividing subgraph are proposed to arrange formula calculation sequence. Then the formal definition of Web report is expounded in detail. After analyzing concrete problem, the Java implementation of algorithm is introduced at last.

Keywords: Web Report, Directed Graph, Report Cell, Divide Subgraph

1. INTRODUCTION

Web report is a typical application of transforming from traditional report system, such as Microsoft Excel, to Web. It transplants traditional business to Web and comes into being a distributed information management system in network. Web report has some advantages comparing to traditional report system. But it has some disadvantages too. In the new application model, client uses Web report system by browser, and most business logical operation is assumed by sever. It must have time delay because of network own attributes. The present problem is how to guarantee that the server can respond rapidly and accurately to client's request. The report system table is composed of cells of n rows and m columns. There are large computation and quotation between cells. It is the key problem how to calculate cell value rapidly to respond to client's following operation.

The rest structure of this paper is as follows. In section 2, the problem analysis of Web report is presented. The solution by establishing directed acyclic graph and dividing sub graph to determine formula evaluation sequence is described in section 3. Finally, the conclusion and future research are put forward in section 4.

2. PROBLEM ANALYSIS

The computation and quotation between cells are called formula in Web report. The value of a cell is depended on other cells if it has formula, if other cells have formula too, the other cells value are depended else cells value. The procedure of calculating a cell value is complex because of the dependence between cells. If the formula calculation sequence is not logical, the computational complexity is increased rapidly because of calculating some cell value repeatedly. Fig.1 is an example of report. It has 8 rows and 7 columns.

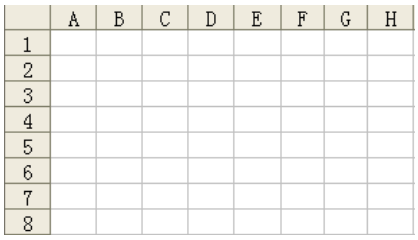


Figure 1. Report Example

Example 1 In Fig.1, suppose formula [A,1]=1+[B,1], it means cell [A,1] value is the sum of cell [B,1] value and 1. In the calculation procedure, the cell [A,1] quotes the cell [B,1] value.

As can be seen from example 1, the formula in one cell quotes other cells name so as to cause the dependence and computation order of priority between cells. Supposed two formula f1 and f2 in different cells, f1 quotes f2. So the computation f1 is prior to f2. The two different cells have one dependence because of formula f1.

Example 2 In Fig.1, suppose formula as follows:

f1: [F,3]=[E,3]+[C,3]
f2: [E,3]=[A,3]+[D,3]
f3: [C,3]=[A,3]+1

It can be seen from example 2 that formula f1 quotes two cells which respectively have formula f2 and f3. f2 and f3 quote the same cell. If computational process does not consider the dependence between cells, then the computational complexity is large. The value of f1 depends on value of cell [E,3] and [C,3]; the value of f2 depends on value of cell [A,3] and [D,3]; the value of f3 depends on value of cell [A,3]. In above computation procedure, the computational complexity of cell [E,3] and [C,3] is 2, whereas the computational complexity of cell [A,3] is 5 and [D,3] is 3. Most computational complexity is repeated. If the dependence is considered, all the computational complexity of cell is 5. It means reducing half computational complexity and increasing efficiency by double.

Suppose F is a formula set of a report, Because of quoting cell name in formula, there must be dependency between cells in a report.. The next task is to find an optimal computation sequence which can release the repetitive computation according to the dependency between cells. Next section is the solution.

3. SOLUTION

If the dependence between cells is defined according to formula set and dependency graph of cells is built, then the above question can be transformed to the relevant problem in graph theory. The solution can be obtained by analyzing the properties of the dependency graph. Firstly some definitions are given.

3.1 Graphical Representation of Web Report

Definition 1. The cells set of report is $V = \{c_i, \cdots, c_n\}$, where c_i is defined as a cell in report, i=1,…,n .

Definition 2. The formula set of report cells is $F = \{f_i, \cdots, f_n\}$, where f_i is the formula of c_i, formula set is a set of formula in different cells.

Definition 3. $E = \{<c_i, c_j>, \cdots, <c_k, c_l>\}$ is dependence set, where $i, j, k, l = 1, \cdots, n$, $<c_i, c_j>$ exists while c_i must have a formula f_i which quotes c_j.

Obviously, any report can be abstracted by graph $G = \langle V, E\rangle$ according to the definition 1,2 and 3, where (1) V is the set of report cells, $V \neq \varnothing$

(2) E is multiple subset of Cartesian product $V \times V$ by definition 2 and definition 3.

Example 3 Suppose that $F = \{f_1, f_2, f_3, f_4, f_5\}$ is the formula set of some web report, where

$$f_1 : [A,1] = ([B,2] + [C,3]) / 2$$
$$f_2 : [B,2] = ([D,5]) * 100$$
$$f_3 : [C,3] = 100 + [E,5]$$
$$f_4 : [D,5] = [E,5] + 100$$
$$f_5 : [E,5] = fun(x, y)$$

The dependence between cells is <$c_{A,1}$,$c_{B,2}$>, <$c_{A,1}$,$c_{C,3}$>, <$c_{B,2}$,$c_{D,5}$>, <$c_{C,3}$,$c_{E,5}$>, <$c_{D,5}$,$c_{E,5}$> by definition 2 and definition 3.

Dependence graph: It is directed graph which is defined by report cells set V and dependence E on V. It is noted as G(V,E), where report cells set V is vertex set of graph G, dependence E is directed arc of graph G. Each cell c of cells set V corresponds to vertex P in graphical representation graph G. $<c_i, c_j> \in E$ means there are a directed arc from vertex Pi to vertex Pj. Fig. 2 is the dependence graph of example 3.

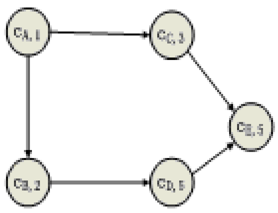


Figure 2. Dependency Graph G_1 of Example3

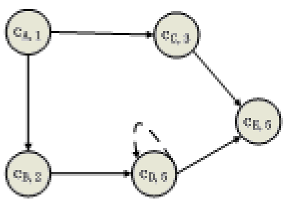


Figure 3. Dependency Graph G_2 of Example4

Example 4 In example 3, if

$$f_4 : [D,5] = [D,5] + [E,5] + 100,$$

the dependence is <cA,1,cB,2>, <cA,1,cC,3>, <cB,2,cD,5>, <cC,3,cE,5>, <cD,5,cE,5>,<cD,5,cD,5> by definition 2 and definition 3. The dependence graph is illustrated in Fig. 3. the broken line is presented the differentia between example 3 and example 4. As is shown in Fig. 3, there is a directed circuit in it.

3.2 Determine the Computation Sequence by Dividing Subgraph

Definition 4 $G = \langle V, E\rangle$ and $G' = \langle V', E'\rangle$ are directed graph or undirected graph, if $V' \subseteq V$ and $E' \subseteq E$, then G' is called sub graph of G, $G' \subseteq G$.

Definition5

$G = \langle V, E\rangle$, $v_0, v_1, \cdots, v_n \in V$, $e_1, \cdots, e_n \in E$, where e_i is an edge joining v_{i-1} and v_i. The alternating sequence $v_0 e_1 v_1 e_2 \cdots e_n v_n$ is called path from vertex v0 to v_n.

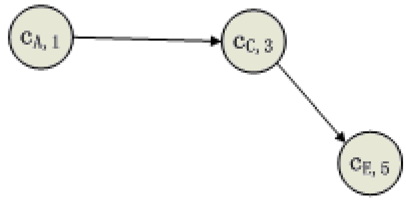


Figure 4. Subgraph G' of G

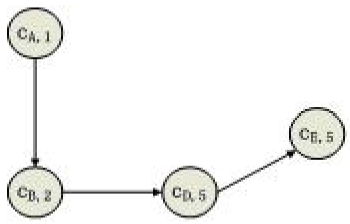


Figure 5. Subgraph G'' of G

As is shown in Fig. 2, suppose that e_1 is the edge from $c_{A,1}$ to $c_{B,2}$, e_2 is the edge from $c_{A,1}$ to $c_{C,3}$, e_3 is the edge from $c_{B,2}$ to $c_{D,5}$, e_4 is the edge from $c_{C,3}$ to $c_{E,5}$, e_5 is the edge from $c_{D,5}$ to $c_{E,5}$ in graph $G_1 = \langle V_1, E_2\rangle$. There are two paths from $c_{A,1}$ to $c_{E,5}$ which are $c_{A,1}e_2c_{C,3}e_4c_{E,5}$ and $c_{A,1}e_1c_{B,2}e_3c_{D,5}e_5c_{E,5}$.

Suppose $E' = \{e_2, e_4\} \subset E_1$, $V' = \{c_{A,1}, c_{C,3}, c_{E,5}\} \subset V_1$, then the graph $G' = \langle V', E'\rangle$ is the subgraph of G1. Similarly, graph $G'' = \langle V'', E''\rangle$ ($V'' = \{c_{A,1}, c_{B,2}, c_{D,5}, c_{E,5}\}$, $E'' = \{e_1, e_3, e_5\}$) is also the subgraph of G1. As is shown in Fig. 4 and Fig. 5, G' and G'' are directed graph. There is only one path from source to sink in each subgraph. The computation sequence q1 of G' is $\langle c_{E,5}, c_{C,3}, c_{A,1}\rangle$ by the dependence between cells. The computation sequence q2 of G'' is $\langle c_{E,5}, c_{D,5}, c_{B,2}, c_{A,1}\rangle$. In graph G, $c_{A,1}$ is crossing point whereas $c_{E,5}$ is meeting point in the two paths. If the vertexes in computation sequences of subgraphs are separately calculated, repetitive computation is not avoided. In order to solve multiple dependence of crossing point and repetitive computation of meeting point, the calculation sequence between vertexes in graph is obtained by cooperation of multiple subgraph vertexes. Then the new computation sequence Q is given. How to obtain Q by q1 and q2 is next emphasis. Following, the method of Pitchingpile is introduced.

Pitchingpile: Firstly, suppose q1 is basic sequence. Then retrieve each element from q1 in turn. For $c_{E,5}$ is a meeting point in q1 and check each element in q2, $c_{E,5}$, $c_{B,2}$ and $c_{D,5}$ which are satisfied calculation constraint, can be inserted into q1 in turn. Then q1 is $<c_{E,5}, c_{B,2}, c_{D,5}, c_{C,3}, c_{A,1}>$. Next crossing point $c_{A,1}$ is not satisfied calculation constraint in q2, it can not insert into the sequence. Therefore, $c_{C,3}$ and $c_{A,1}$ is selected to process orderly in q1. After processing all the vertexes, the basic sequence q1 $<c_{E,5}, c_{B,2}, c_{D,5}, c_{C,3}, c_{A,1}>$ is the new sequence Q.

As is shown in Fig. 3, e_6 is the edge from $c_{D,5}$ to $c_{D,5}$ in graph G2, other edges are same as the edge in graph G1. There are two paths from $c_{A,1}$ to $c_{E,5}$. One is $c_{A,1}e_2c_{C,3}e_4c_{E,5}$ and the other is $c_{A,1}e_1c_{B,2}e_3(c_{D,5}e_6c_{D,5})^n e_5c_{E,5}$. The second path has a simple circuit. If the circuit $(c_{D,5}e_6c_{D,5})^n$ is replaced with $c'_{D,5}$, graph G2 can be converted to directed acyclic graph G2′. In G2′, two paths from $c_{A,1}$ to $c_{E,5}$ are $c_{A,1}e_2c_{C,3}e_4c_{E,5}$ and $c_{A,1}e_1c_{B,2}e_3c'_{D,5}e_5c_{E,5}$. Then the graph can be divided into two subgraphs by two paths and also can calculate the vertex value by aforementioned method. As is shown in Fig.7, vertex $c'_{D,5}$ is self cyclic graph which is consisted of $c_{D,5}$ and e_6 in fact. The calculation procedure of vertex $c'_{D,5}$ is actually the calculation procedure of vertex $c_{D,5}$ by terminating condition. The value of vertex $c_{E,5}$ is prior condition.

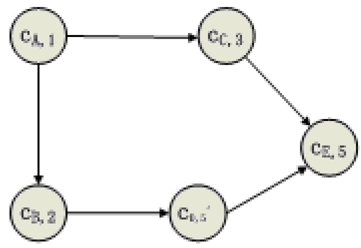


Figure 6. Graph $G_2{}'$

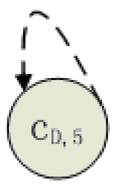


Figure 7. Self Cyclic Graph G_3

3.3 Algorithm for Determining the Formula Calculation Sequence with Dividing Subgraph

3.3.1 Algorithm Idea

Given a report, the cells set is $V=\{c_i,\cdots,c_n\}$, formulas set is $F=\{f_i,\cdots,f_n\}$.

Step 1: The dependence set E is obtained by formula set F, $E=\{<c_i,c_j>,\cdots,<c_k,c_l>\}$, $i,j,k,l=1,\cdots,n$, the directed acyclic graph G is constituted by V and E.

Step 2: Find all paths from source to sink, the paths set $L=\{l_1,\cdots,l_n\}$

Step 3: Divide graph G into some subgraph $\{G_1,G_2,\cdots\}$ by paths set

Step 4: Calculate vertex computation sequence q_i in each subgraph

Step 5: Calculate the computation sequence Q of graph G by inserting piles.

3.3.2 Description of the Algorithm

The algorithm based on Java Language is described as the following procedure.

(1) Establish directed acyclic graph

DAG is directed acyclic graph. GenDAG is a function that can constitute directed acyclic graph by analysis dependence set E. List(E) is linked list generated by E, each e in List(E) is an ordered pair which represents the dependence between two cells. P(e) returns ordered pair inclusive two cells.

```
public DAG GenDAG(E){
    while(List(E)!=null){
        tempe=e;//get a ordered pair from linked list
        p=P(tempe);// return vertex of ordered pair
        DAG.add(p);// add vertex to DAG
        DAG.add(tempe);// add directed edge to DAG
        List(E).remove(tempe);// delete e from linked list
    }
    return DAG;
}
```

(2) Calculate all paths in DAG

Firstly find the vertex that its in-degree and out-degree are zero. Secondly get vertex that its in-degree is zero in turn, and find its next vertex according to dependence set E until the vertex which out-degree is zero, then a path is obtained.

```
public List(L) findPathsInDAG(DAG){
    ListIn0=genIn0(DAG);// return vertex of zero in-degree in DAG
    ListOut0=genOut0(DAG);// return vertex of zero out-degree in DAG
    List paths=genPath(ListIn0，ListOut0，List(E));//if path between vertex of zero in-degree and vertex of //zero out-degree is exist, then get the path
        return paths;
}
```

(3) Get subgraph of DAG

The procedure of getting subgraph of DAG is described as follows: get each path in turn from path linked list, then add vertex and relevant edge to Sub(DAG) according to dependence among vertexes.

```
public Sub(DAG) genSubDAG(List L){}
```

(4) Calculating computation sequence of each subgraph

In subgraph, add each vertex to sequence q in turn by reverse lookup method from sink to source. After all the subgraph is processed , the computation sequence is obtained.

(5) Get DAG's computation sequence by each subgraph's computation sequence

Getting DAG calculation sequence is same to scheduling problem. But the difference is that it sorts subgraph computation sequence by checking dependence set E while normally sorting directly compare each element. The algorithm is described as follows:

```
Public Queue genDAGQueue(List q){
    Queue queue=null;
    While(q!=null){
        List Tempq=q.getList(i);
        While(tempq!=null){
Node=tempq.getNode();// get a vertex from sub computation sequence orderly
            Boolean tag=CheckNode(node);// check node for calculation condition
If(tag){
                queue.add(node);// add vertex to computation sequence
                tempq.remove(node);// delete
```

```
vertex from sub computation sequence
                        delete node from sub sequence;
          }else{
                    add other vertex to queue if it is
    satisfied calculation condition;
          }
        }
      q.remove(tempq);
   } }
```

It can be divided subgraph not only directed acyclic graph but also directed cycle graph by paths. In each subgraph, it must be a unique path, then the computation sequence between vertexes can be obtained by dependence. Thereby repetitive computation on same cell can be avoided effectively and calculation speed can be improved.

4. CONCLUSIONS

There are great deal of formula and complex dependence between cells in Web report. If we do not arrange computation sequence of all formulas reasonably, most computing resource could be wasted. In this paper, the directed acyclic graph and dividing subgraph are proposed to solve the problem. Good results of the application have been obtained.

REFERENCES

[1] John A.Dossey,Albert D.Otto,Charles Vanden Eynden .DSCRETE MATHEMATICS,Fifth Edition,China Machine Press,2007.1

[2] He Fengling,Liu Lei,Zhang Xiaozhi. Solve the Problem of Formula Evaluation Sequence by Directed Graph.Computer Engineering and Application，2003.36:87-89

[3] Joe Francoeur. Algorithms using Java for Spreadsheet Dependent Cell Recomputation (2003). http://arxiv.org/abs/cs/0301036.

[4] Mark Allen Weiss. Data structures and algorithm analysis in C, Posts & Telecom Press, 2005.

[5] Mark Allen Weiss.Data structures and problem solving using Java, Publishing House of Electronics Industry, 2004.

DESIGN AND IMPLEMENTATION DATA PERSISTENCE LAYER MODEL BASED ON J2EE

Xia Sun, Runping Yang, Yan Nie, Rui Huo
College of Science and Technology, Ningbo University
Ningbo, Zhejiang, China
Email: sunxia@nbu.edu.cn

ABSTRACT

Persistent management design in enterprise applications is crucial for performance and flexibility of the whole system. This article proposes an object persistent model that is both extensible and performable. We use proxy model to manage data persistence based on O/R mapping; make access to the underlying data independently encapsulated and make object model and relational database independent of each other by dynamically generating SQL statements and mapping information. In addition, we improve the data accessing efficiency greatly through the dynamic allocation mechanism and the buffering mechanism.

Keywords: Object-oriented, Relational Deatabase, Mapping, Data Persistence

1. INTRODUCTION

The rapid development of IT technology has seen today's object-oriented technology to become the mainstream for enterprise software development. The object model is able to describe and design complex software systems. In software systems, some objects need to be persistent. In most cases this is done via the mature relational database technology. However, due to the difference between the object-oriented model and relational database, there exists the object-relation mapping impedance, forcing the object designer to spend a lot of time to achieve persistence in the target relational database. Therefore, a good data persistent model can free the developer from the complex object-relational mapping and focus on business logic.

The comparison of existing persistence technologies used in J2EE:

1.1 Java Object Serialization

Object serialization is the simplest strategy for Java persistence. Object serialization is a process that turns an object hierarchy into a linear sequence of bytes. However, for complex serialization, object serialization is flawed in many respects: the object serialization must immediately access objects' properties; at the same time the whole object hierarchy needs to be mapped to persistent storage, resulting in a lot of I/O operations. It also lacks a query mechanism once the object is serialized, when changing the attributes of an object, in case of error it cannot "rollback". Therefore object serialization cannot guarantee the data integrity and thus cannot let users to share the data. These shortcomings make a simple object serialization not suitable for most data persistent scenarios.

1.2 JDBC

Compared to the simple object serialization, JDBC is a complex object persistent technology. JDBC requires manual establishment for object-relational mapping, thus it is costly to develop and maintain a data persistence layer. In Java environment, JDBC is the easiest and most basic approach. The advantage of this approach is that it is the most efficient solution; but at the same time it is complex, which requires a lot of manual coding, and is hard to maintain in large projects. Nevertheless, JDBC is still the most popular mechanism to access persistent layer in today's enterprise application development.

1.3 Entity Beans in EJB

Entity Bean is the persistent component in Enterprise JavaBeans (EJB) spec. Entity Beans represents business objects that are to be persistent. It is responsible for automatic persistence of the objects. Entity Bean simplifies framework code and allows for rapid development, hence its prospect is very attractive. As long as the EJB container is configured to know the information about the Entity Beans, the container-managed persistence can complete all of the data access code, which means the code inside Entity Bean itself is greatly reduced. Without JDBC code inside, Entity Bean requires less development and is easier to read and maintain. However, the implementation of the entity Bean mechanism itself has led to another problem: the bean container is not efficient at data access, resulting in poor performance during data access.

As we can see the comparative analysis above, each persistence technology has its own advantages and disadvantages. For the important topic of object persistence in enterprise applications, we need to select a good object-relational mapping technology which will have a direct contribution to the whole system's performance and flexibility. Using any existent technologies, we cannot achieve both development efficiency and performance at the same time, especially with stringent performance requirement in real-world enterprise applications.

2. THE OVERALL DESIGN OF DATA PERSISTENT MODEL

The persistence solution in this article is based on DAO (Data Access Object) data access model. DAO is a model that separates data access logic from business logic, providing a standalone data access interface. The DAO itself hide the persistent layer from the domain business objects, resulting in a better separation of concern. We propose a data persistent model: Data Persistent Manager (DPM). DPM contains three layers: Persistent object layer, persistent manager, and data access layer.

The persistent object layer provides service to business logic layer. From the viewpoint of business logic layer, it implements the data persistence. The major responsibility of persistent manager is to manage the data persistence. This is done via querying the mapping information, interpreting the result from data access layer and returning back to persistent object layer, which in turn return back to business logic layer. The data access layer is what's directly accessing the underlying database, including adding, deleting, modifying, and querying operations.

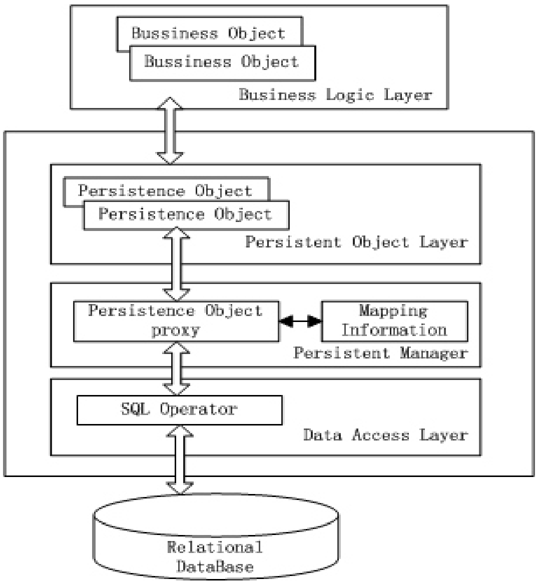


Figure 1. Data Persistence Architecture

3. THE DESIGN AND IMPLEMENTATION OF THE DATA PERSISTENT MODEL

3.1 Persistent Object Layer

Persistence Object is data components that can be stored in persistent storage. From the business logic layer, its interface provides persistent operations. Persistence object uses object-relation mapping to map itself to a relational database. Our design below involves three persistence object classes: Entity class, Relation class, and Rendering class, all of which inherit from the SwellBean class.

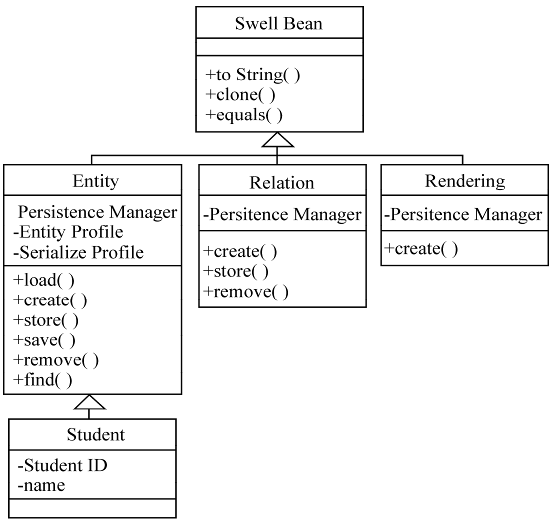


Figure 2. Three Kinds of Persistent Object Class Diagram

SwellBean class: implement the Swell interface; also implement general Java bean’s base class, providing output/compare/clone operations for sub-classes.

Rendering class: base class for view classes. It maps to what users would want to show. The data in memory is not a direct query which results from database, more likely it's a result of logic calculation or analysis. Rendering class is tailored to serve this purpose.

Relation class: base class for relation object classes. It contains object-relation mapping information. Relationship object consists of multiple Entity objects, managing the inter-relationship between these objects.

Entity class: base class for all entity object classes. It represents entities in storage space (e.g., tables and views in relational database). Entity class is the object representation of the relational persistent data. It uses load() and store() to transfer data between Java objects and database.

During the design process, considering performance, DPM does not treat data Bean the same as those EJB container-managed entity beans, whose load/store is totally managed by the container, which may cause many problems including:
- The container would need to evaluate the transaction status of the current entity bean;
- Secondly, if there are multiple entity beans representing the same data, the container would need to deal with the synchronization between these different beans.

Most EJB containers will invoke the "store" operation when the entity bean is passivated and invoke "load" when the entity bean is activated. This may result in unnecessary I/O operations, e.g., even if an entity bean is not changed, it may still be "stored". All these I/O operations will greatly reduce performance. The entity of DPM is designed as explicit persistence, letting the developer to decide when to load data from underlying database and when to store the data back to database. At the same time, the entity bean has hidden the details about persistence and thus developers would only need to decide when to persist, but not how to be persistent. In the sample below, the Student class inherits from Entity class. When instantiating the Student class, the Entity class will also be instantiated. In other words, a Student object contains an Entity object implicitly.

Here is some code excepts when the client application is loading information about a student with id "001".

```
Student student=new Student();
student.setStudentid( “001” );
student.manangerByPersistenceMananager(pm);
student.load();
```

The following code excepts showed the usage of loadEntity.

```
public boolean load() throws CommonException
{    checkManager( "load() #1" );
     return _persistenceManager.loadEntity( this );
       //delegate to the persistentManager to load the Entity
         before it's returned
   }
```

3.2 Persistence Manager

In this Persistent Manager Model, we used proxy as a foundation for persistent manager. In other words, Persistent Object won't directly access the database; instead, this is done via Persistent Manager as a proxy. The load() and store() operations in Entity/Relation/Rendering class delegates to the Persistent Manager, which in turn (using meta-data) find out the corresponding database table, generate sql statement, filling in parameters and execute the sql against the underlying database. The sql result is also passed through Persistent Manager, using meta-data to construct objects, to be finally returned back to Entity classes.

Hence, Persistent Manager defines a set of API to deal with data access with database, hiding the implementation details. This can be seen in Figure 3. The "Entity MetaData" contains meta data about mapping, managed by EntityProfileManager. "SQL Parser" is the parser for sql statements, generating sql statements based on the operations needed by Entity classes. "Accessor" is what's used to execute the sql statement against database.

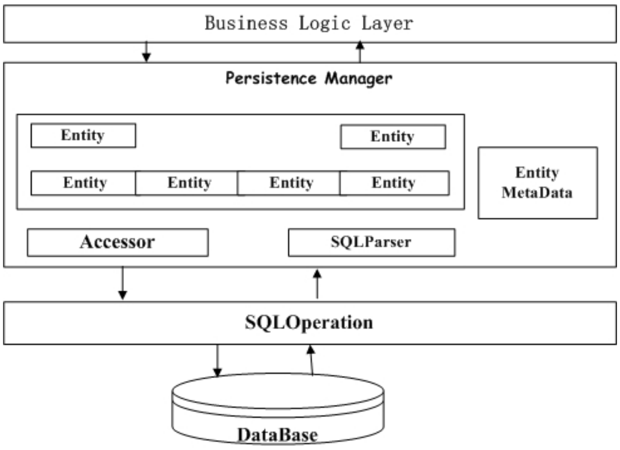


Figure 3. Static Structure of Persistence Manager

Code excepts about how accessing database is done.

```
SqlPersistenceManager class
      //mapping information manager, managing meta data.
_profileManager = profileManager;
      //getting database connection
_connection= dataSource.getConnection();
      //creating generic data accessor
_accessor= new Accessor( _connection );
      //getting sql parser.
SQLParser parser = getSQLParser();
      //create sql parser
 state = new SQLState();
      //set the read operation
state.setName( name );
      //generating sql statement.
state.setSql( parser.forLoad( profile ) );
public boolean queryEntity( Entity entity, SQLState state )
throws CommonException
{ ………..
   setStatement( state.getSql() ); //setting sql statement.
   …………
   accessor.executeQuery();    //execute the sql query
   accessor.first(); //cursor moves to the first record
   accessor.get( this );
   aceessor.close();       //close the cursor.
}
```

3.3 Run-time Model

3.3.1 Data Access Object Caching

When implementing, SQLOperation object is what accomplish communicating with underlying database. During runtime, object creation and deletion are very expensive. Especially for overloaded objects like SQLOperation, if requests from client side come frequently, the creation and deletion of SQLOperation are costly from space and time perspective. Due to this reason, we can design a caching pool mechanism to minimize such penalties. This is depicted in Figure 4:

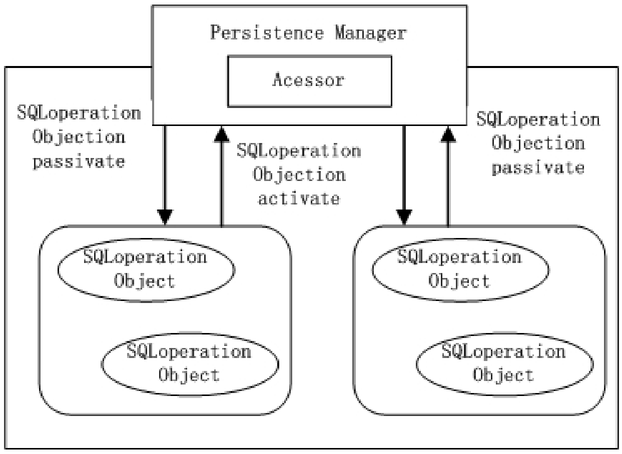


Figure 4. SQLOperation Object Buffer Pools

3.3.2 Data Access Object Dynamic Allocation

From object-oriented design, SQLOperation class belongs to Persistent Manager class; therefore it is bonded with implementation details of communicating with underlying storage. If these details are changed, all related Accessor classes would be impacted. Also it's hard to expand SQLOperation to support different underlying storage. If we isolate SQLOperation class, we can provide multiple implementations for Accessor classes (depicted in Figure 5). This way, SQLOperation objects are dynamically bonded to Accessor object. Thus without modification Accessor objects can communicate with multiple different storage. For example, if we are to store data into an XML file instead of database; all we need to do is to provide a SQLOperation implementation for XML, without touching Accessor classes.

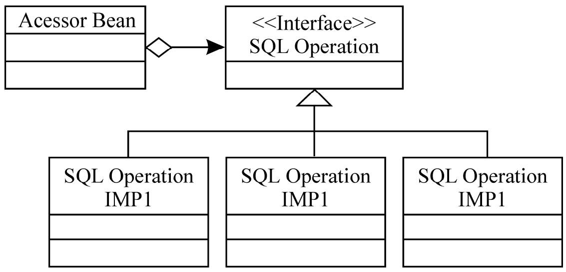


Figure 5. SQLOperation Multiple Realization

During runtime, SQLOperation objects are created by its factory class. When Entity or Accessor objects needs data persistent operation, they would first request a SQLOperation object from SQLOperation Factory class, which in turn would activate an appropriate SQLOperation object from object buffer pool according to the request type. If the buffer pool is out of available objects, SQLOperation Factory would create a new SQLOperation object, activate this instance, then return it back to caller. SQLOperation Factory would also monitor SQLOperation object's status. Once the requestor has finished using the SQLOperation instance, SQLOperation Factory would passivate this instance and put it back to the buffer object pool.

```
 Class SQLOperation Factory
   { public SQLOperation get Operation( String sql ) throws
    Exception
     {
        if( sql is a inquery statement )
           return new SQLQuery();
        else if( sql is an update statement )
           return new SQLUpdate();
```

```
      . . . . . .
      else if( sql is XML query statement)
        return new SQLXML();
      else throw new Exception( “operation not supported” );
    }
  }
```

From description above we can see the runtime Persistent Manager consists of:

SQLOperation: the implementer of communicating with underlying persistent storage.

Buffer pool: a pool containing SQLOperation object instances; different kinds of SQLOperation object have their own buffer pools.

SQLOperation Factory: manager of SQLOperation objects, it manages the creation, deletion, allocation and deallocation.

4. CONCLUSIONS

Data Access Layer is the foundation of an application system, its performance directly impacts the whole system's running efficiency. This article proposes and implements a J2EE based persistent layer with relative complete functionalities. This solution relieves application developers from worrying about details of communicating with underlying persistent data storage, allowing them to focus on business logic design. In reality there will be different questions and challenges in terms of data access layer requirements. The solution in this article has established a simple yet efficient foundation for expansion. Possible future enhancements can be focused on the efficiency of Data Access Layer, especially for the SQL Operator class and the related topics like how to configure buffer pool, re-engineering existing code, and simplify management.

REFERENCES

[1] Ed Roman, Mastering Enterprise JavaBeans and the Java 2 Platform, Enterprise Edition, USA: John Wiley & Sons Inc. pub, 1999.
[2] Erich Gamma, Richard Helm, Ralph Johson, etal. Design patterns: Elements of reusable object-oriented software, New York: Addison Wesley, 1995, p137~139.
[3] Patrick Chan, Java developers Almanac 1.4 volume 1 examples and quick reference, New York: Pearson Education In. pub, 2002.
[4] Jushua Bloch, Effective Java programming language guicle, New York: Pearson Education Inc. pub, 2003.
[5] Wolfgang Keller, Jens Coldewey, Relational Database Access Layers: A PatternLanguage, in Collected Papers from the PloP’96 and EuroPLoP’96 Conference, Washington University, Department of Computer Science, Technical Report WUCS 97-07, February 1997.
[6] IBM Corporation, Object Extender for VisualAge,1998. URL:http://www.software.ibm.com/ad/smalltalk/about/persfact.html.
[7] Wolfgang Keller, Jens Coldewey, Accessing Relational Databases: A Pattern Language, New York: Addison-Wesley,1998.
[8] Wolfgang Keller. “Object/Relational access Layers - A Roadmap, Missing Links and More Patterns”,Submitted to EPLoP ’98.

THE RESEARCH OF GIRD RESOURCE DISCOVERY BASED ON P2P

Qing Wang
Department of Computer Science and Technology, Wuhan University of Technology
Wuhan, Hubei, China, 430063
Email: wangqingniuniu@126.com

ABSTRACT

The current resource discovery model have achieved resource discovery in grid environment at a certain extent, but there is a common problem, that is low efficiency, which lowers the performance of grid. Therefore it is needed to study a new resource discovery model with high performance and efficiency to enhance the grid resource sharing degree.The research of resource discovery mechanism in this paper combines grid and P2P technology, not only considers characteristics of grid resources, but also takes advantage of the structure characteristics and high efficiency resource search technology of the P2P network, discusses the difference between grid and P2P technology and the advantages of their combination, and improves the efficiency of resource discovery, thus realizes the grid resource sharing better.

Keywords: Grid, Resource Discovery, P2P Technology

1. INTRODUCTION

The grid, as a new generation of Internet-based infrastructure, has become a research highlight in the area of distributed systems in recent years. In the late 90s[2], although the network bandwidth grew doubling and the performance of PC grew rapidly in accordance with Moore's Law in speed and processing ability, but the increased computing power has not been fully utilized, and the idle link bandwidth has been wasted. So people began to realize that we can place server on an independent PC, which caused P2P technology come into vogue. Grid and P2P networks have the same goal of sharing resources, and the key to achieve a high degree of sharing resources is to find resources.

2. INTRODUCTION OF GRID

I.Foster, the expert in grid, present a descriptive definition of grid: "Grid, a group of emerging technology built on the Internet, integrating the high-speed Internet, computer, large database, sensor, remote equipment and so on，provides more resources, functions and services for scientists and technicians and ordinary people. While the Internet mainly provides people with email, web browsing and other communication service, the grid can provide more and stronger functions, such as sharing computation resources, storage resources and other resources." Along with the rapid development and application of Internet technology, the grid technology becomes a highlight in recent years.

At present, grid resource discovery mechanism can be divided into two types: centralized type and distributed type. Traditional centralized resource discovery mechanism relies on a central server (node) to discover resources. DNS service is a typical centralized resources discovery process, and Globus's MDS is another instance in principle. Centralized resource discovery mechanism can inquiry efficiently and transmit more responding message through network, but the central server only can provide services to limited clients, lack of scalability, and with the problem of single point failure, so the centralized resource discovery mechanism is not effective in dynamic and wide-range distributed environment.

3. P2P TECHNOLOGY

P2P, or Peer to Peer, known as peer-to-peer connections or peer-to-peer network. IBM gives P2P a definition as following: P2P system consists of a number of collaboration computers, and at least has one of the following characteristics: the system depends on the Marginalization (non-central server) equipments' active collaboration, where each peer benefits from other peers but the participation of the server; Each peer plays the role of server and client at the same time; P2P makes it possible to directly connect to other computer users and exchange files, rather than to connect to the server to browse and download as in the past. Simply put, P2P links people directly, allowing people to directly interact through the Internet, making the network communication, sharing and interaction easier and more directly; Meanwhile, P2P technology has changed the "content" location from "center" to "margine", meaning that the content is no longer stored in the main server, but in a variety of terminal nodes including the PC.

4. THE ADVANTAGES OF P2P AND Grid COMBINED RESOURCE DISCOVERY ALGORITHM

Traditional grid uses centralized management mechanism, which controls information only in one place, so it is not needed to worry about synchronization issues; However, with the massive increase of the number of nodes, the shortcomings of the centralized management turns out to be more and more serious: If the server fails, the system would collapse --a single point failure; without well expandability .And the most important point is that centralized management can not meet the grid's dynamics, and too much inquiries will lead to the failure of client-server mode.

There is difference between the traditional grid and P2P technology at motivation, applications, and resources. At first, the aim of grid was designed to enhance the computing ability, while P2P was to exchange documents directly; Now the application of the grid is expanding gradually, from cutting-edge scientific research to daily life, and the P2P can also effectively balance the load , efficiently search, fully use bandwidth; Grid provides all kinds of services to tens of thousands of users, but P2P can only provides limited special services.

Nevertheless, the network structure characteristics of P2P technology give it a great advantage in the location of resources search. And grid resources can be shared among different organizations and individuals. Many organizations have their own share strategies, so that the middle nodes are only responsible for forwarding the request, without the right to decide the use permission of resources, so the end-to-end structure is comparatively favorable. The most important thing is that P2P and Grid have the same ultimate purpose: collect the large number of network resources appeared in different organizations to achieve share. So these two technologies can be effectively combined.

5. REALIZATION OF THE COMBINATION OF GRID AND P2P

Based on P2P search mechanism, we introduce the idea of super-node in P2P search model to the grid system,and the concept of Super-Peer[10] into grid's supercomputer and common node. Selection of Super-Peer can be based on geographic location or some high-performance computer among common nodes. A P2P middle layer between Super-peers is constituted, the network model is as Figure 1.

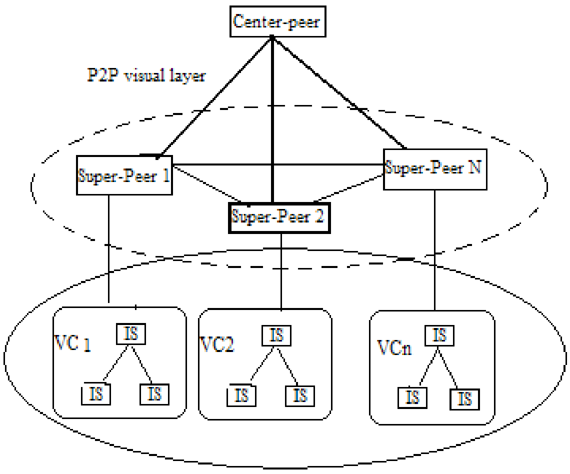


Figure 1

Notes: (1) VC (virtual commonwealth) is a group of independent elements which share the same interest and comply with common principles. In order to support the completion of the task, a virtual community need to provide certain facility for resource information management and corresponding resource release and discovery mechanism. Informations, as well as the resources released and found mechanism.(2) IS (information service) information service nodes gather resource information in different VC, thus can access corresponding VC information through those nodes.

Grid — P2P model integrates the advantages of P2P's unstructured model and Grid-hierarchy resource management. In a Grid-P2P model, each common peer subordinates to a Super-peer, that is in a Super-peer domain, all the other peers are connected to that Super-peer. Low bottom peers need to login information on Super-peer. The peer can be a resource provider or a resources requester. We establish P2P network in the second layer of Super-peer. Super-peer need to register in Center-peer, and it forms neighborhood relationship with other peers in the same layer under certain algorithm.

In P2P network, nodes often join or leave dynamically. In order to avoid that the system's stability and reliability is influenced by dynamic changes of Super-peer, we can consider define Super-peers by location division[4], such as designating a high-performance server as Super-peer node in some regions, and the other computers within these regions as common peer. At the highest level, that is Center-peer layer, an index adjacent table of super-peer is established, in which resources contained in that layer be arrayed reversely so that the discovery can start query from the supper-peer containing most resources.

6 . MODEL CONSTRUCTION

6.1 Interest Measure

One feature of this model is the proposal of introduction of intrest measure[11] from artificial intelligence to resource discovery .If a user requests grid computing resources, this intrest measure will be defined by the number of CPU, basic frequency and memory size; If a grid storage resources are requested, this intrest measure will be defined by the size of effective storage space; if grid document resources are requested , the intrest measure will be defined by the amount and type of documents.

Resources with similar interest are dynamically linked together through self-organizing algorithm dynamically. The main purpose of linking nodes according to interest measure is to establish nodes with the same interest, keeping high interest between nodes, which can reduce the searching range when users request resources, and reduce the amount of redundant messages.

Suppose a query node is p, whose neighbor node is q, and the nodes adjacent to q but p is S. In terms of node p, q is adjacent to p, S is an indirect adjacent node to p, shown as Figure 2:

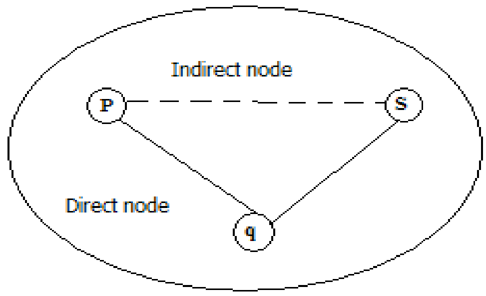


Figure 2

Node p generates a resource request $Query_q(p)$, and sends the query to each of its adjacent points q, and q will send this query to its adjacent nodes S .

Definition 1: $QueryHits_p(q)$ is the number of hit messages arrived at p after query, returned from adjacent node q

Definition 2: $QueryHits_{p,q}(s)$is the number of hit messages arrived at p after query, returned from indirect adjacent node S across q

Definition 3: $RateQueryHits_p(q)$ is the proportion of hit query number p received from q in total hit query number.

$$RateQueryHits_p(q)= QueryHits_p(q)/\sum_r QueryHits_p(r) \qquad (1)$$

r is an arbitrary node of the system.

Definition 4: $RateQueryHits_p(s)$ notes the proportion of hit query number p received from S in total hit query number.

$$RateQueryHits_p(s)= QueryHits_p(s) / \sum_r \sum_q QueryHits_{p,q}(r) \qquad (2)$$

Definition 5: $NumConn_p$ is the number of links from p;MaxConn is the largest number of links from p.Through the Formula (1) and Formula (2) the query hit rate p received from adjacent node q and indirect adjacent node S. p decides whether connect directly to S or disconnect to q. If RateQueryHits p (s) $\geqslant$ RateQueryHits p (q), p sends invitation of connection to St. At this time it is also needed to judge NumConnp, if it exceeded MaxConn, select the smallest link in RateQueryHits p (q)to disconnect.

The scope of NumConnp should be considered is based on the following factors:

(l) Reduce expenses of dynamic reconfiguration of Supper - Peer layer. Grid resource nodes reconfigure with dynamic self-reorganizing algorithm through mutual commusnication, so there exist communication cost of potential re-connected node;

(2) It is good for Suppe-peer layer to reach steady status at a faster rate.

(3) Reduce the resource search space. Resource request is only sent to neighbor node with high "interest messure", which reduces the searching space of resource discovery and improves the efficiency.

Hence, how to determine the size and scope of NumConnp? Reference [13] gives dividing principle of NumConn, which notes the number of nodes in the resource region as X, with the range of $K \leqslant X \leqslant 3K-1$, where K is a constant, and gives qualitative analysis of such division.

6.2 Resources Construction in P2P Layer

6.2.1 Join and Left of Resource Nodes Domain among Domains

In fact, in the P2P grid layer, a phenomenon should be noted that the establishment of the overlay network has experienced the process from small to large, in which the information server must have an order by which they join the P2P layer.Based on this idea, organize the successive nodes depending on the early joining node, so that they can create or join group of proper resource type. When a new information server needs to join in, it will get touch to one or more known active information servers and send Join-in message to them. IS receiving that message continue forwarding the message to its lower neighbor nodes and returning response to the sending node. The response message contains the resource type which it owns and the information of adjacent nodes it maintains. The new joined IS gets information about other nodes from the returning message and fixes its adjacent nodes to create or join in the correct group. Besides, IS can caches the join-in point of resource type not including in IS according to the returning message.

Nodes which has established neighborhood relationship take soft state protocol,they send message"I'm alive" periodically to determine the existence of neighbour relationship, if they find that some neighbors failed or left, they will use the above method to find their necessary neighbors through self-organized manner

In order to meet the needs of join and leave the entry point of the message group ,the central server CS also take soft state protocol.If CS can not receive "I'm alive" message in a certain period ,it will delete the entry Point of information that have preserved.

6.2.2 The Addition and Left of Resources Node within the Domain

When a node needs to join in the grid, it chooses the nearest domain server, and then it can regist the shared origin resources to the domain information server. If new resources make this domain an increase of one or several new types of resources, then the IS needs to adjust division to ensure the accuracy.

Specific algorithm is as follows:

(1) For every resource type which has newly registered on the IS, if IS has such resources on the origin, it will register such resources to the IS's RI directly. There is no need to do other changes.

(2) If the IS did not has such resources originally, the IS needs to join this group. If this IS has cached entrance node t, the IS can join this group successfully.

(3) If the IS does not cache entry point of this type of group, it needs to use the P2P layer's nodes join algorithm to find the t type of group, then join the group successfully.

When a node wants to exit the domain, the resources that it has registered on the domain server IS will be invalid . If the withdrawal of resources makes the original resources of the IS type lack one or more, then the domain server need to withdraw from one or more groups. Withdrawal only means that the server no longer sends "I'm alive" message to adjacent nodes within the group.

7. ALGORITHM DESCRIPTION

In Grid-P2P model, the resource discovery algorithm in Super-peer layer and between peer layers[12] can be generally Described as follows:

(l) When a peer node submits inquiries, first of all, send query to Super-peer. Super-peer node searches in its current information. If the satisfying resources be found, return to the peer, and algorithm end;

(2) If Super-peer node do not find the resources in their own information to meet the requirements, send request to other Super-peer in accordance with the "interest measure";

(3) Super-peer nodes receives the request searches in its current information , if finds the resources to meet the requirements, return to the requesting Super-peer node, then the Super-peer node sends the request forward the information back to the peer, and algorithm ends.

The algorithm description of grid resource discovery based on P2P is given in pseudo-code as follows:

Algorithm 1: Peer node send the request to the Super-Peer nodes

```
When a Peer sent a seareh message to Super一Peer
{
 If (Super-Peer founds resources)
   return to peer
 else call search algorithm of Super-peer layer
}
```

Algorithm 2: Resources search among Super-peers

```
When a SuPer-Peer sents a seareh message to another
Super-Peer in accordance with the "interest measure"
 {
   search index adjacent table
   if (Super-Peer founds resources)
     return to the requesting Super-Peer
   else if (Super-Peer founds no resource)
     continue to call Algorithm 2
 }
```

8. PERFORMANCE ANALYSIS of Grid-P2P MODEL

Assuming there are n Super-Peer in the grid, that is, n super nodes, and further, assuming the i^{th} Super-Peer has N ordinary nodes, in which the probability of the existence of the required resources is the same, then it is clear that time complexity consists of two parts:

(l) Sort the information in the index adjacent table using the optimal sequencing method. The average time complexity and the worst time complexity are both O ( );

(2) Traverse the selected adjacent table of Super-peer. If using serial search method, the worst time complexity is 0 (N), and the average time complexity is 0 ( ). Hence, the

average time complexity of the Resource Discovery Algorithm is 0 ( ) +0 (1).

Under the best circumstances, only one Super-peer index adjacent table needs to be traversed after the resource requirements is submitted, and one search can meet the needs. As a result, O (O ( ) + O (1)) = O ( )。

Under the worst case, all the Super-Peers in the grid have to be searched to find the resource. In this case, the time needed is the sum of sequencing time, adjacent tables traversing time and locating time in super-peers, thus the time complexity is 0 (0 ( ) + m * 0 (N))) 。

9. CONCLUSIONS

This paper proposes a grid resource discovery model, typical of using interest measure and information service, based on P2P by introducing the resource discovery algorithm, grid resource management technique and P2P. Discuss the advantage of combining network with the emerging P2P. However, model design is a complicated and huge problem, restricted by time and practical environment, this paper just gives a preliminary analysis of the combination of network and P2P, some functions have not come into reality, and some are still in theory to be further researched and improved.

REFERENCES

[1] Reynolds R G,Michalewica Z,Cavaretta M.Using Cultural Algorithms for Constraint Handling in Genocop[C] // McDon-nell J R,Reynolds R G,Fogel D B.Proceedings of the 4thAn-nual Conference on Evolutionary Programming. Cambridge, Massachusetts: MIT Press,1995:289-305.

[2] GLAS G, CAO P. Adaptive page replacement based on memory re-ference behavior[C] //Proc of ACM SIGMETRICS Conference on Measuring and Modeling of Computer Systems. Seattle: ACM, 1997:115-126.

[3] Rychtyckyj N,Reynolds R G.Using Cultural Algorithms to Improve Performance in Semantic Networks[C]// Proceedings of 1999 IEEE Congress on Evolutionary Computation. Washington D.C:[s.n.],1999:1651-1656.

[4] Foster I,Kesselman C,Tuecke S. The anatomy of the grid:Enabling scalable virtual organizations[J]. International J.Supercomputer Applications,2001,15(3):1-10.

[5] Wu Yongwei, Wu Song, Yu Huashan, et al. CGSP:an Extensible and Reconfigurable Grid Framework,Lecture Notes in Computer Science, 2005, 3 756:292-300.

[6] M.Ripeanu.Peer-to-Peer Architecture Case Study: Gnutella.in:Proceedings of International Conference on P2P Computing.Skyways,Sweden:IEEE,2002.99-100

[7] Frey J, Tanenbaum T, Livny M, et al. Condor-G: A Com-putation Management Agent for Multiinstitutional Grids[J]. Cluster Computing, 2002, 5(3):237-246.

[8] BENERS-LEE T, HALL W, HENDLER J, et al. Creating a science of Web[J]. Science,2006,313(11):769-771.

[9] Iamnitchi A,Foster I.On fully decentralized resource discovery in grid environments[C].Proceedings of International Workshop on Grid Computing , Denver, Colorado, 2001,51-62.

[10] LüQ,Cao P,Cohen E,et al.Search and Replicating in Un-structured Peer-to-peer Networks[C]//Proc. of 16th An-nual ACM Int. Conf. on Supercomputing (ISC’ 02).NewYork, USA:[s.n.],2002.

[11] JON K Leinberg,The Small-World phenomenon;An algorithmic perspective,ACM Symp on Yheory of Computing,2000,28(4),820-828.

[12] Jiang Haitao,Prof.Yang changxing.Rearch of Grid resource Discovery Based on P2P Mechanism,The full text of China's dissertation database ,2008.2.

[13] Le optical, LI Ren-fa, Zude. P2P-based calculation of multi-layer structure Region network model〕 J Journal of Software, 2005, 16 (6) :1140-1150

RESEARCH ON NETWORK EDUCATION RESOURCE DATABASE CONSTRUCTION BASED ON INTELLIGENCE RECOMMENDATION

Jian Liu
Hubei University of Technology
Wuhan, 430068City, P.R.China
Email: liujian1965715@163.com

ABSTRACT

With the development and combination of web and multimedia, the network education and training becomes more and more popular because of its convenience and novelty. Existing network education production are based on traditional exam-oriented in our country, and it's difficult for the user to find out the information efficiently from massive resource on the internet. In order to change this confusion and find suitable network education mode for our country, we construct a network education resource database based on intelligence recommendation. The network structure, framework of the system and the user and authority function relationship are suggested, and the principle of intelligence recommendation system is illustrated. In our opinion, this kind of network education system can solve many existing problems and will fit for the current situation of China.

Keywords: Education, Network, Intelligence Recommendation, Resource Database.

1. INTRODUCTION

With the fast development of network and multimedia technology, digitalization and network of mass education resource and construction of the network education resource database system become necessary. At present, there are a lot of products from each company and college, but still can't satisfy the need of the society because the system structure, function design and the mode that the developer provides the resource directly, bring many problems in reality. For example, the content of these production pay more attention to the quantity instead of the quality of the resource, the frame work based on the traditional exam-oriented education, and the functions of these system are just accumulated simply. Besides, each standard lacks unifying normative constraint, and structure planning has great diffidence. All these reduce the resource database construction efficiency, and cause the resource island phenomenon more and more serious. Therefore, the design and realization of the rational normative practical Network Education Resource Database (NERD) become very important.

2. NERD SYSTEM FUNCTIONAL REQUIREMENT ANALYSIS

2.1. Necessity Analysis

The speed of the development of network and multimedia technology amaze us all along, which make us able to extend our communication way a lot by video, audio, and human-computer interaction. The hardware of the computer improves fast; the storage, compression, searching and rendering technology become mature, and the multimedia technology become a study focus. These provide solid technical support for the NERD. So it's necessary to make the massive video and audio education resource digital, networked and construct the network resource management platform so as to guarantee its security and carry out the normative, scientific management. The NERD can improve the server function of the education resource, and becomes the necessary foundation to develop network information education.

The video and audio education resource has a great data amount, its strong real-time characteristic require high data transfer rate and bandwidth of the multimedia education network information resource system. The content of the multimedia information resource mainly include the video and audio information flow, multimedia courseware and material which contain picture, sound, cartoon, video, text, etc. These information resources can be compressed to standard information which can be transferred on the internet. Because this kind of media is different with the traditional text information, the video and audio education resource stress on the vision effect, and its storage is enormous, it's difficult to realize the high quality and fast play on the internet. Fortunately, with the development of the general application of broadband technology and continuously improvement of the decompressing technology, it's possible for us to view the video on the internet. However the video quality still cannot satisfy our requirement, we can solve the problem that the video always intermit while playing by streaming media technology. So these technologies provide foundation for constructing a network resource platform to serve the education and improve the innovative education based on network resource.

2.2. System Characteristic

(1) Integration. The construction of the NERD system should follow certain software development system, and each function of the system should be able to be combined closely.

(2) Standardization. The system should follow certain development standard in the resource structure and database attribute, to make it convenient for system expanding and for other system visiting.

(3) Compatibility. The distributed system deploying should be able to run on many kinds of systematic platforms.

(4) Flexibility. The system should be able to release resource according to different levels of user's authority. This flexibility can reduce the complexity of resource management and the work of the system administrator. Besides, it also helps to carry out the level management and distributed construction of the resource.

(5) Security. The information resource in the system should be observed, managed during the whole life cycle from production, storage, use and stop. Likewise, the relevant information with resource such as author, provider, person who verify, user and so on should be recorded and managed.

(6) Convenient. The system should be able to let users add the resource information they collect to the database, and release these information according to the verifying of the authority in the system. This makes the system to expand

itself.

2.3. Intelligence Recommendation Technology

Because the education resource always be multi-specialty, inter-disciplinary and very huge, it's important to select suitable knowledge resource for learner. So, if the NERD system has the knowledge navigation function, the efficiency can be improved a lot. As the Figure 1 shown below, we construct the framework of the intelligence recommendation function of the NERD system. The ontology services layer connects the user interface layer and accepts request from user. The user give the information searching server a searching request, then the server will mapping the searching information to its ontology knowledge database cluster after the data mining ontology service and knowledge integration ontology service. After that, the intelligence recommendation system gets relevant knowledge and association from its knowledge database, and then finds out each field relevant with the keywords from disciplines and specialties knowledge ontology database, and calculated out the information which the learner and most other learner care about. The intelligent searching engine in the knowledge grid layer is a network information intelligent acquiring and processing system, which is a integrate solution for acquire and process the information. This system supports the distributed network information parallel searching and content filtering. Furthermore, the intelligent recommendation system can identify text sort and understand the text information from concept level including text classification, cluster and concept abstracting. Then it can utilize the efficient unify text in ontology space of the ontology service layer to manage the information. At last, it provides high efficient text information searching based on ontology and image searching of the content, and personalized special information service. Thus, the NERD system can realize the disciplines and specialties intelligence recommendation expert navigation system during the education resource learning.

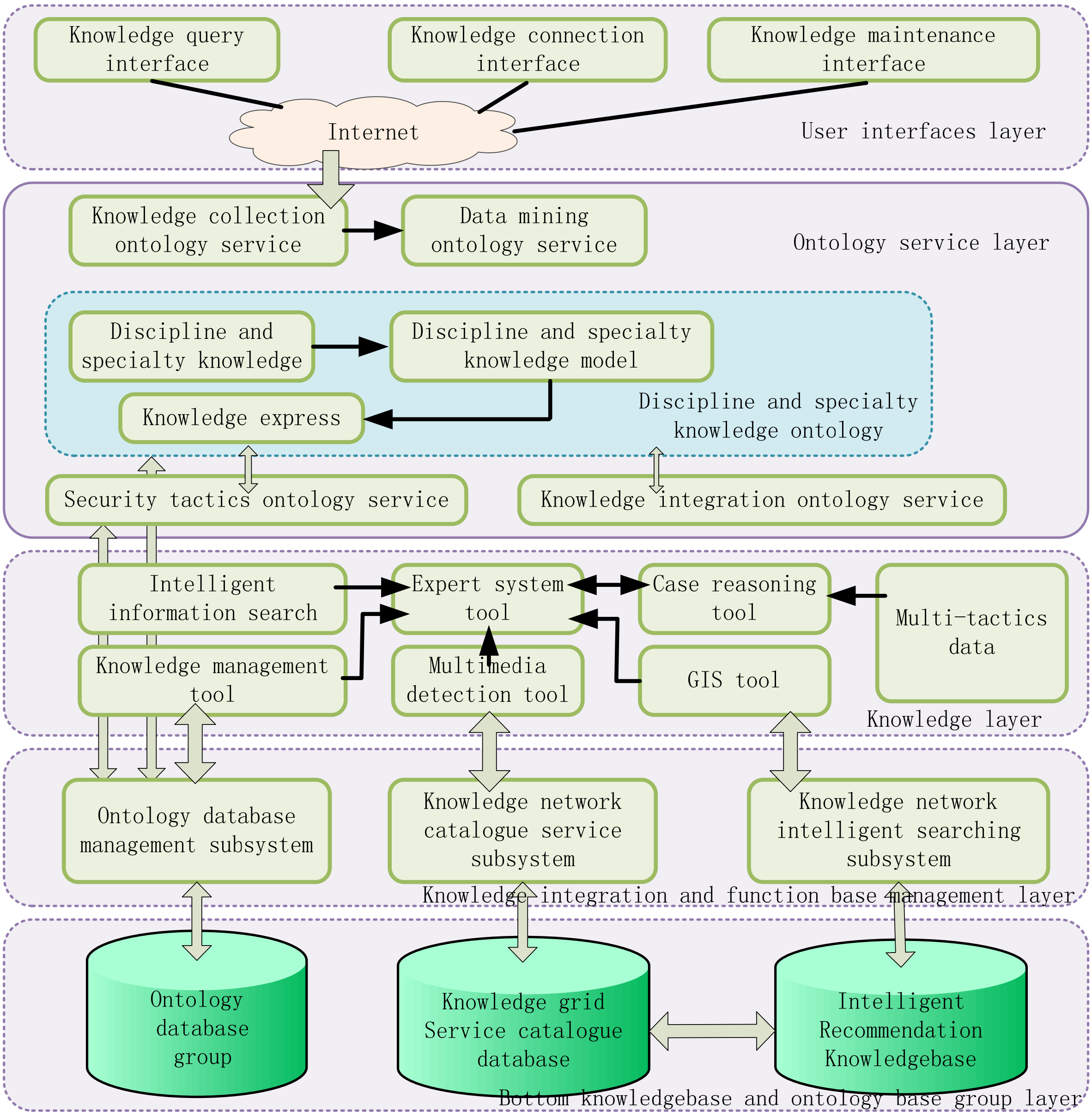


Figure 1. Framework of The NERD Intelligence Recommendation Knowledgebase System

The intelligence recommendation system of the NERD is available on two aspects. One is education resource recommendation. While the learner hopes to learn certain knowledge, the resource database can recommend relevant resource actively. Besides, the system will recommend some other knowledge relevant to what the learner is interesting in, in order to help the learner understand profoundly. The other is learning method suggestion. The system provides relevant learning method and actively guidance for the learner, and some advice about how to do or what to do in next step.

The intelligence recommendation system is important part of the NERD system, which help the learner search the education resource accurately and fast. By this technology, it's possible for the learner choose the education resource from massive information on the internet.

The customer interface layer consists of knowledge query interface, knowledge dispatch interface and knowledge maintenance interface. Customer interface layer offers some function of advanced navigating, semantic explaining, semantic reasoning, man-computer communication to users. According with the standard of the Natural User Interface (NUI), it offers not only the humanistic and semantic guides service to users, but also the friendly human-computer interface to experts of discipline and specialty, knowledge engineer and ordinary user to help to apply, manage and expend the knowledge. Meanwhile, knowledge dispatch interface and knowledge maintenance interface can define the new interface, meet more advanced and specialized application program, and serve the ontology services layer directly. The ontology services layer and the knowledge grid layer are the key parts of this framework. The ontology services layer is charge for the management and query of ontology and the maintenance and control for ontology database and ontology buffer base. It offers the ontology services of data integration, data mining and knowledge integration. The knowledge grid layer can provide the operation of Registry and Discovery of data source, Federated Access, Federated Management, Collaboration effectively. Knowledge integration layer and function base management layer, which deal with visiting knowledge, mapping semantic and finding resources and matched services, play the important role in continuing the second layer, the third layer and the fifth layer.

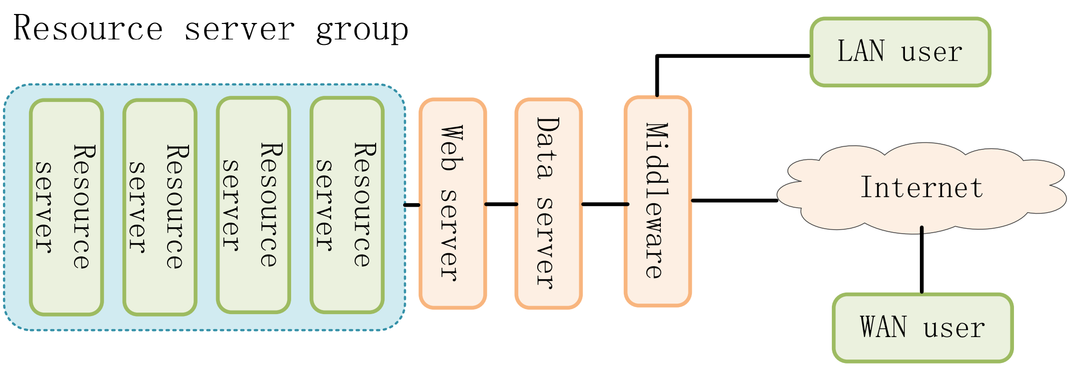


Figure 2. Network Structure of the NERD System

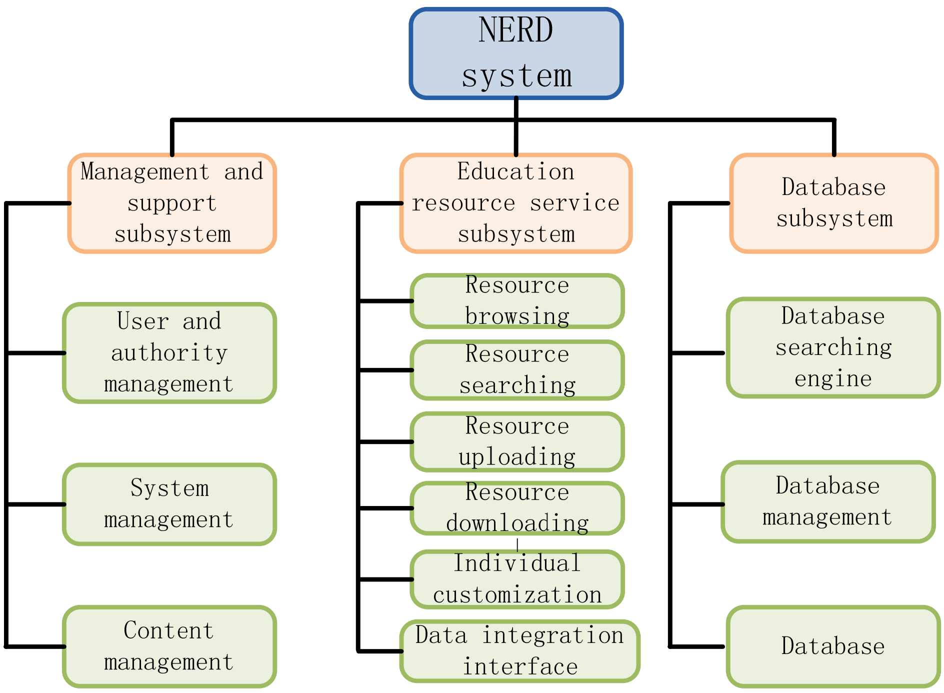


Figure 3. Framework of the Subsystem of NERD

3. FRAMEWORK DESIGN OF THE NERD SYSTEM

3.1. Network Structure

The network structure of the NERD system is shown as the Figure 2 below.

From the Figure 2, we can see the portal server provides access for the user, the data server store the education resource and user information, and the resource server, which consist of several servers and storage devices, store the resource content distributed.

The traditional two layers Client/Server frame fits for running in small-scale, less-user single-database, safe and fast environment such as LAN. With the popularization of the network technology and the increase of the users, the complexity increase too. In large-scale, massive-use multi-database, and unsafe environment such as Internet, the three layers frame is developed to meet the needs. The NERD uses the three layers mode, which can be divided into three subsystems including management and support system, education resource service system and database system. The subsystem framework is shown in Figure 3.

3.2. System User Role Analysis

Seeing from management aspect, the education resource database includes resource management, user management and statistical analysis; From using aspect, it includes browsing, searching, uploading, downloading, commenting of the resource and so on; Form security aspect, it includes security of system running, security of information system, security of information transmitting and the security of the information content.

Generally speaking, the user role of the NERD system can be divided into three levels including super administrator, ordinary administrator and ordinary user. Accordingly, in some subsystem of NERD, the user role includes administrator, teacher and student. Each role and its function relationship are shown in Figure 4

The super administrator possesses highest authority in the whole system which almost covers every basic function provided in the system. Its main task is the normal maintenance and configuration of the system, statistical analysis of the whole using situation, and the ordinary administrator management.

The authority of the ordinary administrator is granted by the super administrator and is second only to the super administrator. Its main task is resource management such as the adding, deleting, revising and verifying of the resource, etc), and the ordinary user management.

The ordinary user is the final user and beneficiary, biggest user group and the motive force of the development of the system. In order to bring convenience for the user to use the resource, the system should provide some function for the ordinary user, such as browsing, searching, downloading, etc. In addition, in order to make the interacting communication among administrator and user easy, and user can expand the resource conveniently; the system should provide resource evaluation and uploading functions.

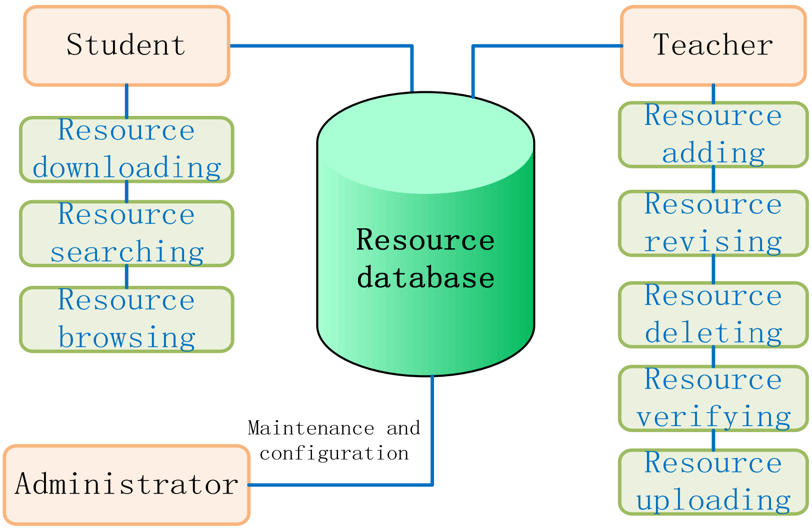


Figure 4. User And Authority Functional Relationship

3.3. System Function Module Design Analysis

Database system function

- Education resource classifying management
- Education resource verifying. It includes the detailed verifying, batch verifying and multiple delayed verifying functions.
- Education resource releasing. For verifying passed resources, the platform realizes different media types real-time dynamically releasing and upgrading.
- Education resource using tracking.
- Education resource filing
- Education resource recycle bin. It includes the deleting and recovery of the resource so as to prevent the mistake operation of the administrators.

Education resource service system function

- Education resource searching
- Education resource browsing
- Education resource uploading
- Education resource downloading and watching online
- Education resource individual customization

The NERD management system mainly consists of three function including resource inquiring, resource uploading and downloading function. Its concrete function includes the inquiring, recording, revising function to all kinds of resource.

The material searching engine should include the functions of boolean query, association query, accurate query and fuzzy query. The system should have the material remote submitting function, so the user can submit the material by the internet. Moreover, the relevant material recommendation function makes the user able to know the relevant material; the commenting function help the user make comment to any material and see the others' comment. And the system should provide the content transferring management function, and can support the uploading and downloading of each kinds of multimedia and so on.

4. CONCLUSION

The development of NERD system is important to the network education in China. This paper construct the a framework of NERD system, which realizes the functions such as uploading, downloading, browsing, searching and so on. Besides, it realizes the management of the user and resource, such as register, revising, verifying. It utilizes the intelligence recommendation to find out useful information efficiently from massive resource on internet. However, we still need to expand our study on systematic theory of NERD, consider more integrative function system, and improve the interacting communication module of the system. And this system rains to be tested in the practice. We will carry on further researching on this subject so as to improve the network education in the future.

5. ACKNOWLEDGEMENT

We are grateful for the data resources provided by Library of Wubei University of Technology. We thank Professor Yang Qifeng for framework guidance.

REFERENCES

[1] Gomez-Perez. (2006). A Technical report Some ideas and examples to valuate ontologoies.KSL-99-65 Knowledge systems laboratory，Standford University, CA:5～6.

[2] Yang Qifeng, Hou Zhenhua. (2007). Research on Knowledge Resource Base of University Discipline and Specialty.The 13th Cross-Strait Academic Conference on Information Management Development & Relevant Strategy. Englishi Part,P893-899

[3] Hou Zhen hua. (2008). Construction and Application on the Knowledge Resource Base of University Discipline and Specialty. Wuhan University of Technology.

[4] Zhuge H. (2002).A knowledge grid model and platform for global knowledge sharing[J]. Expert System with Applications ,,22 (4) :313～320.

[5] Zhuge H. (2004). Knowledge grid [M]. Singapore: World Scientific Publisher.

[6] Sheng Q J, Shi Z Z. (2004). A knowledge2based data model and query algebra for the next-generation web[J].Lecture Notes in Computer Science , 3007 :489～499.

[7] Xiao L, Fu W.(2003). SDPG: Spatial data processing grid[J].Journal of Computer Science and Technology, 18(4):523～531.

RESEARCH ON INTEGRATION OF DEEP WEB QUERY INTERFACE

Wei Jie Kuai
School of Automation, Guangdong Technology Normal College
Guangzhou, Guangdong 510630, China
Email: kuaiweijie@21cn.com

ABSTRACT

With the rapid development of World Wide Web (WWW), there is an overflow of information on Web could be utilized, and the amount of information is growing quickly. To utilize these we need to access the backstage Web database through query interface. Although the Deep Web contains plentiful information, it is still a challenging task to utilize them efficiently because of their isomerism and dynamics. Integrating Deep Web dataset is still a rising research area up to now, which contains several issues must be resolved. Totally speaking, numerous research works have been done in this area, but the developments of each aspect are not balanced. In this paper the current research situations of several crucial issues to integrate Deep Web query interface are reviewed and summarized, and the direction of intending research is thoroughly investigated and analyzed.

Keywords: Deep Web, Query Interface, Integration.

1. INTRODUCTION

With the rapid development of the World Wide Web, there is an overflow of information on Web could be utilized. According to the up-to-date investigation of [1], the amount of information of the whole Web has surpassed 200,000TB, and the number is growing quickly. Researches of Web aim at developing new techniques to gain useful information from Web efficiently. The information in Web is issued mainly through web page, which composed by texts and hyperlinks has its unique: astonishing amount, rich information; developed by different person or team with huge differences in form and content; distributed on everywhere of the earth where connected with Internet which make the Web data heterogeneity and un-constructed. With all the above, it becomes a challenging task to obtain the valuable information and data automatically. Up to now, in order to utilize the Web information efficiently, there are many approaches developed which involve extensive research fields such as data mining, machine learning, natural language processing, statistical analysis, database and information indexing.

The whole Web seems disorderly and unsystematic, but can be divided into Surface Web and Deep Web according to their implicated information. Surface Web is the set of Web pages that could be indexed through hyperlink using the traditional Internet search engine. In fact, there are about 21.3% of pages could not be indexed because the hyperlinks to them lost, which also be viewed as Surface Web in this paper. As to the Deep Web, there is no uniform definition yet, according to [2], Deep Web is defined as the content which could not be indexed by traditional search engine, especially the dynamic pages that are generated online for query. But with the increasing of search engine Crawler's ability, the definition of Deep Web has become complicated and hard to give a definition with a long-term and consistent identification.

Here we try to make a perfect description of Deep Web as: Deep Web is defined as the addressable online database on the web, Web DB or WDB for short here and the content is stored in the real database.

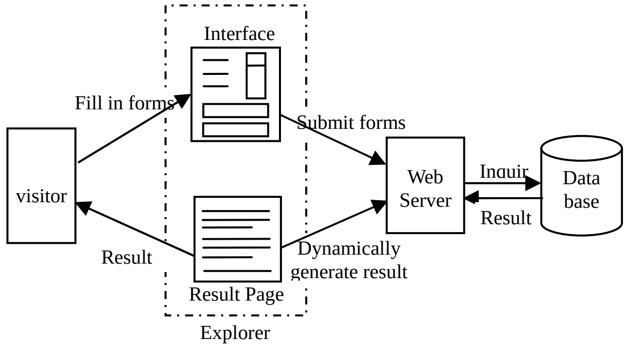


Figure 1. The Process of Getting Data Form Web DB

The contents will be presented to the visitor through the web page dynamically generated by Web server only when they are inquired (Figure 1), so there is no hyperlink connecting to these pages, this is the basic difference between Deep Web and the static page visited directly. According to the structuring degree of stored information, Deep Web could be divided as structured information, document information and non-text information. For example, the information of online shopping web belongs to structured information while the information of news web belongs to document information, and the technical difference for query caused by structuring degree are also huge. The non-text information mainly includes multimedia files, image files, software and some files with special formats (e.g. PDF files). In the ordinary meaning, to gain Deep Web information we pay more attention on the obtaining of structured information, but not document or non-text files, the reason is easy understanding, the integration of structured data is more valuable and the techniques are more plentiful. The integration of Deep Web data is mainly to integrate the structured information. With the of growing up of Web techniques and rapid increasing of information implicated in Deep Web, visiting the Web DB becomes the main way to obtain information, and the research on Deep Web is being gotten more attention.

2. GENERATION OF INTEGRATED QUERY INTERFACE

In order to visit data from several database simultaneously, the Web DB integrated system must provide a uniform access path. Each of the Web DB provides its query interface, so what we should do is to integrate each of the Web DB query interface to a uniform interface, which is named as integrated interface. Submitting query through the integrate interface to achieve the aim of Submitting query through the query interface of several Web DB simultaneously. There are mainly four steps to get the integrated interface. Firstly, find the query interface will be integrated on the Web. Secondly, resolve the interfaces and obtain their model information, i.e. query ability. Thirdly, divide them according to different fields and fourthly, integrate the interfaces belongs to the same field to a uniform interface. Since the 21 century, with the rapidly development of Internet technology, software technology and hardware technology, obtaining useful information automatically from Web is not assume any more. Deep Web has attracted more and more attention, and many related research achievements have been published. The final aim of Deep Web research is obtain and integrate the rich information of Deep Web on the whole Web automatically. Though the whole Deep Web contains almost all information we need, in fact it is difficult to utilize them in a manual way, and the integrating of Deep Web DB is to efficiently utilize information of Web DB as automatically as possible.

Finding the Web DB on the Web and sorting them according to fields, integrating all the query interfaces according to every sort to provide a uniform query interface and make it is possible to submit query to several real query interfaces simultaneously, by which to achieve the aim of visiting several Web DB of the same field simultaneously. There are four sub-modules: finding of Web DB, extracting of query interface model, sorting according to field Web DB and integrating of query interfaces. Finding of Web DB means find a real Web DB website on the Web, and then find the query interface of the Web server. Extracting of query interface model means to analyzing and extracting the attribute of query interface found previously, and decomposing a query interface into a set of attributes. Sorting of Web DB according to field means determine which field does the Web DB belong to according to the obtained attribution of query interface. Integrating of query interfaces is to integrate the query interfaces which belong to the same field to obtain a whole query interface, through which to visit several local query interfaces simultaneously.

When a custom fill and submit query through the integrated query interface, the query are transformed to query of every local query interface, and each of Web server will return their result pages that satisfy the query, from which the query results are extracted and appended semantic comment. Because of the independence and heterogeneity of each Web DB, the formats of result data are different and it is necessary to uniform the data formats into a final format which could be processed automatically. This contains 7 modules as field mapping, Web DB selecting, query assigning, results extracting, results commenting, entity recognizing and results combining. Field mapping is to automatically select the right field according to the custom's query and send the query into the right field integrated query interface. Selecting of Web DB is to select the right subset from all the Web DB of this field, getting the satisfied query results at the same time reducing the cost as much as possible. Query assigning means transforming the query submitted to the query interface into query to the every local query interface of Web DB to be visited. Results extracting is to extract the query results exactly and completely from the query result page and form a storage model for processing at next step. Results commenting are to add semantic description to the extracted results, i.e. adding metadata information. Entity recognizing means to find the data which describe the identical entity of real world from results of different Web data, this step could delete the repeating data of the results and reduce the redundancy of results. Results combining are to transform the results from different Web DB into a uniform expression, and store them as the same format.

3. EXTRACTION OF SCHEMA OF QUERY INTERFACE

The schema of query interface is a set of properties related to a set of neighborhoods, and a query on the Web DB which indicates a query interface is done by evaluating on some of the properties. The query power of the query interface can be obtained by extracting from the schema of the query interface. The schema of query interface can be considered as a view which is built on the corresponding Web DB. The extraction of the schema means the collection and analysis on the properties of the query interface. The main purpose of the extracting from the schema of the query interface is for the next classification of Web DB and integration of query interface, and the key work is to exactly extract each property included in the query interface.

The extraction of the schema of query interface is achieved by using grammar analysis. Some In-depth work has been done on the analysis of the structure of a full page. However, the study of the analysis of the structure of query interface has not been begun. In the method, a hypothesis that each query interface is constructed by hidden grammar is proposed firstly by observation and statistics. In order to accurately identify the compound mode of each element representing the property from a specific query interface, a parse tree is constructed to interpret the whole query interface, definite their grammar rules, and resolve the possible conflict among the patterns of group by using priority. Thus, the properties of the query interface can be listed as far as possible. Obviously, it cannot satisfy the need of practical application fully if the precision is 80% and the recall is 89%.

After extracting the properties, query interface will be represented formally in order to propose the model solution on next step. The expression of the formal query interface is related to the purpose of application. If the purpose is to the classification of the Web DB, the whole information of the query interface will be paid more attention, i.e. the information in which domain can be queried. If the purpose is to the integration of the query interface, the detail of each property of query interface will be paid more attention, which means the best matching relation between different query interfaces will be founded. The most intuitive method is to consider query interface as a set of properties. A kind of more complete formal expression is proposed, the whole query interface is expressed as a triad firstly including the related information about Web site which of the query interface, the set of the properties and relations between the query condition formed by properties, such as conjunction, non-conjunction, repulsion, etc. The set of the properties is the description of the information of each property which is expressed as a septet, including name, position in the query interface, domain type, default value, value type, and value unit of the property. From above we can see all related detail information is included, and enough information is supplied for the next classification of Web DB and integration of query interface.

4. INTEGRATION OF QUERY INTERFACE

Property analysis is the most primary approach for the integration of query interface, and many researches on this have been conducted up to now. In this approach, the information of schema and grammar for the specific query interface is mainly mined, the matching relation between properties of different query interface is identified by using the grammar information, an integrated query interface is obtained on the basis of some of the specific query interfaces, and thus more Web DBS can be accessed simultaneously. Schema matching and integration are the key technologies for achieving this approach and the data union of the following. We will not give more detailed description for that is mainly the application of existing technology. The most popular approach for the integration of query interface presently is by manual approach, which is more exactly. However, the efficiency is difficult to ensure in the case of large scale of integration of query interface. So this will be achieved by automatically. In the past, the automatic integration of query interface can be classed two types, one is local approach, and the other is whole approach. The local approach is based on the given set of query interface which will be integrated, we

analyzing the implied information of the properties especially the grammar information, matching the properties among them, and a new global interface is obtained. The whole approach is based on some specific topic on which much interfaces are processed and the general query interface is found.

Local integration approach: Wise-integrator is a system which is for the data integration of E-commerce, and integration of interface in one of the important parts of the system. It is an integrated solution which will analyze the query interface first and obtain the property information. An important tool, Wordnet3, will be used in the process of grammar analysis. Then, the matching of properties will be done. After that on all query interface, the global name, type, and range of value will be confirmed on the integrated query interface for the matched properties. In the experiments, we evaluate the quality of integration by correctness and integrality, and the experimental results are 95.25% and 9791% respectively. In summary, the study has implemented the integration of interface. However, the other side of the approach shows weakness in two aspects. Firstly, the query interface in considered as a flat structure, but actually the query interface has rich structural information. Secondly, only the case of the mapping of 1:1 between the query interfaces is taken into account, but there exist a lot of complex mappings in the practical query interface. To aim at the weakness, we propose the expandedness and improvement on the integration of query interface. First, the schema of query interface is considered as a hierarchical tree structure. Second, a more precise matching and cluster are carried out on the properties of the query interface by “building bridge”. Since the frequent appearance of the complex mappings of 1:m, the complex mapping are divided into two types, one is aggregate and the other is is-a, and the users can take part in the procedure of integration and instruct the integration. Two approaches of full automation and user taking part in are compared by experiments. The average precision and recall of full automation approach are 88.2% and 91.1 respectively, yet the two numbers increase 7.8% and 2.9% respectively by the approach of user taking part in.

Whole integration approach: This approach is different from the local integration approach. Some literatures propose the solution by statistics of schema matching, and a common schema which can describe the general characters of all query interfaces in the domain is hidden in the data source of the same domain of Deep Web. The general schema will be used as a whole to match the entire schema in the same domain. Based on the idea, a general schema matching framework called MGS which includes hypothesis modeling, hypothesis generation, and hypothesis selection is proposed. In the step of hypothesis modeling, a schema which aims at specific problems will be defined. Hypothesis generation is to set the parameters for defined schemas, and hypothesis generation will select the suitable schema from the possible schemas because more than one schema is generated. A viewpoint of property relativity of query interface is proposed, that means all properties will be divided into three types: positive relativity, negative relativity and independent relativity. We give the criteria called H-measure which can be used to decide two properties belongs to one of the three relativities. The most suitable matching will be selected after the relations between the properties of different query interfaces are determined. The experimental results demonstrate that it can find all the relativities accurately for the matching of 1:1 in the query interface, all the negative relativities for the matching of m:n, and only one wrong in the testing data set of the positive relativities matching.

5. CONCLUSION

Query interface is the main method for accessing Web DB. Presently, the research work mainly focuses on classification and integration, and there are two approaches for the implements, one is property analysis, and the other is based on schema. But regardless of the division, the analysis for the query interface mainly includes two sides: schema information and property grammar. The current research work can give the ideal results for the query interface of specific topic, but there is no suitable solution for the simple query interface (e.g. the interface only has one keyword) and for the integrative topic (e.g. E-Commerce). In the future work, we can try to determine the topic and query ability by analyzing the query results, and this will need to design a sample DB related to the topic and can fill the query interface and submitting query automatically.

The other issue we should pay attention is, with the growing of Web service, more and more large business web sites have provided Web Service which makes obtaining and processing information easy and stable. To access information in this way becomes a trend of the future research, but today the web sites could provide Web Service are still rare, and large scare integration based on web page will still be a necessary method.

REFERENCES

[1] Fetterly D., Manasse M., Najork M., Wiener J. L.. A large-scale study of the evolution of web pages. In: Proceedings of the 12th International World Wide Web Conference, Budapest, 2003, 669-678
[2] Chang K. C., He B., Li C., Patel M., Zhang Z.. Structured databases on the web: Observations and Implications. SIGMOD Record, 33, 3, 61-70
[3] Cope J., Craswell N., Hawking D.. Automated discovery of search interfaces on the Web. In: Proceedings of the 14th Australasian Database Conference(ADC 2003), Adelaide, 2003, 181-189.
[4] Zhang Z., He B., Chang K. C.. Understanding Web query interfaces: best-effort parsing with hidden syntax. In: Proceedings of the 23th ACM SIGMOD International Conference on Management of Data, Paris, 2004, 107-118.
[5] Arasu A., Garcia-Molina H.. Extracting structured data from Web pages. In: Proceedings of the 22th ACM SIGMOD International Conference on Management of Data, San Diego, 2003, 337-348
[6] Crescenzi V., Mecca G., Merialdo P.. RoadRunner: towards automatic data extraction from large web sites. In: Proceedings of the 27th International Conference on Very Large Data Bases, Italy, 2001, 109-118.
[7] Wittenburg K. Weitzman L.. Visual Grammars and Incremental Parsing for Interface Languages. In: Proceedings of the IEEE Symposium on Visual Languages (VL), Skokie, 1990, 111-118
[8] He H., Meng W., Yu C. T., Wu Z.: WISE-Integrator: an automatic integrator of Web search interfaces for e-commerce. In: Proceedings of the 29th International Conference on Very Large Data Bases, Berlin, 2003, 357-368

A UML-BASED COSMIC-FFP MEASUREMENT METHOD

Chun-lei Ji,[1,2] Ye-yong Tan[1]

[1.] Department of Computer Science and Engineering , East China University of Science and Technology,200237

[2.] School of Electronic Information,Shanghai Dianji University,200240

Shanghai ,China. Email: jclei@sdju.edu.cn; ronaldo0903@126.com

ABSTRACT

Software Estimation is one of the most important parts of software project management，and the functional size measurement of software is an important part of software estimation. There are several popular methods of functional size measurement, including IFPUG FPA、Mark II FPA、NESMA FPA and COSMIC-FFP. This paper analyzes the semantics of UML artifacts (use case diagram, class diagram and sequence diagram) and primary elements of COSMIC-FFP(Functional User, Layer, Boundary, Data Group and Functional Process) and establishes the mapping from UML artifacts to COSMIC-FFP elements. As a result, 15 mapping rules are proposed. This paper also proposes a UML-based measurement process of COSMIC-FFP, and meantime proposes 11 principles and 8 measurement rules while doing measurement work. This paper also presents a case study for our work.

Keywords: Functional Si， Measurement， COSMIC-FFP， UML Mapping rules

1. INTRODUCTION

As the size and complexity of software increase, the functional size measurement(FSM)method has become more and more popular. There are several popular FSM methods, including IFPUG FPA、Mark II FPA、NESMA FPA and COSMIC-FFP, and they are extended from the original IFPUG FPA. As a next-generation FSM method, COSMIC-FFP [1] was designed to be able to meet the constraints of the many new (and complex) types of data-driven and event-driven software as well as the type of software served by first generation FSM methods, like IFPUG FPA. However, it is manually operated and requires the measurer should have some kind of professional qualities, and different measurement processes and results are made by different measurers, which results in a low repetitiveness of measurement results and limits its applied range. So we need some kind of "auto-measurement" method.

The Unified Modeling Language (called UML)[2] was developed to provide a common language for object-oriented modeling. It was designed to be extensible in order to satisfy a wide variety of needs and was also intended to be independent of particular programming language and development methods. Its wide applied range makes it possible to sizing software in an impersonal and automatic way.

This paper proposes detailed mapping rules between UML artifacts (use case diagram, class diagram and sequence diagram) and primary elements in COSMIC-FFP .It is organized as follows: Section 2 describes the overview of COSMIC-FFP method and the several diagrams by the UML. Next, Section 3 proposes detailed mapping rules between UML artifacts and COSMIC-FFP model, and presents an applicable measurement process based on them. Section 4 describes a case study that evaluates the applicability of our work. Finally, Section 5 concludes this paper and discusses some possible future work. needed to be done.

2. PRELIMINARIES

2.1 COSMIC-FFP Measurement Method

The COSMIC-FFP method measures the functional size of software by analyzing the four data flow types: Entry, Exit, Read and Write. The unit of measurement is 1 data movement, referred to as 1 CFP(COSMIC Function Point).Two distinct and related phases are necessary to perform the measurement: mapping the functional user requirements (FURs) for the software to be measured onto the COSMIC-FFP software model and then measuring the specific elements of this software model(Fig1).

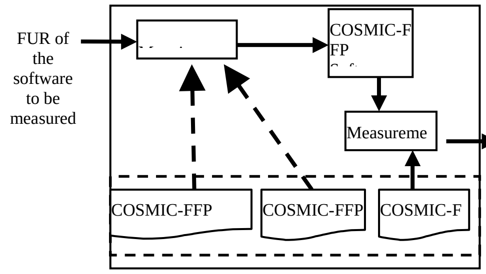


Figure 1. COSMIC-FFP Measurement Process Model

The COSMIC-FFP software model distinguishes four types of data movement sub-process: in the “front end” direction, two types of movement (Entry and Exit) . allow the exchange of data attributes with the users (or other layers); in the “back end” direction, two types of movement (Read and Write) allow the exchange of data attributes with the storage hardware. Every resultful data movement is regarded as a CFP(Fig2).

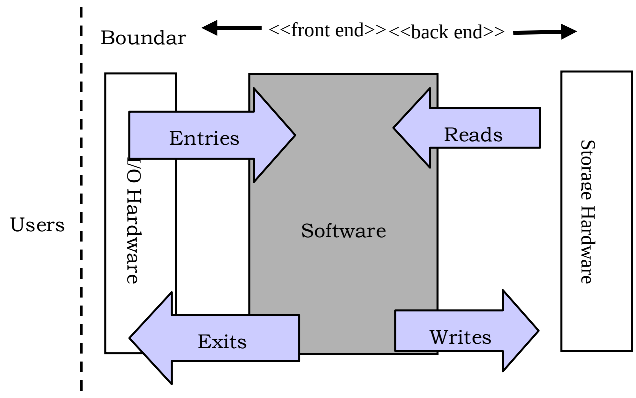


Figure 2. COSMIC-FFP software model and data movement type

2.2 Unified Modeling Language

UML is a universal graphical modeling language used in object-oriented software development process .It is clearly defined, easy to express, functionally powerful and most widely used.

In order to calculate the CFPs from the above diagrams, we use the use case diagram, sequence diagram and class diagram. Because these diagrams include the information about all functions and data manipulated in the system. In subsections 2.2.1,2.2.2 and 2.2.3, we briefly explain the use case diagram, class diagram and sequence diagram.

2.2.1 Use Case Diagram

It is "A diagram that shows the relationships among actors and the subject (system), and use cases"[2].Use case diagram describes a functional unit of the system and it is usually used to express the system-level advanced function. A path through a use case is known as a scenario.

2.2.2 Class Diagram

It is "A diagram that shows a collection of declarative (static) model elements, such as classes, their content and relationships"[2].Class diagrams describe the static structure of the model, that is objects, classes and the relations between these entities including generalization and aggregation. They also represent the attributes and operations of the classes.

2.2.3 Sequence Diagram

It is "A diagram describes an Interaction by focusing on the sequence of Messages that are exchanged, along with their corresponding Occurrence Specifications on the Lifelines".[2] A sequence diagram shows an interaction arranged in time sequence. In particular, it shows the objects participating in the interaction by their "lifelines" and the messages that they exchange arranged in time sequence.

3. MAPPING RULES BETWEEN UML ARTIFACTS AND COSMIC-FFP

The COSMIC-FFP Method mainly takes consider of these concepts: FU(Functional User),OI(Object of Interest), DG(Data Group),DM(Data Movement) and FP (Functional Process).And they can be indicated in UML artifacts. For example, we can use case diagram to identify FUs and system boundaries; class diagram to identify OIs and DGs; sequence diagram to identify DMs and FPs. Detailed mapping rules between UML artifacts and primary COSMIC-FFP elements will be discussed in the following section.

3.1 Use Case Diagram

Use case diagram can help us to identify "measurement scope" and FU, it is also helpful in identifying FPs. The elements in use case diagram which are helpful to counting CFPs include: Actor, Relationship and Use case. The detailed mapping rules are concluded as follows:

R1: FU is the object which initiates a functional use case, and FU includes Human Actor and Non-Human Actor.

According to the definition of FP in COSMIC-FFP Manual,"A functional process is an elementary component of a set of Functional User Requirements comprising a unique, cohesive and independently executable set of data movements ." [1]It is triggered by a data movement (an Entry) from a functional user. We could identify FPs from use case sets through these rules:

R2: For every FU, we take the use cases directly initiated by it into candidate FPs.

R3: For every FP identified by Rule2,we take the use cases which extends it into candidate FPs.

R4: For use case UC_0 that is identified by Rule2 and Rule3 and UC_0 uses or extends UC_1, if UC_1 is an abstract use case,we don't identify it as a FP; otherwise, we identify UC_0 and UC_1 as two separate FPs.

R5: All use cases that are identified by above rules constitute the "measurement scope" and the other use cases will be excluded from our CFP analysis.

3.2 Class Diagram

Use case diagram describes the detailed information about the classes, the objects and the relationships between them. It is mainly used to identify the OIs and DGs in COSMIC-FFP model. These rules are concluded as follows:

R6: Usually the classes which have a stereotype of "Entity" are taken into consideration of OIs; For Real-Time System and embedded system, classes that have a stereotype of "control" could also be considered as candidate OIs. The group of attributes of a class is considered as a DG.

R7: if a class A has a composition relationship with other classes, we will take the "whole" as one DG.

R8: When class A has a aggregation relationship with class B, if A is dying with B, they are identified as one DG; otherwise, they are considered as two different DGs.

R9: When class A has a inheritance relationship with class B, that is, A inherits from B, if B is abstract, B will not be identified as candidate DG; otherwise, B is also identified as DG.

R10: If a class A isn't abstract and there are no relationships between One and other classes (composition, aggregation and inheritance), A is considered as candidate DG; otherwise if A is abstract, it won't be considered as DG.

R11: Rule6-Rule10 help us to identify candidate DGs generally, practically DG is associated with special DM(s): A DM may be associated with several DGs, and a special DG may also be associated with several DMs. This depends on a special system.

3.3 Sequence Diagram

The sequence diagram describes the detailed information about the implementation of a use case, mainly the information about the message exchange, functional invocation and collaboration among the objects participating in the interaction. It puts emphasis on the time sequencing of the messages exchanged among objects. The communication and collaboration between objects are implemented by messages exchange. So the "messages" are what we mainly take into consideration and the rules are as follows:

R12: An ordered, functionally meaningful sequence of messages describe a FP, and this sequence is triggered by an event which is called "triggering event", and ends with a message to FU.

R13: Every message that is exchanged between objects is identified as a DM, and it can have four types: Entry, Exit, Read and Write; and the message which go across the boundary from FU to the system's internal object is considered as the "triggering event" of the FP.

R14: (The determination of the type of DM)The messages which are from FU to interface object are identified as "Entry"; those from interface object to FU are identified as "Exit"; and the messages exchanged between internal objects are identified as "Read" or "Write".

R15: (The determination of the size of DM) The size of a DM is the count of different DGs that are concerned with this DM. It is usually the count of different OIs concerned with this DM.

3.4 Measurement Process

3.4.1 Measurement Purpose

A general measurement matrix model was presented in COSMIC Manual[1].This paper has designed the measurement matrix based on the model and the connections between UML and COSMIC-FFP, and the matrix is showed as Table 1.

According to this table1, the measurement process needs to finish 4 tasks, including: Identify Functional Processes; Identify relevant Data Groups; Identify Messages and Message Type; Count CFP.

3.4.2 Preconditions

Software development is a top-down and decomposing process. If the software to be measured has a overhigh level of abstraction, COSMIC-FFP is not suitable for its measurement[1]. So the software to be measured should be[1]: (1) The Functional User should be Human, Device or other software; (2) The software should response to every occurrence of an event. The COSMIC-FFP requires a "functional process" level of granularity of software decomposition.

3.4.3 Principles

This paper proposes 11 principles that run through the whole measurement process. And they are used to handle several situations in UML sequence diagram, such as Reference, Branchiness, Loops and Parallelism.

P1 Named Principle. It defines the naming rule for Functional processes, Messages and Data Groups.

a) Functional Process: [FU_Triggering Event_Sub-condition1_Sub-condition2_...]

b) Message: [Classname: Operation_ Loop Iterator]

c) Data Group: [Classname]

P2 Started Principle. FP is started by input messages from FU.

P3 Multi-input Principle. A FP may receive input from different FUs, such as application wizard.

P4 Stop Principle. A FP ends with the last occurrence of the message that is relevant to it.

P5 Sequence Principle. FP sand sub-FPs is identified according to the sequential order of the messages in sequence diagram.

P6 Retrospective principle. For every message invocation, we should check if there is a sub-sequence diagram that defines this invocation, and if we have found one, we should merge the two sequence diagrams into one and make analysis on the merged one.

P7 Branchiness Principle. If some conditional branchiness exist in the sequence diagram, we should regard each branch as separate FPs.

P8 Loop Principle. The recurrent messages can be regarded as a group of messages. And they can be separated by postfix of message serial number when recording.

P9 Parallelism Principle. The parallel messages that occurred at the same moment belong to the same FP.

P10 Relevance Principle. If the parameters or the return values of a message is a subset of one DG, we regarded that the message is relevant to the DG.

P11 Eliminate Repetitive Principle. When identify Entry and Exit, if two Entry(Exit)s are relevant to one same DG, we regarded them as one Entry(Exit), and they are only recorded once in the measurement matrix table.

3.4.4 Measurement Rules

This paper presents 8 measurement rules based on the mapping between UML and COSMIC-FFP.

MR1 According to Class Diagram and Rule6-Rule11 in section 3.2, we write down every class that are already identified as one DG in Measurement Matrix Table.

MR2 For every sequence diagram that A FU has participated, we use Start Principle to recognize FPs. And the FPs is recorded in Measurement Matrix Table.

MR3 The messages in a FP are recognized using Sequence Principle. After a message is recognized, Stop Principle is used to check if the FP has ended.

MR4 Every message in a FP is recorded in the corresponding location of the Measurement Matrix Table.

MR5 Determine message type.

a) If a message flows from FU to Software and it is relevant to some DG, then it is recognized as type of Entry.

b) If a message flows from FU to Software but there is no DG relevant to it, then it is recognized as type of Control Command.

c) If a message flows from Software to FU, then it is recognized as type of Exit.

d) If a message flows from storage hardware to Software, then it is recognized as type of Read.

e) If a message flows from Software to storage hardware, then it is recognized as type of Write.

f) The messages which are exclude from case a) to e) are recognized as Data Manipulation.

MR6 Record relevance of messages and DGs. According to Relevance Principle, the relevance of messages and DGs is recorded in the "Data Group" Column of Measurement Matrix Table, and the relevance factor is recorded in the "Message Type" column of Measurement Matrix Table. The record format of the relevance of messages and DGs are as follows:

a) If a message is relevant to a DG, we recorded a "Y" in the corresponding location.

b) If a message is not relevant to a DG, we recorded a "X" in the corresponding location.

MR7 When identifying FPs, we should keep an eye on the Retrospective Principle, Branchiness Principle, Loop Principle and Parallelism Principle.

MR8 For every sequence diagram that a FU has participated, Use MR1-MR7 to fill the Measurement Matrix Table until all sequence diagrams have been analyzed.

4. CASE STUDY

In this section we analyze a simple case to demonstrate the applicability of our rules. A module in a Student Information Management System is considered. The Module has 4 Persistent Classes: Student, Course, Institute and StudentCourse. The Class Diagram are as Fig3.

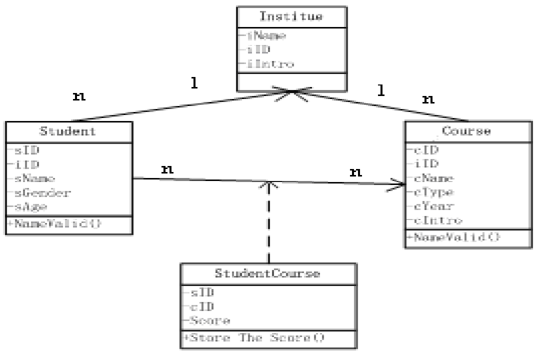


Figure 3. Module Class Diagram

A "Student Score Data Entry" scenario is analyzed here. When a user try to input RegisterScore (Student, Course, Score) in UIScoreRegister UIScoreRegister makes an data consistency check ConsistencyCheck (Student, Course, Score). If succeed, store the score with StoreTheScore(Student,Course,Score), and return a "success" message to user. Otherwise return a "warning" message. The sequence diagram that describes this scenario is as Fig4.

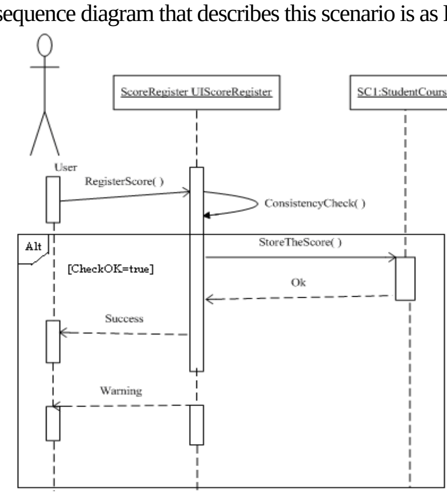


Figure 4. The Sequence Diagram for the Case Study

The ConsistencyCheck in UIScoreRegister has a sequence diagram describing it as Fig5.

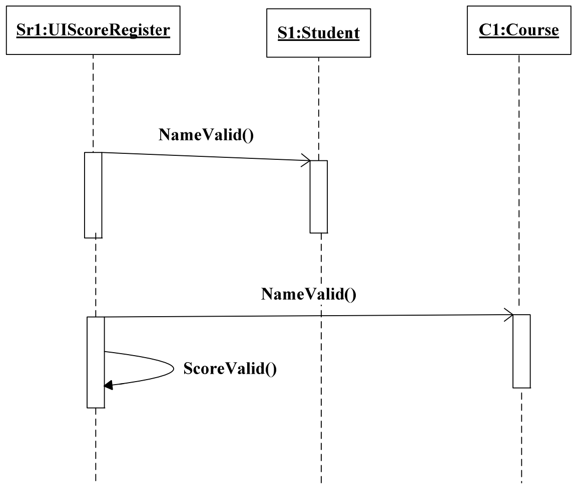


Figure 5. The Sequence Diagram for Consistencycheck

According to Retrospective Principle, Fig5 can be regarded as a sub-diagram of Fig4, So we merge them into Fig6.After analyzing Fig6 using MR1-MR8, we get the measurement result matrix table Table2.

5. CONCLUSION AND FUTURE WORK

This paper has discussed the UML artifacts, COSMIC-FFP method and the connections between them, and proposed 16 mapping rules, which make it possible to measure the software size automatically. The future work may include:

1. Analyse other UML artifacts(for example, the activity diagram, the state diagram) and find the mapping rules between them and COSMIC-FFP model. We will get an all-around knowledge of our software after that work.

2. Design and implement a kind of automatic tool for FSM based on the proposed rules.

REFERENCES

[1] Alain Abran,Jean-Marc Desharnais,Serge Oligny, Denis St-Pierre,Charles Symons,"COSMIC FFP Measurement Manual,Version 3.0",The Common Software Measurement International Consortium,Sept. 2007.

[2] OMG Unified Modeling Language (OMG UML), Superstructure, V2.1.2 .

[3] Alain Abran,Jean-Marc Desharnais,Fatima Aziz. "Measurement Convertibility - From Function Points to COSMIC-FFP".

[4] Geert Poels."Definition and Validation of a COSMIC-FFP Functional Size Measure for Object-Oriented Systems",QAOOSE 2003.

[5] Manar Abu Talib, Adel Khelifi, Alain Abran and Olga Ormandjieva."A case study using the COSMIC-FFP Measurement Method for Assessing Real-Time System Specifications",Proceedings of the IWSM - Mensura 2007.

[6] Çigdem Gencel, Luigi Buglione, Onur Demirors, Pinar Efe."A Case Study on the Evaluation of COSMIC-FFP and Use Case Points",Proceedings of Smef,2006.

[7] Frank W. Vogelezang,Sogeti Nederland B.V. "Implementing COSMIC-FFP as a replacement for FPA",IWSM/MetriKon 2004.

[8] Klaas van den Berg, Ton Dekkers, Rogier Oudshoorn. "Functional size measurement applied to UML-based user requirements".

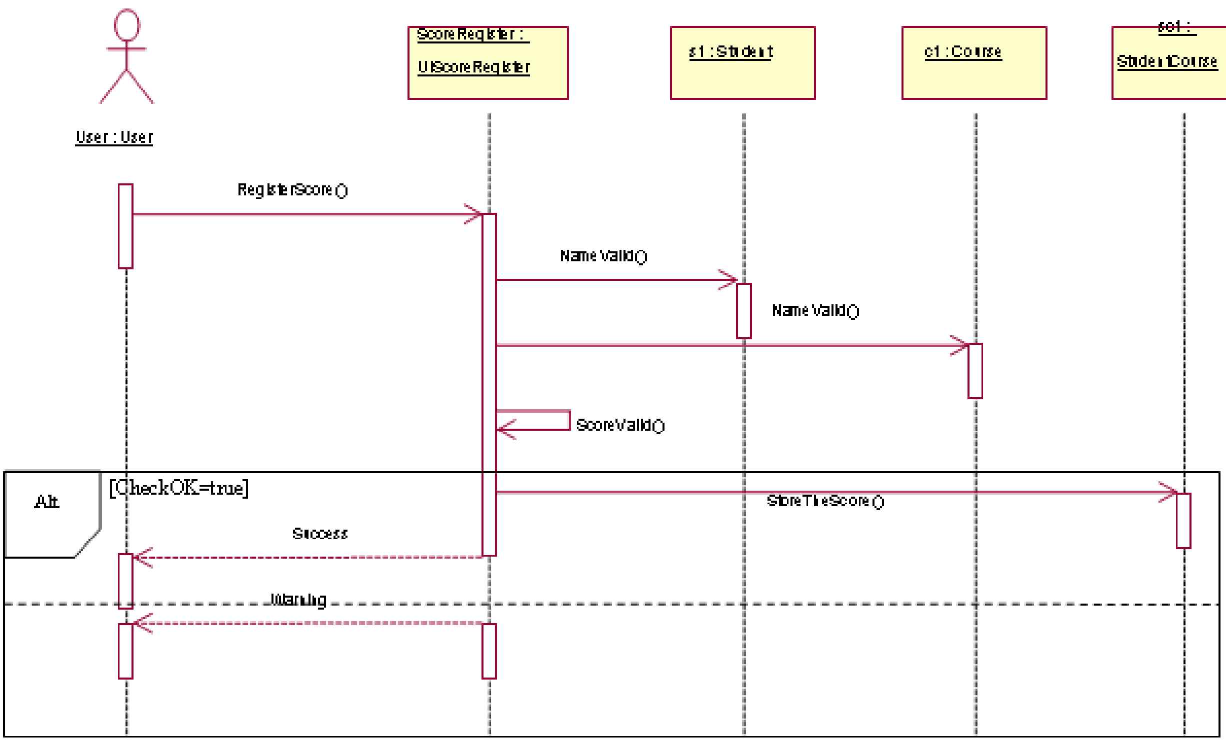


Figure 6. The Merged Sequence Diagram

Table1 Measurement Output Matrix

Fuctional Process	Message	Data Group				Message Type					
		Data Group1	Data Group2	Data Group3	..	Entry	Exit	Read	Write	Data Manipulation	Control Command
Fuctional Process1	Message 1	X	Y	X	..	1	0	0	0	0	0
	Message 2	Y	Y	X	..	0	2	0	0	0	0
	Message 3	X	X	Y	..	0	0	1	0	0	0
	Message 4	X	X	Y	..	0	0	0	0	1	0
	…	…	…	…	..	…	…	…	…	…	0
Fuctional Process2	Message 1	X	Y	Y	..	0	2	0	0	0	0
	Message 2	Y	X	X	..	0	0	0	0	1	0
	Message 3	X	X	Y	..	0	0	1	0	0	0
	Message 4	X	X	Y	..	1	0	0	0	0	0
…	…	…	…	…	..	…	…	…	…	…	…
Total	…	…	…	…	..	…	…	…	…	…	…

Table2 Measurement Table For Case Study

Functional Process	Message	Data Group				Message Type					
		Student	Course	StudentCourse	Institute	En	Ex	R	W	DM	CC
User_RegisterScore_Success	ScoreRegister :RegisterScore	Y	Y	Y	X	3	0	0	0	0	0
	S1:NameValid	Y	X	X	X	0	0	1	0	0	0
	C1:NameValid	X	Y	X	X	0	0	1	0	0	0
	ScoreRegister :ScoreValid	X	X	X	X	0	0	0	0	1	0
	Sc1:StoreTheScore	X	X	Y	X	0	0	0	1	0	0
	User:Success	X	X	X	X	0	1	0	0	0	0
User_RegisterScore_Warning	S1:NameValid	Y	X	X	X	0	0	1	0	0	0
	C1:NameValid	X	Y	X	X	0	0	1	0	0	0
	ScoreRegister :ScoreValid	X	X	X	X	0	0	0	0	1	0
	User:Warning	X	X	X	X	0	1	0	0	0	0
Total		3	3	2	0	3	2	4	1	2	0

PRINCIPLE OF SYMMETRY FOR NETWORK TOPOLOGY *

Wenjun Xiao, Yanxia Liu, Qin Zhang
School of Software Engineering, South China University of Technology
Guangzhou, Guangdong 510641, P.R. China
Email: wjxiao@scut.edu.cn

ABSTRACT

General networks such as Internet are complex heterogeneous networks, which are constructed by many different organizations, and so become non-effective ones. Therefore one constructed a level of software over networks which is called overlay or virtual topology. In this paper we present principle of symmetry for general network topology by using cayley graph models and show its necessity to improving network performance. We explain the main conclusions of the paper by many examples in optical, wireless and peer-to-peer networks.

Keywords: Principle of Symmetry, Network Topology, Overlay, Virtual Topology, Cayley Graph, Performance

1. INTRODUCTION

We know that general networks such as Internet have become more and more complex, which are constructed by many different organizations, heterogeneous, and so very non-effective[6,8]. Hence one constructed the level of software over networks which are called overlay or virtual topology to improve its performance[17-24]. In this paper we want to solve the following important problem: How should we design overlay (or virtual) topology for networks in order to improve network performance? We try to present an answer to the problem by principle of symmetry. We show that we should use a class of overlay topology of higher symmetry such as Cayley graphs in construction of the level of software. Thus we propose a general method of constructing overlay topology by means of Cayley graph model.

The fact that Cayley (di)graphs and coset graphs are excellent models for interconnection networks, studied in connection with parallel processing and distributed computation, is widely acknowledged [1], [2], [6-7]. Many well-known interconnection networks are Cayley (di)graphs or coset graphs. For example, hypercube, butterfly, cube-connected cycles and honeycomb networks are Cayley graphs, while de Bruijn and shuffle-exchange networks are coset graphs [5-6], [9-16].

Much work on interconnection networks can be categorized as ad hoc design and evaluation. Typically, a new interconnection scheme is suggested and shown to be superior to some previously studied network(s) with respect to one or more performance or complexity attributes. Whereas Cayley (di)graphs have been used to explain and unify interconnection networks with many ensuing benefits, much work remains to be done. As suggested by Heydemann [4], general theorems are lacking for Cayley digraphs and more group theory has to be exploited to find properties of Cayley digraphs.

By many examples in optical, wireless and peer-to-peer networks we explain the main conclusions.

2. PATTERNS OF GENERAL NETWORK DESIGN

We consider patterns of general networks design should possess the following properties:

lower network algorithm cost;
lower network hardware cost;
lower communication delay;
higher reliability and fault-tolerance.

Lower network algorithm cost denotes lower software development one and lower network hardware cost may be expressed as the number of network node ports. Lower communication delay may be denoted as network average distance or its maximum-diameter. But we should further consider trade-off between the number of network node ports and network average distance. One extreme example is complete graph, which is of the smallest diameter but the greatest node ports; another extreme example is line, which is of the greatest diameter but the smallest node ports. Networks of higher reliability and fault-tolerance usually are ones of higher symmetry[6,8]. We shall further explain it in the section 3.

3. PRINCIPLE OF SYMMETRY FOR NETWORK TOPOLOGY

We have known that symmetry for network topology possess very important influence for interconnection network performance in parallel computing[6-8]. We use interconnection networks of higher symmetry such as hypercube, mesh, torus, butterfly and De Bruijn graphs which all are Cayley graphs or their subgraphs and quotient graphs[4]. It is not strange because these networks of higher symmetry possess many excellent properties such as simpler routing algorithm and higher reliability and fault-tolerance[4]. We have known that Cayley graphs are vertex transitive and possess uniform routing and other communication algorithms[4]. They also have higher reliability and fault-tolerance[4]. We also known that some Cayley graphs have used as static model for overlay(or virtual) topology in optical, peer-to-peer and wireless networks[17-24]. Thus we may present Cayley graphs as static model for overlay topology of general networks and give uniform processing to them.

In the following, we give some notations and definitions for group and graph[3]. Let G be a finite group and S a subset of G. The subset S is said to be a generating set for G, and the elements of S are called generators of G, if every element of G can be expressed as a finite product of their powers. We also say that G is generated by S. The Cayley digraph of the group G and the subset S, denoted by $Cay(G, S)$, has vertices that are elements of G and arcs that are ordered pairs (g, gs) for $g \in G$, $s \in S$. If S is a generating set of G then we say that $Cay(G, S)$ is

* This Work is supported by Guangdong Key Laboratory of Computer Network (CCNL200705) and Guangdong Laboratory of Software and Applied Technology (2006B80407001) and the National Basic Research Program of China under Grant No. 2003CB314805.

the Cayley digraph of G generated by S. If 1 ∉ S (1 denotes the identity element of G) and $S = S^{-1}$, then Cay(G, S) is a simple graph.

Assume that Γ and ∑ are two digraphs. The mapping ϕ of V(Γ) to V(∑) is a homomorphism from Γ to ∑ if for any (u, v) ∈ E(Γ) we have (ϕ(u), ϕ(v)) ∈ E(∑). In particular, if ϕ is a bijection such that both ϕ and the inverse of ϕ are homomorphisms then it is called an isomorphism of Γ to ∑.

4. MATHEMATICAL METHORS FOR NETWORK TOPOLOGY DESIGN

4.1. Combinatorial Method

We first give the definition of generalized homomorphism of graphs. Given two graphs $G_o(V_o, E_o)$ and $G_p(V_p, E_p)$ .The generalized homomorphism is the mapping f from V_o to V_p such that for any $(m,n) \in E_o$ there are a path $P_{m,n}$ of G_p such that $f((m,n)) = P_{m,n}$. Here $P_{m,n}$ may be a path of length zero(that is, it is a vertex). The generalized homomorphism has many applications in networks. We propose three examples in the following.

Example 1. In WDM optical networks we often consider virtual topology design problem[25]. G_o and G_p are virtual and physical networks respectively. The light path(an edge) in G_o is mapped to a path in G_p. It has a constraint that any two light paths of the same wavelength in G_o can not through the same edge in the physical topology G_p.

Example 2. In wireless mesh networks we often consider logical topology design problem[26]. G_o is a multigraph and logical links between two nodes must be assigned different channel number.

Example 3. In peer-to-peer networks we often consider overlay topology design problem[17]. G_o and G_p are dynamic overlay network and static network respectively. Furthermore the mapping f is often a contraction(merging) which maps an edge to a vertex.

4.2. The Optimization Method

The optimization method is important in network research[6]. We consider the following general network parameters: traffic matrix, the maximum number of ports and the diameter of virtual topology, and the maximum number of ports and the diameter of physical topology. We also consider some special parameters such as the number of wavelengths for optical networks and the number of channels for wireless networks. We first consider virtual topology design problem for general networks.

We use the following notations:

s and *d* used as subscript or superscript denote source and destination of a packet, respectively. *i* and *j* denote originating and terminating nodes, respectively, in a virtual path. *m* and *n* denote terminating nodes of a physical path.

Given: Number of nodes in the network = *n*.

$T = (\lambda^{sd})$ is the traffic matrix, i.e., λ^{sd} is the arrival rate of packets *s* that are destined for *d*.

$\lambda_{ij} = \sum_{sd} \lambda_{ij}^{sd}$, where λ_{ij}^{sd} is the arrival rate of packets from *s*-*d* pair (*s*, *d*) on virtual path (*i*, *j*).

$\lambda_{ij}^{sd} = a_{ij}^{sd} \lambda^{sd}$, where a_{ij}^{sd} is the traffic which is routed from *s*-*d* pair (*s*, *d*) on virtual path (*i*, *j*).

$\lambda_{\max} = \max_{ij} \lambda_{ij}$, it is called congestion. It is an important parameter[25].

$b_{ij} \in \{0,1\}$ is binary variables, one for each possible virtual path such that $b_{ij} = 1$ if there is a virtual path from node *i* to node *j* in the virtual topology and $b_{ij} = 0$ otherwise.

Maximum number of wavelengths per fiber =*W* (a system-wide parameter).

Maximum number of ports = Δ in virtual topology.

Diameter of virtual topology = *D*.

Maximum number of ports = Δ_p in physical topology.

Diameter of physical topology = D_p .

For virtual topology design of general networks, we can consider the following optimization problem[25].

Objective:

Minimize: $\lambda_{\max}$

Subject to:

Flow conservation at each node:

$$\sum_j \lambda_{ij}^{sd} - \sum_j \lambda_{ji}^{sd} = \begin{cases} \lambda^{sd}, & if\ s = i \\ -\lambda^{sd}, & if\ d = i,\ \forall s,d,i \\ 0, & otherwise \end{cases} \tag{1}$$

Total flow on a virtual path:

$$\lambda_{ij} = \sum_{sd} \lambda_{ij}^{sd},\ \forall i, j \tag{2}$$

$$\lambda_{ij} \le \lambda_{\max},\ \forall i, j \tag{3}$$

$$\lambda_{ij}^{sd} \le b_{ij} \lambda^{sd},\ \forall i, j, s, d\ . \tag{4}$$

Degree constraints:

$$\sum_i b_{ij} \le \Delta,\ \forall j \tag{5}$$

$$\sum_j b_{ij} \le \Delta,\ \forall i\ . \tag{6}$$

Bidirectional virtual path constraint:

$$b_{ij} = b_{ji}\ . \tag{7}$$

Non-negative and integer constraints:

$$\lambda_{ij}^{sd},\ \lambda_{ij},\ \lambda_{\max} \ge 0,\ \forall i, j, s, d \tag{8}$$

$$b_{ij} \in \{0,1\} \quad \forall i, j\ . \tag{9}$$

It is noticed that objective function and constraints are a linear function of variables (λ_{ij}^{sd}, λ_{ij}, $\lambda_{\max}$, b_{ij}), and hence it is a mixed-integer linear programming problem (MILP). In general, it is a NP-hard problem.

In the following, we consider a special case when $\lambda^{sd} = 1$ for *s*≠*d*.

We conclude the following.

Theorem 1. when the traffic is even (λ^{sd} = constant), $\lambda_{\max}$ is inverse proportion to degree Δ approximately, where virtual topology is any Cayley graph.

5. SOME VIRTUAL TOPOLOGY DISIGNS BY CAYLEY GRAPH MODELS

5.1. Suppose that physical topology in WDM optical networks is arbitrary, we proceed to design virtual topology and

wavelength assignment and routing problem[25].
In the following, we consider it when virtual topology is Cayley graph.

Definition 1. Let $C_{ij}^k = 1$, if the lightpath from nodes *i* to *j* uses wavelength *k*, else $C_{ij}^k = 0$. Let $C_{ij}^k(m,n) = 1$, if the lightpath from nodes *i* to *j* uses wavelength *k* and is routed through physical link (*m*, *n*), else $C_{ij}^k(m,n) = 0$. Let P_{mn} be a physical link in the physical topology, if $P_{mn} = 1$ then there is a fiber link between nodes *m* and *n*, otherwise $P_{mn} = 0$. Let D_p be maximum of hops in virtual link between any two nodes *i* and *j*.

(a) Unique wavelength constraints

$\sum_{k=0}^{W-1} C_{ij}^k = b_{ij}, \quad \forall (i,j)$, where the definition of b_{ij} is as that in the section 4.2.

Note that $k \in \{0,1,2,\ldots,W-1\}$,

$C_{ij}^k(m,n) \le C_{ij}^k, \quad \forall (i,j),k,(m,n)$.

(b) Wavelength clash constraints

$\sum_{ij} C_{ij}^k(m,n) \le 1, \quad \forall (m,n),k$.

(c) Conservation of wavelength constraints

$$\sum_{k=0}^{W-1}\sum_m C_{ij}^k(m,n)P_{mn} - \sum_{k=0}^{W-1}\sum_m C_{ij}^k(n,m)P_{nm} = \begin{cases} b_{ij}, & if\ n=j \\ -b_{ij}, & if\ n=i,\ \forall (i,j),n \\ 0, & if\ n \ne i\ and\ n \ne j \end{cases}$$

(d) Hop Bound Constraints

$\sum_{mn} C_{ij}^k(m,n) \le D_p, \quad \forall (i,j),k$

(e) Symmetry Constraints

$C_{ij}^k(m,n) - C_{ji}^k(n,m) = 0, \quad \forall k,(i,j),(m,n)$.

By means of the method in the section 4.2 we may solve the problem on virtual topology design and wavelength assignment with routing in WDM optical networks. The problem is NP-hard one in general and needs to use heuristic methods to find solutions[25].

5.2. Suppose that physical topology is $G_p(V_p,E_p)$ in wireless networks, we proceed to design virtual topology and channel allocation and routing by Cayley graphs[26].

Definition 2. Let *C* be orthogonal frequency channels available and frequency channel $k \in \{1,2,\ldots,C\}$, and *I* be the number of network interface cards(NICs) of nodes.

We define a link channel allocation variable x_{ij}^k. In the virtual topology, if node *i* communicates with node *j* over the *k*th frequency channel, then $x_{ij}^k = 1$, otherwise, $x_{ij}^k = 0$.

To establish the virtual links, nodes *i* and *j* should assign the same frequency channels to communicate with each other. This requires that,

$$x_{ij}^k = x_{ji}^k, \quad \forall i,j \in V_p, e_{ij} \in E_p, \forall k = 1,2,\ldots,C. \qquad (11)$$

For any node $i \in V_p$ and any channel $k \in \{1,2,\ldots,C\}$, we define y_i^k to be as following:

$$y_i^k = \begin{cases} 1, & if\ \exists j \in V_p\ and e_{ij} \in E_p, such as\ x_{ij}^k = 1 \\ 0, & otherwise \end{cases} \qquad (12)$$

We refer to y_i^k as the node channel allocation variable corresponding to node *i* and channel *k*. $\sum_{k=1}^{C} y_i^k$ indicates the total number of channels that are being used by node *i* to establish virtual links with its neighboring nodes. Since each NIC operates on a distinct frequency channel, $\sum_{k=1}^{C} y_i^k$ cannot be larger than the total number of available NICs on node *i*. That is,

$$\sum_{k=1}^{C} y_i^k \le I, \quad \forall i \in V_p \qquad (13)$$

y_i^k is a continuous real variable for all nodes $i \in V_p$ and all channels $k \in \{1,2,\ldots,C\}$ and also requiring that:

$$0 \le y_i^k \le \sum_{j \in V_p, e_{ij} \in E_p}^{C} x_{ij}^k \qquad (14)$$

$$x_{ij}^k \le y_i^k \le 1, \quad \forall j \in V_p, e_{ij} \in E_p \qquad (15)$$

The flow conservation requires that for $s,d,i \in V_p$,

$$\sum_{j \in V_p, e_{ij} \in E_p} \sum_{k=1}^{C} a_{ij,k}^{sd} \lambda^{sd} - \sum_{j \in V_p, e_{ij} \in E_p} \sum_{k=1}^{C} a_{ji,k}^{sd} \lambda^{sd} = \begin{cases} \lambda^{sd}, & if\ s=i \\ -\lambda^{sd}, & if\ d=i,\ \forall s,d,i \\ 0, & otherwise \end{cases} \qquad (16)$$

In (16), the term on the left-hand side is the net flow out of node *i* for the flow from source *s* to destination *d*. The net flow is the difference between the outgoing flow and the incoming flow. The term on the right-hand side is equal to 0 if node *i* is neither the source nor the destination for that specific flow. If node *i* is the source, then the net flow is equal to λ^{sd}. If node *i* is the destination, then the net flow is equal to $-\lambda^{sd}$.

We define a variable $a_{ij,k}^{sd}$, and $0 < a_{ij,k}^{sd} \le 1$ if the traffic from source *s* to destination *d* is being routed via link (*i*, *j*) over channel *k*, otherwise it is equal to 0.

$$\sum_{k=1}^{C} a_{ij,k}^{sd} \le 1, \quad \forall s,d,i,j \in V_p, e_{ij} \in E_p \qquad (17)$$

Let $a_{ij}^{sd} = \sum_{k=1}^{C} a_{ij,k}^{sd}$, we have (16):

$$\sum_{j \in V_p, e_{ij} \in E_p} a_{ij}^{sd} \lambda^{sd} - \sum_{j \in V_p, e_{ij} \in E_p} a_{ji}^{sd} \lambda^{sd} = \begin{cases} \lambda^{sd}, & if\ s=i \\ -\lambda^{sd}, & if\ d=i,\ \forall s,d,i \\ 0, & otherwise \end{cases} \qquad (18)$$

We need to solve the following optimization problem.

Minimize: $\lambda_{\max}$

Subject to (11),(13),(14),(15), and (18),

where $x_{ij}^k, y_i^k \in \{0,1\}, a_{ij}^{sd}, \lambda_{\max} \ge 0, a_{ij}^{sd} \le 1,$ $\forall i,j,s,d \in V_p, e_{ij} \in E_p, \forall k = 1,2,\ldots,C$ is variables and x_{ij}^k, y_i^k is integer variables.

It is a mixed-integer linear programming problem (MILP). The problem is NP-hard one generally and may be used by heuristic methods to find sub-optimal solutions[26].Here as an example we may consider the simple case that the virtual topology is ring[or torus or butterfly graph].

5.3. In peer-to-peer networks we often consider overlay topology design problem[17]. G_o and G_p are dynamic overlay network and static network respectively. Furthermore the mapping f is often a contraction(merging) which maps an edge to a vertex. In the following we suppose that static network is Cayley graph which definition is as follows.

Example 4. we research P2P networks in detail. Almost all recent researches on P2P system focus on how to build a highly usable P2P overlay network. Researchers include small routing table, short query path and robustness into their design objectives of overlay topology. In this paper, we define a new cayley graph, and based on this cayley graph, a set of P2P DHT protocols have many excellent properties such as small routing table and short query path and high clustering and robustness. Our system has simpler routing(search) and many other properties than the most former systems because its symmetry. The performance is evaluated in this paper, indicating that the new protocols can reach the theoretical lower bound of routing table size and query path length. Furthermore, the robustness of the new Cayley graph is also better than most the P2P DHTs recently proposed.

Centralized computing is not suitable for those applications because it's not scalable. P2P computing is invented to solve this problem and *P2P DHT* (Distributed Hash Table) which belongs to the structured P2P branch is introduced to overcome the inefficient problem of centralized and unstructured branches. One of the fundamental problems in P2P systems is resource locating. Almost all recent researches of P2P DHT [17-19] concentrate on how to lower the length of query path and reduce the size of routing table. Therefore the uses' download behavior and requirement are not taken into consideration when they design the systems**.** In addition to searching, browsing is one of other important requirements when people use P2P file-sharing systems, but as we know, none of current P2P DHT can support efficient file browsing service. The main cause is that all P2P DHTs are based on old graphs such as ring, butterfly and CCC, which are not designed for P2P systems, and thus don't provide the ability of semantic clustering. Therefore those graphs may not ease the overloading problem on the underlying network.

To remedy the disadvantages of the recent structured P2P system, it is necessary to introduce some characteristics of human society, i.e. small-world phenomena [20], into the overlay network. The phenomena of small-world lead to the phenomena of community(clustering), which means that people with the same interests know each other with high probability (i.e. highly clustered). In order to introduce small-world features into P2P DHT, we define a new cayley graph Γ, and then based on Γ, a new P2P DHT called EBu is also designed. In addition to efficient resource searching mechanism, EBu supports explicit grouping of peers, and thus supports effective resource browsing service. Both theoretical analysis and experimental evaluation show that EBu can reach the lower bound of routing table size and query path length at the same time [19]. Moreover, its robustness is also better as compared to Ulysses [17].

Design Objectives

We are primarily interested in the following features of P2P DHT network.

1) *Short query path:* A desirable P2P network should achieve a short query path so that it can respond to a query quickly and thus reduce the load on the underlying network. A P2P network may consist of thousands or even millions of computers, so the scalability of our system should be ensured so that the average length of query path would not increase significantly as the number of peers in the overlay network become larger.
2) *Reasonable size of routing table:* Keeping the routing table in a reasonable size is another challenge when designing a scalable P2P overlay network. The minimal routing table size is beneficial to ensure fault-tolerance while the maximal one is relevant for ensuring bounded maintenance cost [21].
3) *Reasonable cluster coefficient:* [20] defines the clustering coefficient (CC1) as follows. Suppose that a vertex v has k_v neighbors; then at most k_v (k_v - 1) / 2 edges can exist between them. Let C_v denote the fraction of these allowable edges that actually exist. Then CC1 can be defined as the average of C_v over all v. Non-zero cluster coefficient leads to the phenomena of clustering and community. Nevertheless, the greater CC1 is, the worse connected the network is, because a lot of links are wasted to increase CC1 of the graph. Therefore the balance between connectivity and CC1 should be achieved by EBu.
4) *Grouping:* Applications based on topic-based publish/subscribe model need explicit grouping of peers. And although grouping is not included in the functional requirement of other software, the performance will be significantly improved if it is introduced. For example, the performance of file sharing system can be remarkably improved by structured grouping, and with which, users can easily browse the files they are more interested in.
5) *Self-configuration:* It is not possible for a large scale P2P system to employ centralized server to provide joining, departing and searching services. For the sake of scalability, distributed network services are preferred.
6) *Robustness:* Robustness refers to the influence to the system performance due to faulty peers in the network. Our performance evaluation shows that the robustness of EBu is excellent.

5.3.1. Related Research

The motivation for our research stems from the following fields:

1) Ulysses (see[17])
2) Content-based Shortcut Gnutella(see [22])
3) Small-world Network

Many biological, technological and social networks lie somewhere between regular and random graph. These systems can be highly clustered, like regular lattices, yet have small path lengths, like random graphs. Watts and Strogatz call this type of network small-world networks, by analogy with the small-world phenomenon [20]. Large CC1 and small path lengths are the most significant features of small-world networks.

4) Cayley Graph as Models of Small-world Networks

[23] presents a model of deterministic small-world graph. Let G be a finite group. Assume that graph $\Sigma = Cay(G, S)$ for some generating set of S of G, where the identity element $e \notin S$ and $S^{-1} = S$. Because Σ is symmetric, the CC1 of all nodes are the same, that is to say, only the CC1 of the vertex e is needed to be computed. The neighbor set of e is S. Assume that $H \subseteq S$ and $H \cup \{e\}$ is a subgroup of G. Then if $h_1, h_2 \in H$, there must exists $h_3 \in H$, so that $h_1 = h_2h_3$. Consequently, at least $|H|$ ($|H|$ - 1) / 2 edges exist among the neighbor set S of vertex e. If H can be chosen to be large, the CC1 of Σ will be large.

5.3.2. The Definition of Static Graph of EBu

Every cayley graph is vertex transitive, and thus using cayley graph as the static graph of P2P DHT has the benefit of distributing loading to all peers evenly [24]. In this section, definition of the static graph Γ of EBu is presented and then some properties of the graph which are essential to P2P systems are explored.

A. Terminology and Notation

In this section, x and y are assumed to be strings composed of digital or wildcard "*". Throughout this paper, vertex, node and peer are used alternatively. So are graph and network. In addition, we will use vertex identifiers or peer identifiers for themselves respectively.

B. The Definition of Cayley Graph Γ

The definition and related terminology of Γ are presented in this section.(we omit the proof because of the limit of space)

Definition 3. Let $H = (G, \bullet)$, where $G = (Z_r, Z_k)$. The operation $\bullet$ on G is defined as $\forall(\mathbf{c}_1, r_1), (\mathbf{c}_2, r_2) \in G$,

$$(c_1, r_1) \bullet (c_2, r_2) = (c_1 \oplus \sigma^{r_1}(c_2), r_1 + r_2)$$

Here, σ is cyclic right shift operation. $\oplus$ is component-wise addition mod r in Z_r. Unless noted otherwise, + is modulo-k addition throughout this paper.

Proposition 1. The operation $\bullet$ on G is associative.

Corollary 1. $(G, \bullet)$ is a group.

For any $(\mathbf{c}, r) \in G$, $\mathbf{c}$ is referred to as *group identifier*, r as *region identifier*, and $(\mathbf{c}, r)$ as *vertex identifier*. In order to define Γ, a subset S of G should be given first. We compose S in two steps:

First, let $S = \varnothing$. Then, links are added by letting,

$$S_c = \{(c0^{k-1}, 1), (0^{k-1}c, -1) \mid c \in Z_r\},$$
$$S = S \cup S_c$$

Finally,

$$S_r = \{(0^k, \gamma) \mid \gamma \in Z_k \setminus 0\}$$
$$S = S \cup S_r$$

Given a group G and its subset S, the definition of our graph Γ is clear:

Definition 4. Γ is a cayley graph defined on G and S, that is $\Gamma = Cay(G, S)$.

Clearly, $S = S^{-1}$, and so Γ is a simple graph.

C. Some Properties of Γ

The degree of a cayley graph equals the cardinality of S, that is $|S| = |S_c| + |S_r| = r + k - 1$.

Proposition 2. Γ is a $(r + k - 1)$-regular graph

The routing from $(\mathbf{c}, r)$ to $(\mathbf{c}_3, r_3)$ proceeds in two phases. In the first phase $\mathbf{c}$ successively change to $\mathbf{c}_3$. In the second phase, one step is required to correct the region identifier to r_3. The pseudo-code for forwarding in a vertex is shown in algorithm 1.

Input: The identifier of current vertex $(\mathbf{c}_1, r_1)$ with $c_1 = x_1 x_2 \ldots x_k$ and the destination vertex $(\mathbf{c}_3, r_3)$ with $c_3 = y_1 y_2 \ldots y_k$

Output: The identifier of next-hop vertex $(\mathbf{c}_2, r_2)$

```
if (c1 = c3 ∧ r1 = r3 ) then
    the destination has been reached
else
    if (c1 = c3 ) then
        (c2, r2) := (c3, r3)
    else
        c2 := x1...x_{r1} y_{r1+1} x_{r1+2}...xk
        r2 := r1 + 1
    End
End
```

Algorithm 1: Routing algorithm at a vertex in Γ

Then we obtain,

Proposition 3. The diameter of Γ is k+1.

Note that the number of vertices n of Γ is kr^k. Assume $k = \log n / \log\log n$, then $r = (n/\frac{\log n}{\log\log n})^{1/\frac{\log n}{\log\log n}} < \log n$. According to proposition 2, the degree of a vertex is $r + k - 1 < r + \log n / \log\log n - 1 < \log n + \log n / \log\log n - 1$, thus

Proposition 4. The degree and diameter of Γ can reach $O(\log n)$ and $O(\log n / \log\log n)$ respectively.

Proposition 4 shows that Γ reaches the theoretical lower bounds proposed by paper [19].

Proposition 5. CC1 of Γ equals C_{k-1}^2 / C_{k+r-1}^2.

CC1 of graph Γ and random graph is listed in table 1 for comparison. All this graphs show the small-world phenomenon: $CC1_\Gamma$ is larger.

Table 1. Comparison of CC1 of Γ and Random Graph

Parameters of Γ	*n*	*deg*	*$CC1_\Gamma$*	*$CC1_{random}$ = deg / n*
$r_c = 8, k = 3$	1536	10	0.0222	0.0065
$r_c = 8, k = 5$	163840	12	0.0909	0.0000732
$r_c = 8, k = 7$	14680064	14	0.1648	0.000000953
$r_c = 4, k = 9$	2359296	12	0.4242	0.00000508

5.3.3. EButterfly(EBu) Protocols

Our P2P DHT uses Γ as its static graph. In this section, algorithms on how to embed peers into Γ to form a P2P DHT overlay called UBu are presented.

The Identifier Space

Every peer in EBu is identified by a unique 2-tuple $(\mathbf{c}, r)$ which is defined as

$$c \in \{(c_1 c_2 \ldots c_l *^{k-l}) \mid c_i \in Z_r\}$$
$$r \in Z_k$$

r, k and l are three integral parameters of EBu and $l<k$. In this paper, we use EBu(a, b, c) to denote that a system with parameter $r = a$, $k = b$ and $l = c$.

The identifier of EBu differs from Γ. $\mathbf{c}$ are strings that consist of "0", "1", …, "r-1" or "*". The "*" wildcard will match any character of "0", "1",…, "r-1" .

There is correspondence between the EBu topology and graph Γ: the former can be obtained by vertices merging on the latter. Therefore, a peer $(\mathbf{c}, r)$ in EBu is *responsible for* a vertices set $VSet(\mathbf{c}, r)$ in Γ, and is obtained by merging the vertices in this set. In this paper $VSet(\mathbf{c}, r)$ is called the *zone of responsibility* of peer $(\mathbf{c}, r)$. It may be expressed as $VSet(c, r) = \{(c_1 c_2 \ldots c_l *^{k-l}, r)\}$, where c must contain $c_1 c_2 \ldots c_l$ and the "*" wildcard will match any character of "0", "1",…, "r-1" .

The Topology of EBu

The topology of EBu captures the link structure of Γ. Because a peer is obtained by merging vertices in Γ, so two peers P and P′ is adjacent if there exits two vertices $(\mathbf{c}, r) \in VSet(\text{P})$ and $(\mathbf{c}', l) \in VSet(\text{P}')$, so that $(\mathbf{c}, r)$ and $(\mathbf{c}', l)$ is adjacent in Γ.

Distributing the Hash Table

In order to distribute the hash table items among peers, file names and searching keywords are hashed to a 2-tuple key (α, γ) where α is a fixed-length binary string, and $\gamma \in Z_k$. Actually, the keys' identifier space equals the vertices' identifier space in Γ. The real network location of a resource(i.e. the value of the key) with hash key (α, γ) is stored at a peer (c, γ) and (α, γ) belongs to the zone $VSet(c, \gamma)$ of responsibility of (c, γ). Here

$c_1c_2...c_l$ is a prefix of α when $c = c_1c_2...c_l *^{k-l}$.

Our performance evaluation shows that the robustness of EBu is excellent.(omitted)

6. CONCULSIONS

We know that general networks such as Internet have become more and more complex, which are constructed by many different organizations, heterogeneous, and so very non-effective. Hence one constructed the level of software over networks which are called overlay or virtual topology to improve its performance[17-24]. In this paper we have solved the following important problem: How should we design overlay (or virtual) topology for networks in order to improve network performance? We present an answer to the problem by the principle of symmetry. We show that we should use a class of overlay topology of higher symmetry such as Cayley graphs in construction of the level of software. Thus we propose a general method of constructing overlay topology by means of Cayley graph model. We explain the main conclusions of the paper by many examples in optical, wireless and peer-to-peer networks. There are many problems to need further researched such as routing and scheduling in network virtual topology design by the method of symmetry.

REFERENCES

[1] S.B. Akers, and B. Krishnamurthy, "A Group Theoretic Model for Symmetric Interconnection Networks," IEEE Trans. Computers, Vol. 38, No. 4, pp. 555-566, April 1989.

[2] F. Annexstein, M. Baumslag, and A.L. Rosenberg, "Group Action Graphs and Parallel Architectures," SIAM J. Computing, Vol. 19, No. 3, pp. 544-569, June 1990.

[3] N. Biggs, Algebraic Graph Theory, New York: Cambridge University Press, 2nd ed., 1993.

[4] M. Heydemann, "Cayley Graphs and Interconnection Networks," in Graph Symmetry: Algebraic Methods and Applications, ed. by G. Hahn and G. Sabidussi, Dordrecht: Kluwer Academic, 1997, pp. 167-224.

[5] D.M. Kwai, and B. Parhami, "Pruned Three-Dimensional Toroidal Networks," Information Processing Letters, Vol. 68, No. 4, pp. 179-183, November 1998.

[6] F.T. Leighton, Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes, San Mateo: Morgan Kaufmann, 1992.

[7] F.G. Nocetti, I. Stojmenovic, and J. Zhang, "Addressing and Routing in Hexagonal Networks with Applications for Tracking Mobile Users and Connection Rerouting in Cellular Networks," IEEE Trans. Parallel and Distributed Systems, Vol. 13, No. 9, pp. 963-971, September 2002.

[8] B. Parhami, Introduction to Parallel Processing: Algorithms and Architectures, New York: Plenum Press, 1999.

[9] B. Parhami, and D.M. Kwai, "A Unified Formulation of Honeycomb and Diamond Networks," IEEE Trans. Parallel and Distributed Systems, Vol. 12, No. 1, pp. 74-80, January 2001.

[10] B. Parhami, and D.M. Kwai, "Incomplete k-ary n-cube and Its Derivatives," J. Parallel and Distributed Computing, Vol. 64, No. 2, pp. 183-190, February 2004.

[11] B. Parhami, "Swapped interconnection networks: Topological, performance, and robustness attributes," J. Parallel and Distributed Computing, Vol.65, pp. 1443–1452, 2005.

[12] B. Stong, " On Hamiltonian cycles in Cayley Graphs of Wreath Products," Discrete Mathematics, Vol. 65, pp. 75-80, 1987.

[13] W.J. Xiao, and B. Parhami, "Some Mathematical Properties of Cayley Digraphs with Applications to Interconnection Network Design," Int'l J. Computer Mathematics, Vol. 82, No. 5, pp. 521-528, May 2005.

[14] W.J. Xiao, W.D. Chen, M.X. He, W.H. Wei and B. Parhami, " Biswapped Networks and Their Topological Properties," 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD2007), 193-198, 2007.

[15] S.M. Zhou and W.J. Xiao, "A New Family of Interconnection Networks of Fixed Degree Three," J. Computer Science and Technology, Vol. 19, No.2, pp. 218-223, 2004.

[16] W.J. Xiao and B. Parhami, "A Group Construction Method with Applications to Deriving Pruned Interconnection Networks," IEEE Trans. Parallel and Distributed Systems, Vol. 18, No. 5, pp. 637-643, May 2007.

[17] A. Kumar, S. Merugu, J. Xu, and X. Yu, "Ulysses: A robust, low-diameter, low-latency peer-to-peer network," European transaction on telecommunications, vol. 15(6), pp. 571–587, 2004.

[18] D. Malkhi, M. Naor, and D. Ratajczak, "Viceroy: A scalable and dynamic emulation of the butterfly," Proc. of ACM PODC, 2002.

[19] J. Xu, "On the fundamental tradeoffs between routing table size and network diameter in peer-to-peer networks," Proc. of IEEE Infocom 2003, vol. 1-3, pp. 2177–2187, 2003.

[20] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature, 1998.

[21] K. Aberer, L. O. Alima, A. Ghodsi, S. Girdzijauskas, S. Haridi, and M. Hauswirh, "The essence of p2p: a reference architecture for overlay networks," Fifth IEEE International Conference on Peer-to-Peer Computing, 2005.

[22] K. Sripanidkulchai, B. Maggs, and H. Zhang, "Effcient content location using interest-based locality in peer-to-peer systems," Proc. of IEEE Infocom 2003: The conference on computer communications, vol. 1-3, pp. 2166–2176, 2003.

[23] W. Xiao and B. Parhami, "Cayley graph as models of deterministic small-world networks," Information Processing Letters, vol. 97(3), pp. 115–117, 2006.

[24] C. Qu,W. Nejdl, and M. Kriesell, "Cayley dhts - a group-theoretic framework for analyzing dhts based on cayley graphs," Parallel and Distributed Processing and Applications. Second International Symposium, ISPA 2004, 2004.

[25] B. Mukherjee, Optical WDM Networks, Springer, 2006.

[26] Mohsenian-Rad, A. H. Joint logical topology design, interface assignment, channel allocation, and routing for multi-channel wireless mesh networks, IEEE Trans. Wireless Communications,Vol.6, No.12, pp.4432-4440 , 2007.

SLAVE-MODEL PROFIBUS-DP PROTOCOL AND ITS FPGA IMPLEMENTATION

Wang Bin,Cai Rui,Xu Wenbo
School of Information Technology,Jiangnan University
Wuxi, Jiangsu, 214122, China
Email: wangbin@163.com

ABSTRACT

This article introduces the Profibus-DP Protocol first, then illustrates the way of data reception and transmission of this type of bus data link protocol(Field link) and its concrete implementation based on FPGA technology in detail, and runs on Spartan-3 XC3S400 of XILINX Company successfully. The experimental result shows the design is flexible,reliable and compatible to relative accessories, and it comes to the expected results in the application of data receiving and processing and has certain value to popularize.

Keywords: Profibus-DP Protocol, Field Link, Format, FPGA , Slave-Model.

1. CURRENT SITUATION AND AIM OF THE PROJECT

Fieldbus is one of the highlights in field of industrial automation, as the highest market share fieldbus Profibus has been accepted in the world, and become an international fieldbus standard, which is widely used in processing, manufacture and construction automation, and won a great group of customersbecause of its excellent performance. Profibus is open fieldbus technology invented by joint companies such as SIEMENS , and became one of eight kinds Fieldbuses of types of the international standard IEC61158 in 1999. November 20, 2006, Profibus officially became the first China's national standards in area of fieldbus technology for industrial communications. Following with the further development of industrial automation, research on Profibus technology and its products will have a greater development prospect and be of a specially great practical significance.

As a series of Profibus, Profibus-DP is designed for the automatic control system and communications between distributed Input/Output devices and gets implementations of the fast cyclic data transmission in the distributed control system, and applications have been made in the field of industrial control in China. Profibus protocol is comparatively complex, so most of Profibus interface chips are provided by a certain number of foreign manufacturers at present in our country. Domestic research on Profibus technology is still limited in application and it is difficult for Chinese practitioners to master key technology in automation industry At the same time, FPGA (Field Programmable Gate Array) has been the rapid popularization quickly and widely applied in the fields of electronics, communications, computers and so on. Meanwhile,as the Programmable Logic Controller technology develops rapidly,famous manufacturers ,such as Xilinx and Altera, has designed many powerful, cost-effective FPGA products, making using a FPGA chip to complete the digital design of the embedded system a reality.PSystem on a Programmable Chip technology has been applied in the area of communication, electronic, computer, industrial control and other fields.

In order to explore the way of localization and Autonomation of Profibus Products in the field of automation and obtain further understanding and research of Profibus-DP technology, based on a serious analysis of bus protocol this article will introduces a design of the slave-station model Profibus-DP protocol FPGA chip.

2. THE INTRODUCTION OF PROFIBUS PROTOCOL

2.1 Brief Introduction of Profibus

Profibus protocol is German national standard DNI19245 and European standard ENSO17O fieldbus standard. According to the different applications,there are three versions of PROFlBUS protocol adapted:Profibus-DP, Profibus-FMS and Profibus-PA.The transfered data services can be acyclic or cyclic.Acyclic ones send data with or without confirmation (SDA:Send Data with Acknowledge and SDN: Send Data with No acknowledge) and send and wait data in response frame (SRD:Send and Request Data). Cyclic service sends and waits for data in a response frame (CSRD: Cyclic Send and Request Data with reply). As one important basic version, PROFIBUS-DP is applied for the high-speed data transmission among sensors on execution level,whose speeed of transmission is up to 12Mbit/s.It provides intrinsic security and power supply of the stations through the bus.Data transmission between the master station and slave stations runs by cyclic manner. In this way, the central controller communicates with the decentralized field devices (such as I/O drives , valves and etc) to implement most of the data exchange with these devices periodically.

Profibus-DP is a Master / Slave fieldbus system, in which the master station masters data communication.The devices Profibus-DP serves includes 2 categories: the master station and slave station.When the master station has right of control, there is no external request can also take send a message actively and the master station will be a PLC ,PC or etc, while slave stations are peripheral devices including input/output devices, valves, drives and measuring transmitters, etc.Slave stations are not allowed to own right of control and they are only able to receive the information from the master station or confirm request when the master station sends information to it. In multi-master station system, the token is passed between master stations in the token holding time,when the master station can require data from a slave station or send data without a remote request. Slave stations has no right of bus control, but are able to respond to a request from a master station or confirm data sent to the master station. Therefore, Profibus protocol includes token-passing protocol and master-slave protocol between the master station and slave station. Token-passing protocol ensures that each master station has enough time to achieve his task of communication, while the maximum of time the token is passed among master stations is certain, the longest cycle are the must.And the master-slave protocol ensures the communications between the master station and slave stations in the period of token holding time.

2.2 Analysis of Profibus-Dp Protocol

Profibus-DP protocol are found on ISO/OSI OSI model(ISO 7498), defines the first layer (Physical Layer, PHY), the second layer (Field Data Link layer, FDL) and the seventh layer (Application Layer). Layers from the third to the sixth is not adopted to minimize cost of system and improve efficiency. The data link layer is called Fieldbus Data link Layer(FDL), and is responsible for regulating the bus access control, data security, and transmission protocol and packet processing. For different communication requirements, FDL defines five different packet formats: fixed length frame without data field (SD1), variable-length frame with data field(SD2), fixed length frame with data fields(SD3), token frame(SD4), short response frame (SD5). For slave stations, three types are needed;fixed length frame without data field (SD1), variable-length frame with data field(SD2), fixed length frame with data fields(SD3).

Frame of fixed Length with no Data Field:

SYN	SD1	DA	SA	FC	FCS	ED

Frame with variable Data Field Length:

SYN	SD2	LE	LEr	SD2	DA	SA	FC	DATA	FCS	ED

Length with Data Field:

SYN	SD3	DA	SA	FC	DATA-UNIT	FCS	ED

Token Frame:

SD4	DA	SA

Short Acknowledgement Frame:

SC

where:

SYN：Synchronization Period, a minimum of 33 line idle bits

SD2：Start Delimiter, value: 68H

LE：Octet Length, allowed values from 4 to 249

LEr Octet Length repeated

DA：Destination Address

SA：Source Address

FC：Frame Control

DATA_UNIT Data Field, variable Length (L-3), max: 246 octets

FCS：Frame Check Sequence

ED：End Delimiter, value: 16H

L：Information Field Length,variable number of octets: L = 4 to 249

These Profibus-DP frames are formed by characters consisting of 11 bits shown in the following figure. The least significant bit corresponds to the start bit (ST) and its value is always 0. Next, there are eight data bits followed by an even parity bit (P). And finally the stop bit (SP) that is always 1.

3. FPGA IMPLEMENTATION OF SLAVE-MODEL PROFIBUS-DP

In the design of slave-mode Profibus-DP fieldbus chip, according to the international fieldbus standards strictly, we will adopt the module-design method,satisfy with function and requirements of system while choosing the sophisticated technology and devices as possible in order to ensure system reliability, improve system performance and reconfigurability.Based on the above design principles, the circuit design of Profibus-DP fieldbus chip have been decided to adopt the FPGA device that is fairly sophisticated by now. Implementation using FPGA has merits such as low cost, easy development,high reliability,reconfigurability and etc.

This design uses a top-down (Top - Down) technical Design way,based on an analysis of the results of division of function modules and the design features of the object for analysis, the description of circuit module is done to implement the function of circuit in Verilog language,and functional simulation is finished.then. After synthesis and optimization of all function modules, the implementation is achieved using FPGA chip.

Based on the foregoing analysis, this design can be divided into the following six functional modules: watchdog module; UART transceiver module; interrupt control module; FDL data receiving logic module; FDL data transmitting logic module and user interface module as follows:

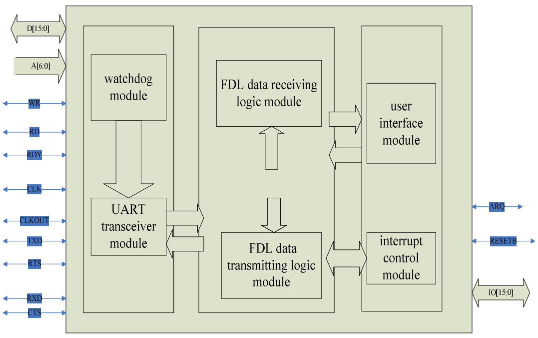


Figure 1. The Whole Structure of System

3.1 Watchdog Module

Watchdog module consists of the Baud rate automatically-detecter module and the Baud rate generator module. It supplies chip for various modules to provide the clock. Profibus-DP supports more than 10 levels of Baud rates including: 9.6Kbps, 45.45Kbps, 93.75Kbps, 187.5Kpbs, 500Kbps, 1.5Mbps, 3Mbps, 6Mbps and 12Mbps.The baud rate of communication in the same network must be unique.In order to avoid errors in communication caused in the design,the baud rate generating module module is needed to detect data on bus. Thereby,the baud rate is determined, and the baud rate generator module is triggered then.

The Baud rate detecting course is as follows: the initial baud rate is assumed equal to the highest baud rate at the beginning of inquiry. Through detecting time,if no SD1, SD2 or SD3 frames are received and there are no errors, the baud rate is adjusted equal to the next level of Baud rate, until the correct baud rate is comfirmed. Once the correct baud rate is assured, the baud rate generator module starts to provide baud rate to other modules and monitor the baud rate. After receiving the correct message the Baud rate automatically-detecter module is reset and re-search for the baud rate.

3.2 Uart Transceiver Module

UART transceiver module and watchdog module work together in the implementation on functions of Profibus physical layer protocol: after the baud rate is detected,conversion between serial and parallel data works and parity is done to ensure the reliability and correctness of data transmissions.

3.3 Fdl Data Receiving Logic Module

FDL data receiving logic module congsists of start frame detector module, end frame detector module, frame receiving logic control module , receiving frame shift register, and receiving data FIFO register. Modular design of the structure as follows:

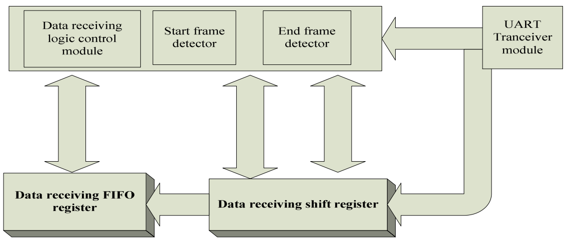


Figure 2. The Structure of FDL Data Receiving Logic Module

The function of FDL data receiving logic module is to receive data sent by the UART transceiver module and do analysis and treatment in accordance with the requirements of the protocol specification for SDN and SRD Services . The course of work is: First, start frame detector module analyzes start frame of message to identify the received message type, then in the assistance of receiving data FIFO register and receiving frame shift register ,frame receiving logic control module according to protocol specification to treats different types of message. This process runs until the termination of end frame detector module detectes the end frame. The process is shown as follows:

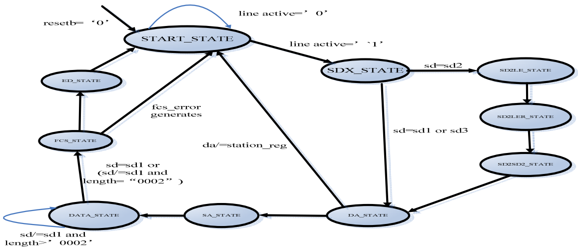


Figure 3. The Work Flow Diagram of FDL Data Receiving Logic Module

3.4 Fdl Data Transmitting Logic Module

FDL data tranmitting logic module consists of data tranmitting logic control module, data transmitting shift register as well as data tranmitting FIFO register module. Modular design of the structure as follows:

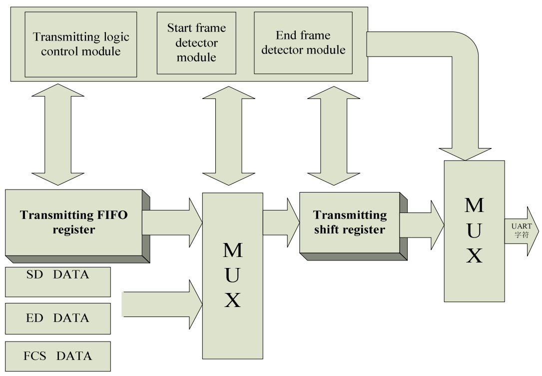


Figure 4. The Structure of FDL Data Transmitting Logic Module

The function of FDL data tranmitting logic module is to analyze and process data in accordance with the requirements of the protocol and generate a message in line with the norms, and then send processed message to the UART transceiver module. The course of their work is: First, the start frame detector module analyses the message at the start frame to carry out analysis to determine the type of message, and then in assistance of the data transmitting shift register and data tranmitting FIFO register , data tranmitting logic control module frames sent by the logic control module according to protocol specifications treat with the message ready to send, and finally is sent to the UART transceiver modules. The process is shown as follows:

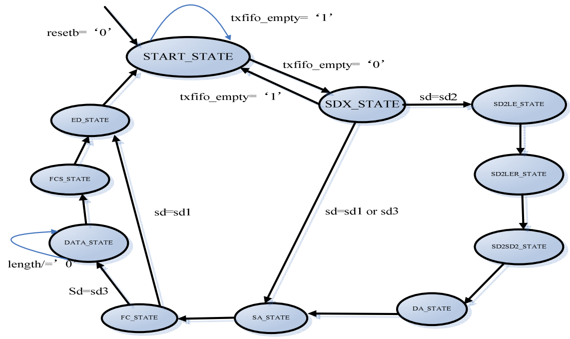


Figure 5. The Work Flow Diagram of FDL Data Transmitting Logic Module

3.5 Interrupt Control Module

Interrupt control module is to receive interrupt request signals from other modules (such as the overflow interrupt signal from FDL data transmitting FIFO register,the overflow interrupt signal from FDL data receiving FIFO register and FCS error interrupt signal),respond to these signals and form flexibly configurable interrupt control logic. Interrupt control module consists of interrupt request register, interrupt Mask Register, interrupt status register and interrupt handling module.

3.6 User Interface Module

Userinterface module is responsible for the data communication between chip and local MCU aiming to receive and transmit data with field data link as well as send message such as overflow signal from watchdog or baud rate searched signal to local MCU in real-time. The way of communication between the chip and MCU is: after chip receives a message in line with the norms of protocol , the chip will gives notice to the local MCU;then MCU get the message from data transmitting FIFO register and do response according to information revealed in message, data prepared to be sent will be sent to data transmitting FIFO register and send out of chip as a bit stream ultimately. User interface module consists of FC (Function Code) decoder, address recognition module, FCS (Frame Check Sequence) generator / detector, length of data transmit/ received counter,logic comparison of length of data transmit/received module. Address recognition module is responsible for detection of the received message's address characters (SA) and destination address characters (DA) as well as generation of address characters (SA) and destination address characters (DA) when a message is to be transmit. length of data transmit/ received counter and logic comparison of length of data transmit/received module in accordance are responsible for controlling the length of the message with the protocol norms and the formation of message in line with the protocol norms. FCS generator is responsible for generation of FCS characters in line with protocol norms. FCS detector is responsible for detedtion of FCS characters in a received message.

4. SIMULATION AND VERIFICATION OF SYSTEM

This design uses the Verilog hardware description language, after functional simulation and timing simulation on ModelSim simulation software, result shows that the chip can implement the protocol function. To further verify the correctness of the design,verification environment is designed as follows: in a platform of XILINX's Spartan-3 XC3S400,the maximum chip operating frequency is up to 320MHZ which can fully satisfy the requirement of 16 times of highest baud rate of the protocol(12Mbps). Cygnal company's product C8032 is used as a local MCU while users layer of the PROFIBUS-DP protocol is compiled in JAVA language completes the corresponding implementation of PROFIBUS-DP protocol combined with the design of Profibus-DP protocol chip.

5. CONCLUSIONS

This paper introduces data communication bus for fieldbus communication equipments :Profibus-DP and designs Profibus-DP bus chip in accordance with the actual needs and requirement of protocol norms ;terminal simulation is done finally.Application of FPGA simplifies the system architecture and shortens the design cycle, improves the system performance and reconfigurability. After debugging =,result shows the board functions properly,works stably and is of full competence in data processing applications.

REFERENCES

[1] Gyu-Sung Bae,Jong-Bae Kim,Byoung-Wook Choi and, Kye-Young Lim, “Implemetation of a Fieldbus System Based on Profibus-DP Protocol”.Journal of Control, Automation and System Engineering, No10, October, 2000

[2] PROFIBUS Specification, Normative parts of Peodivus -FMS- DP- PA according to the europeanstandard EN 50170. Edition 1.0, March 1998

[3] Zhu Bo, Xia Ji-qiang, Zhou Li-ping, Man Qing-feng, “The Design of PROFIBUS- DP Slave Station Link Layer Controller based on FPGA”,2007

[4] Joyce Van de Vegte, Fundamentals of Digital Signal Processing, Beijing: Electronics Industry Publishing House,2004.

[5] Emmanuel C. ifeachor,Barrie W.Jervis, Digital Signal Processing A Practical Approach, Beijing:Electronics Industry Publishing House,2004.

[6] Liang Ruiyu,Jiang Bing, “The Structure Design of An Efficient IIR Filter and Its FPGA Implementation”, JOURNAL OF HOHAI UNIVERSITY CHANGZHOU, 19 (3),2005,pp.39-41

[7] Zhang Chi, Guo Lili, Sun Yan, “A high speed FIR filter design based on MATLAB and FPGA device”, Information Technology,7,2006,pp.31-33.

AN ENERGY EFFICIENT CLUSTERING PROTOCOL FOR WIRELESS SENSOR NETWORKS *

Zhou Zeshun, Li Layuan, Xu Yi , Gao Yu
School of Computer Science and Technology, Wuhan University of Technology
Wuhan China

ABSTRACT

Wireless sensor networks will enable the reliable monitoring of a variety of environments such as battlefields and disaster fields. Traditional routing protocols do not take into account that a node contains only a limited energy supply. Optimal routing tries to maximize the duration over which the sensing task can be performed. Sensors are typically disposable and expected to last until their energy drains due to their battery-constrained power. The energy constrained and limited computing resources of the sensor nodes present major challenges in gathering data. In this paper, we propose an Energy Efficient Clustering Protocol (EECP) based on the distance from the sink in wireless sensor networks. We use the distance vector from the sink, which affects the energy depletion of the network. The simulation results show that the EECP reduces energy and the end-to-end delay is also improved significantly.

Keywords: Wireless sensor networks, EECP, Improvement

1. INTRODUCTION

Wireless Sensor Networks (WSNs) are multi-hop ad hoc network of hundreds or thousands of sensor devices. The sensor nodes collect useful information such as sound, temperature, and light; moreover, they play a role as the router by communicating through wireless channels under battery constraints. Such a network normally consists of a large number of distributed nodes that organize themselves into a multi-hop wireless network. Each node has one or more sensors, embedded processors and low-power radios, and is normally battery operated. Typically, these nodes coordinate to perform a common task. To design a good protocol for the wireless sensor networks, we have firstly considered the energy efficiency. Sensor nodes are likely to be battery powered, and it is often very difficult to change or recharge batteries for these nodes. In fact, someday we expect some nodes to be cheap enough that they are discarded rather than recharged. The second important factor is the scalability to the change in network size, node density and topology. Some nodes may die over time; some new nodes may join later; some nodes may move to different locations. The network topology changes over time as well due to many reasons. A good protocol should easily accommodate such network changes. The third important factors include throughput and bandwidth utilization. Various issues in the design of wireless sensor networks, design of low-power signal processing architectures, low power sensing interfaces, energy efficient wireless media access control and routing protocols, low-power security protocols and key management architectures have been areas of extensive research in recent years. A lot of clustering algorithms have been proposed for wireless sensor networks. Grouping a large number of sensors into clusters and keeping them communicate regularly are quite complex. In [1], Heinzelman et al. developed and analyzed LEACH, an application specific protocol architecture for microsensor networks. LEACH divides time into rounds. Clusters are organized at the beginning of each round and data are transferred from the nodes to the cluster head and on to the base station after the set-up phase. As LEACH is such a typical clustering protocol, several modifications have been made [2, 3]. LEACH [2] is tried to preserve and balance the energy dissipation of the network using cluster-based architecture, thus prolong the network lifetime.

This paper presents a new protocol explicitly designed for wireless sensor networks. While reducing energy consumption is the primary goal in our design, our protocol also has good scalability and collision avoidance capability. It achieves good scalability and collision avoidance by utilizing a combined scheduling and contention scheme. To achieve the primary goal of energy efficiency, we need to identify what are the main sources that cause inefficient use of energy as well as what trade-offs we can make to reduce energy consumption.

The paper is organized as follows. Section II provides the system model. Section III gives the core algorithm and the Implement of EECP. Section IV gives Correctness and Complexity Analysis. Simulations are given in Section V and Section VI concludes the paper.

2. THE SYSTEM MODEL

A sensor network consists of a large number of sensors distributed over an area of interest. There is a command node in charge of the networks mission. In our work, we have focused on maximizing the life span of the entire network, rather than individual nodes. To achieve this goal, we use the advantages of clustering and self-organizing methods in our research work coupled with routing mechanisms that strive to achieve equalization in the available power throughout the network. Figure 1 depicts a sensor network as a conceptual model. Clustering can be defined as the partition of the sensor network nodes into several groups, where each group will have one elected node to be the cluster-head, and the remaining nodes will carry out the sensing. The cluster-head will act as a local controller for the traffic between the cluster and the base station or the sink. Thus the lifetime will be strongly related to the failure of the cluster-head. We also rely on autonomous routing distribution and self-organization in order to minimize the overheads associated with routing protocols. In order to prolong the lifetime of the whole network, a more rational

* Foundation items: Supported by the National Natural and Hubei Science Foundation (60672137, 2008CDB347), and the Specialized Research Fund for the Doctoral Program of Higher Education of China (20060497015, 200804971030)

distribution of clusters should he taken of great consideration.

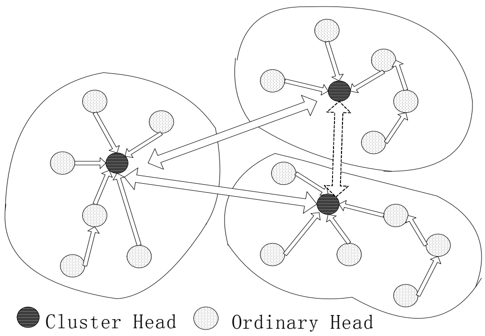


Figure 1 Wireless Sensor Networks

3. EECP

3.1 Overview of EECP

In EECP, multi-hop communication is expected to consume less power than single hop communication in sensor networks. This is because even though more nodes are involved in the transmission of each packet, the transmission energy increases with the square of the distance with omni-radial transmissions. In addition, the coverage areas of sensor networks can very well be so large that the wireless interfaces are unable to reach the sink from any point in the network, and multi-hop becomes necessary. However, there is a reason for using single hop communication. Consider the nodes closest to the sink. Regardless of routing scheme in the multi-hop network, these nodes will always be involved in the forwarding of traffic and therefore, they will drain faster than the other nodes in the network. In order to further increase the life-span it could therefore be useful to allow nodes further away from the sink to bypass these nodes by increasing their transmission power and reach the sink directly. We intend to extend out modeling with such a mechanism in future work.

According to [4], all the member nodes are within one hop away from their corresponding cluster head, the free space (d^2) power loss model will be used. To transmit a *m*- bit packet over a distance *d*, the radio expends:

$$E_{tx}(m,d) = E_{elec} \cdot m + \varepsilon_{fs} \cdot m \cdot d^2$$

On the other hand, to receive this message, the radio expends:

$$E_{rx}(m) = E_{elec} \cdot m$$

As seen above, since the energy cost consists of two components, one for transmitting or receiving a packet in the radio electronics and the other depending on the distance in transmissions, the routing protocols must minimize the number of data receptions and while minimizing the transmit distance for a specific node.

In this paper, we propose a routing protocol to deliver data packets collected from sensor nodes to the base station, which maximizes the lifetime of the sensor network. In order to maximize the network lifetime, the total energy should be minimized in a round of communication, while balancing the energy consumption among the nodes. An ideal aggregation tree can be thought as a balanced tree with a minimum number of edges and the same number of child nodes in order to maximize the network lifetime.

3.2 The Selection of Cluster-Head

In EECP, Each sensor in the network becomes a cluster head (CH) with probability *p* and advertises itself as a cluster head to the sensors within its radio range. This advertisement is forwarded to all the sensors that are no more than *k* hops away from the cluster head. Any sensor that is neither a cluster head nor has joined any cluster itself becomes a cluster head. If a sensor does not receive a CH advertisement within time duration *t* (where *t* units is the time required for data to reach the cluster head from any sensor *k* hops away) it can infer that it is not within *k* hops of any cluster head. Moreover, since all the sensors within a cluster are at most *k* hops away from the cluster-head, the cluster head can transmit the aggregated information to the processing center after every *t* units of time. This limit on the number of hops thus allows the cluster-heads to schedule their transmissions. The energy used in the network for the information gathered by the sensors to reach the processing center will depend on the parameters *p* and *k* of our algorithm. Since the objective of our work is to organize the sensors in clusters to minimize this energy consumption, we need to find the values of the parameters *p* and *k* of our algorithm that would ensure minimization of energy consumption. We derive expressions for optimal values of *p* and *k* in the next subsection [5].

3.3 The Implement of EECP

To determine the optimal parameters for the algorithm described above, we make the following definition:

Definition 1: All sensors transmit at the same power level and hence have the same radio range. Data exchanged between two communicating sensors not each others radio range is forwarded by other sensors.

Definition 2: A distance of *D* is the distance from the sink and smaller index means closer to the sink.

Definition 3: Each sensor uses 1 unit of energy to transmit or receive 1 unit of data. When a sensor communicates data to another sensor, only the sensors on the routing path forward the data.

The basic idea of the derivation of the optimal parameter values is to define a function for the energy used in the network to communicate information to the information-processing center and then find the values of parameters that would minimize it.

The operation of EECP is broken up into several rounds, where each of them begins with a cluster set up phase, which is followed by a steady state phase, when data transfers to the sink occur. In order to minimize overhead, the latter phase need to be longer than the former phase. In more detail, a cluster set up phase is divided into three sub-phases:

(1) Advertisement: The self selected cluster heads broadcast advertisement messages in their clusters, and then the sensor nodes decide which clusters they belong to based on the received signal strength. Initially, when clusters are being created, each node decides whether or not to become a cluster-head for the current round. This decision is based on the suggested percentage of cluster heads for the network and the number of times the node has been a cluster-head so far. This decision is made by the node n choosing a random number between 0 and 1. If the number is less than a threshold T(n), the node becomes a cluster-head for the current round[2]. The threshold is set as:

if $n \in G$ then $T(n) = P/1 - P(r \bmod \frac{1}{P})$, where *P* = the desired percentage of cluster heads, *r* = the current round, and *G* is the set of nodes that have not been cluster-heads in the last 1/P rounds. Using this threshold, each node will be a cluster-head at some point within 1/P rounds.

(2) Clusters Set-up: Each node informs its cluster head its decision using CSMA MAC protocol. During this phase, all cluster head keep their receivers on. The unbalance of energy

depletion is caused by different distance from the sink. We improve the probabilistic fair clustering which is motivated from the LEACH protocol. We divide the whole network into a few group based on the distance from the sink and the strategy of routing. Each group has their own number of cluster number and member node. Then, we apply the scheme to clustering in terms of the routing strategy [6].

$$g = g * \frac{D_{n-j+1}}{\sum_{j=1}^{n} D_j} (1 \le j \le n)$$

Where g is the number of cluster in each group.

$D_j = \sqrt{(x_j - x_{jd})^2 + (y_j - y_{jd})^2}$ is the distance from each node to the ideal cluster head, where x_{jd} and y_{jd} represent the coordinate of the optimal position where the ideal cluster head locates. If minimizing the summation of the distance D_j, we can get the smallest value of the total power to minimize the energy dissipation each round. Here, we treat the whole cluster from the sink and smaller index means closer to the sink.

Based on the geometry knowledge, the ideal cluster head position is the mass center of the cluster. After the cluster head collect all the coordinate information of the member nodes, it can select the cluster head candidate for the next turn, which is the closest to the mass center. Then pass its cluster head authority to the candidate. If itself is the closest to the mass center, it remains to be the cluster head.

(3) Schedule Creation: The most direct factor that the clustering protocol can affect energy is the distance between the nodes to the cluster head. The less the distance is, the more energy can be saved during the data transmission. Based on the number of nodes in the cluster, the cluster heads create TDMA schedules and broadcast them back to their members. The distribution of the network that is controlled by the self-organized protocol is not likely to be very uniform and rational. Some nodes might be locally huddled, while some might be isolated. We design a merge algorithm to uniform the clusters distribution. Each cluster head tries to listen the timely message from other cluster heads. Based on the coordinate information provided by the nodes that send the message, the cluster head can calculate the distance from itself and the other cluster head. If such distance is below a certain threshold, these two clusters will merge into one cluster. Likewise, if the member node receives a message from a closer cluster head, it will try to leave the present cluster and join the closer cluster head.

In the word, we choose a cluster head to manage the communication schedule, merge and partition the clusters. There are geographical clusters right after deploying sensor network and then clusters are merged and partitioned based on data operating. Data-centric cluster scheme can achieve the accuracy of aggregated data and thus make user to become aware the situation of sensor field.

4. CORRECTNESS AND COMPLEXITY ANALYSIS

4.1 Correctness Analysis

Lemma1: In EECP, the probability that two nodes within each other's cluster range are both cluster heads is small, i.e., cluster heads are well distributed.

Proof. Consider the following worst case. Assume that *v*1 and *v*2 are two isolated neighboring nodes. In the worst case, neither of the two nodes decides to be a cluster head. Otherwise, one of them will concede to the other.

Lemma2: In EECP, a node is either a cluster head or a regular node that belongs to a cluster.

Proof. Assume that a node terminates its execution of EECP without electing to become a cluster head or joining a cluster. In this case, the node will become a cluster head, which is a contradiction. To prolong the sensor network lifetime, cluster head selection is primarily based on the residual energy of each node.

4.2 Complexity Analysis

The computational complexity of the proposed protocol is analyzed below [7]. To calculate its working schedule, a sensor first looks at its neighbor table and extracts reference times of its neighbors. Suppose that a node has at most n neighbors. Sorting these reference times takes time $O(n \log n)$. The maximum number of intersection points covered by a sensor is $O(n^2)$. For each intersection point, a sensor has to find out which nodes cover the point, which takes time $O(n)$. So, the calculation of working schedule for all intersection points takes time $O(n^3)$. Therefore, a complexity of $O(n^3)$ is incurred on each node to decide its working schedule.

5. EXPERIMENTAL RESULTS

We use the network simulator OPNET to model our protocol. OPNET provides a fairly realistic simulation environment for WSN among the available network simulators. In order to compare with original LEACH we have also built a model for LEACH using OPNET and used the same power model for both models to evaluate and compare their energy efficiency. Our simulation is based on a network with 40 nodes distributed in a 1km*1km area [8].

We compare our approach with LEACH on network life time and total energy dissipated in system.

As we can see from Figure 2 and 3 the average energy dissipated energy is not significant improved when the network size is 40 and 80. The reason is that the network range is 5 times larger than the node's radius. When the network size rises to 200, we can see the average energy drop from 0.061 to 0.051, and 0.031 to 0.21, meaning that nodes are employed denser, the improvement is more significant. Figure 3 shows with two sources to send data, the latter starts at 10th second.

6. CONCLUSIONS

In this paper, we have presented a novel network protocol for WSNs and compared it to the LEACH protocol. Results from our simulations show that our protocol provides better performance for energy efficiency and network lifetime.

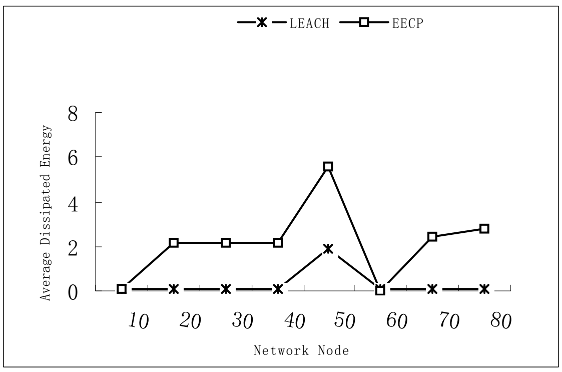


Figure 2 Average Dissipated Energy with One Source

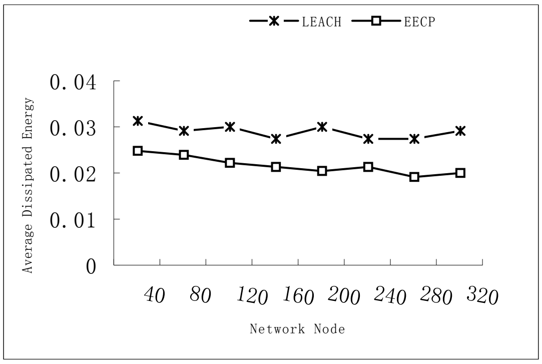


Figure 3 Average Dissipated Energy with Two Source

REFERENCES

[1] Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An application specific protocol architecture for wireless microsensor networks. IEEE Trans. Wireless Comm.**1** (2002) 660–670

[2] Xiang Hong; Qilian Liang.An access-based energy efficient clustering protocol for ad hoc wireless sensor network. Personal, Indoor and Mobile Radio Communications, 2004. PIMRC 2004. 15th IEEE International Symposium on. Volume 2,5-8 Sept. 2004:1022 - 1026 Vol.2

[3] Pham, T., Kim, E.J., Moh, M.: On data aggregation quality and energy e.ciency of wireless sensor network protocols - extended summary. In: Proc. BROADNETS 2004. 730–732

[4] W.B.Heinzelman,A.P.Chandrakasan.Anapplication specific protocol architecture for wireless microsensor networks. IEEE Transaction on Wireless Communications, vol1,no. 4, Oct.2002,pp.660-670

[5] Seema Bandyopadhyay; Coyle, E.J.; An energy efficient hierarchical clustering algorithm for wireless sensor networks. INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE. April 2003 3(3):1713 - 1723

[6] SangHak Lee; JuneJae Yoo; TaeChoong Chung; Distance-based energy efficient clustering for wireless sensor networks. Local Computer Networks, 2004. 29th Annual IEEE International Conference on. 2004:567 – 568

[7] Chi-Fu Huang; Li-Chu Lo; Yu-Chee Tseng; Wen-Tsuen Chen; Decentralized energy-conserving and coverage preserving protocols for wireless sensor networks. Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on.2005 (1):640 – 643

[8] Hongju Cheng; Xiaohua Jia; An energy efficient routing algorithm for wireless sensor networks. Wireless Communications, Networking and Mobile Computing, 2005. Proceedings. 2005 International Conference on. 2005,2:905 - 910

THE STUDY OF TD-SCDMA NETWORK ACCESS AND OPTIMIZATION SCHEME

Xu-Qing Chai [1], Qing-Ping Guo [2]
[1] Network Center, Henan Normal University, Xinxiang, Henan 453007, China
[2]School of Computer Science and Technology, WUT, Wuhan,Hubei 430063, China
Email: cxq@htu.cn

ABSTRACT

The paper summarizes the access problems that was encountered in the process of optimizing TD-SCDMA 3G network, in light of the physical layer process, analyzes the access problems, gives out the analysis of ideas and solution to the problem which was encountered in the actual engineering project.

Keywords: TD-SCDMA, Network Physical Layer Process, Access Procedure

1. INTRODUCTION

TD-SCDMA (Time Division - Synchronous Code Division Multiple Access) system is the 3G communication network which is the independent intellectual property rights China owns now, has been in the country for large-scale network. Because the TD-SCDMA industry chain is still immature, there will be some unexpected problems in the process of building networks. Therefore, the need for network optimization to ensure TD-SCDMA network to run smoothly is essential.
In the network optimization, to analyze the availability of the network and the access performance are very important. The success rate of access will directly affect the users' intuitive feelings of the services provided. From the view of system, the success rate of access has a great relation with the working course of the system physical layer. Therefore, it is very important to analyze the physical layer process about the access performance from the physical layer of the TD-SCDMA system.

2. THE PHYSICAL LAYER PROCESS OF THE TD-SCDMA NETWORK AND RELEVANT PARAMETERS

2.1 Physical Layer Process

Physical layer process means some acts about physical layer which occurred between the UE (User Equitment) and the NodeB. Mainly includes: cell search, random access process, synchronization process. In these acts of physical layer, the cell search, synchronization process is the basis of access; random access process is involved in the specific implementation. Here's a briefly introduction.

2.1.1 Cell Search

When the UE boot, NAS (Non Access Stratum) layer will be initiated at the request of PLMN (Public Land Mobile Network), the aim of the choice is to select an available, the best PLMN. The selection of PLMN has two modes: automatic, manual. Automatic selection means that UE automatically selects a PLMN according to the priority list of PLMN, then reports to the NAS layer. Manual selection means all available networks currently are displayed to users; the user can select a PLMN.
Whether manual selection or automatically selection, PLMN selection process is the same.
For TD-SCDMA system, UE first tests the carrier frequency of the broadband power in the TDD (Time Division Duplex) frequency band, (broadband power means the broadband power is measured directly from the end of RF connector or the output of pulse wave Oscilloscope, the basic TDD band is 2010 ~ 2025MHz, according to the bandwidth of 1.6MHz, the band can accommodate a total of nine frequency points.) , sequences by the strength of the measuring broadband power, and then from the strongest power of frequency points to start the cell search. In the TD-SCDMA mobile communication system, the unique structure of the frame determines the terminal of cell search algorithm. According to the realization of the process, it can be divided into the following 6 steps:

(1) To use of power characteristic, identify the location of pilot, reach to the sync symbols;
After electrified, the terminal read channel number (is frequency point), at the specified frequency point which TD-SCDMA works to find the strongest DwPTS (Downlink Pilot Time Slot) of the approximate location, in order to establish downlink synchronization, synchronous accuracy at symbol-level.

(2) The use of SYNC-DL (Synchronization Channel Downlink) as related to identification of the cell SYNC-DL, to reach CHIP (chip) level synchronization;
DwPTS in the pilot position is largely determined on the basis of the accurate determination of the use of correlation SYNC and SYNC code group position, down to the chip-level synchronization accuracy.

(3) Coarse tune frequency complete;
In the end system design, consider the frequency of the base station carrier frequency deviation, base stations, mobile station relative movement of the Doppler frequency shift, and frequency of quasi-stable device, the total of the maximum () of the error, the need for frequency adjustment . Step-by-step use of the SYNC-DL code offset different section of different methods used for the relevant frequency offset estimate is calculated, and then determine the amount of feedback to the hardware, step by step in order to reduce the carrier frequency offset, this step is complete, carrier adjusted to the offset below 1KHz.

(4) Under the impulse response to identify MIDAMBLE code (intermediate code);
SYNC location after the code group to determine, in accordance with SYNC code and the corresponding relationship between the midamble code, then determine the basic midamble for one of four yards. Four yards with a rotational basis with the received data segment midamble channel impulse response so estimates comparing the 4-channel impulse response value, when the correct pattern, the channel has been the biggest shock value, which found

the district's basic midamble code.

(5) Solution P-CCPCH (Primary Common Control Physical Channel), for the frequency of fine-tuning, that is, to achieve synchronization 1/8CHIP level;
This step has been identified by the use of fine-tuning code Midamble frequency offset, precision partial FM seeks to around 1kHz carrier frequency offset was adjusted to below 200Hz. Joint detection-based method to correct the smaller frequency difference to the base-band demodulator tolerable limits. In TD-SCDMA system, joint detection (JD) techniques, each frame in training sequences add Midamble, can be used to estimate the actual channel response. In this way, terminal joint detection techniques can be used to eliminate multipath and multiple access interference, and the middle of training sequence code demodulator (Midamble) data symbols near the (Data Symbol), the use of these symbols in the frequency difference contains information to guide the hardware automatic frequency control (AFC) to adjust.

(6) To determine the phase information, read the BCH (Broadcast Channel) channel information, the completion of the search area.
Broadcasting information mapping system in the P-CCPCH physical channel, in order to facilitate identification of terminals of the P-CCPCH frame intertwined its starting point, TD-SCDMA system to a different phase modulation DWPTS slot synchronization of SYNC-DL code. In the QPSK modulation mode, the phase modulation to a total of four: 45,135,225,315, will be four sub-frame continuous phase modulation combined with a phase sequence can be as long as found in the physical channel P-PCCPCH sign of the phase sequence, but also found on the P-CCPCH. According to the P-CCPCH system information integrity, UE decide whether the district area as a service.

2.1.2 Random Access Procedure

Random access process, started at UE uplink synchronization, the uplink synchronization is behind the success of the foundation initiated signaling connection; UE from which to access the area used by 8 of the SYNC-UL may randomly select a code, physical channel in UpPTS send it to the base station. UE and then to determine the launch time and UpPTS power (open-loop process) in order to launch UpPTS selected physical channel signature. Once the Node-B to detect the UpPTS information from the UE, it arrived in time and to receive power will know. Node-B transmitting power to determine from time to time to update and adjust instruction, and in subsequent sub-frame 4 through F-PACH (in a burst / sub-frame message) send it to the UE. Once when the UE from the selected F-PACH (with selected signature corresponding to the F-PACH) in the control of the receipt of such information, show that the Node-B has received UpPTS sequence. And then, UE will adjust the launch time and power, and to ensure that in the next two after that corresponds to the F-PACH Send PPACH channel RACH. In this step, UE is sent to the Node-B of the RACH will be synchronized with higher accuracy. After, UE will be in the corresponding S-CCPCH of the FACH channel to receive the response from the network, the instructions issued by UE random access has been received, if received, will be allocated in the network of UL and DL dedicated channel through FACH to establish links from top to bottom line. In the use of allocated resources prior to send a message, UE can send UpPTS and wait for the second F-PACH from the response, which could be the next phase of the transmitter power and the update command SS. Next, the base station in the conveyor channel FACH channel allocation information is news base stations, and signaling between the UE and business interaction information.

2.1.3 Synchronization

Synchronous communication system is the basis of ability to work properly. TD-SCDMA systems in various different aspects of the synchronization process, all in the data exchange between devices must be synchronized before; and uplink synchronization is the unique feature of TD system and key technologies; uplink synchronization is the TD-SCDMA one of the key technologies, the quality of the performance of uplink synchronization is directly related to the quality of the whole system performance. For TD-SCDMA, UE sent Burst signal is different between the UE is the use of OVSF (Orthogonal Variable Spreading Factor) codes to distinguish between the. Taking into account the orthogonality of the OVSF code, if it is not considered multi-path delay, if a different UE at the same time slot sent Burst signal receiver in the base station to maintain synchronization between the UE is different without mutual interference. This can take full advantage of Road Resources Code to increase the system capacity. If you can not maintain synchronization between the UE, it will result in interference between the UE and lower system capacity. Midamble code base station can be in the same slot for different UE channel estimation, channel impulse response according to different estimates of the timing offset information UE. Then, the next available time slot in the downlink, base stations will be offset through synchronous (synchronization shift) instruction, respectively, adjusted for each UE to send their own time, to ensure that different from the same time slot of the uplink synchronization between the UE.

2.2 Physical Layer Parameters of Wireless Access

UE access network in the process, mainly based on physical layer measurements, synchronization, random access; and physical layer measurement parameters, access parameters are able to access the normal course of the guarantee; in the system design, network optimization the course of the design of physical layer parameters, set on a crucial role. Physical layer parameters of wireless access are:

(1) The Guard Period of wireless links simultaneous
Receive from the physical layer of the "In Sync" the success of the number of simultaneous instructions. Node B uses this parameter to access / re-synchronize the uplink interface Uu.

Node B to use this IE to obtain / re-synchronize the uplink interface Uu. The parameter values will affect the Node B to determine the criteria for recovery in RL, if inappropriate; Node B will not be able to detect the RL Restore, or RL Restore frequent, resulting in abnormalities.

(2) PRXUpPCHdes (base station uplink pilot channel power of hope to receive)
UE according to the system to receive broadcast messages, as well as PRXUpPCHdes configuration path loss calculated by the ultimate power to launch UpPCH

(3) Step power Gradeability
Gradeability power step, the previous uplink synchronization fail, UE according to this parameter to increase power, re-launched UpPCH.

(4) The largest number of simultaneous attempts to the maximum number of synchronization attempts. When the

maximum time synchronization attempt failed, the physical layer informs the MAC layer. MAC layer based on the value of Mmax decides whether to continue to access. Therefore, the whole process could be launched access code uplink synchronization for the total number Mmax * Max SYNC_UL Transmission value

(5)T300
UE to the UTRAN when the RRC CONNECTION REQUEST message sent after the T300 to start; UE in the downlink received RRC CONNECTION SETUP message termination timer T300.

If the T300 and the UE re-issued overtime counter V300 ≤ N300, UE will be re-issued RRC CONNECTION REQUEST message, or else turn into idle mode UE.

T300 settings should be slightly larger than under normal circumstances the process of establishing the RRC connection (UE from the RRC CONNECTION REQUEST sent to the RRC CONNECTION SETUP or received RRC CONNECTION REJECT) Length.

(6)Activation time
RNC to the UE re-configured to send messages (for switching or adjustment of district, or business establishment, etc.), according to this parameter to determine the values of activation time.

(7) The initial signal to noise ratio
The target signal to noise ratio of the upstream business. Closed-loop reactive power control for uplink at the time of the target signal to noise ratio, business process can ring up to change the reactive power control.

3. TD-SCDMA NETWORK ACCESS PROBLEMS AND THE REASONS FOR COMMON ANALYSIS

TD-SCDMA network, which has passed the road test data analysis software, correct analysis of test data to determine the occurrence of the Call Fail time, and before and after the Call Fail Pilot Scanner collected information, phone the information collected in adjacent areas, as well as signaling processes. it Collected Information through the alignment of the signaling cell phone and RNC (Radio Network Controller) of the single-user time tracking to determine the time point, combined with the single-user tracking RNC and UE signaling processes, determine which lead to Department abnormal failure. Then TD-SCDMA network is in accordance with the follow-up analysis of the various sub-processes and solve problems, including the issue of paging, RRC establishment of the issue, the question of encrypted authentication, RAB and RB establishment of the issue, equipment abnormalities, network coverage issues.

3.1 Paging Problem

Paging performance issues are generally as follows: Calling completed RAB (Radio Access Bearer) assignment as well as the CC Setup (Call Control Setup), waiting for news Alerting receive CN (Core Network) sent a message of Disconnect Direct. From the called UE signaling processes in general we can not see the anomaly, but there have been news UE has not received Paging initiated the establishment of RRC connection request. From the RNC called single-user tracking we can see that CN received news Paging issued, but there is no follow-up information.

3.2 RRC (Radio Resource Control) to Establish the Issue of

(1) Specific analysis of the process of being frequently asked questions are as follows:
UE issued RRC Connection Request message, but RNC did not receive.
Problem analysis and solution：①If the downlink P-CCPCH of the C / I is lower, it is the issue of coverage, through the RF optimization solution. If the covering is normal, the problem may be RACH, and then you can increase the P-RACH uplink expecting to receive power and increase the terminal transmission power to improve the success rate.②If the terminal transmit power is limited, the performance belongs to the UE own problems which are impossible to solve.③UpPTS (Uplink Pilot Time slot) result in interference, a slight increase uplink interference can expect to receive P-RACH power to resolve a serious UpShifting technology, which can be used by the corresponding parameters of the background configuration, adjust the location of UpPCH avoid interference, or simply to avoid sources of interference;④Node-B equipment may be required to check whether there are warning RRU channel.

(2) RNC received RRC issued UE requests the establishment of news, made under the RRC Connection Reject message.
Analysis and solutions: ① Congestion, Road-constrained code; ②slot status check whether the code Road occlusion; ③ equipment failures, parameter configuration error.

(3) RNC received RRC establishment of the news requests issued by UE , then issued a RRC Connection Setup message but the UE has not received.
Analysis and solutions: ① if the P-CCPCH of the C / I is lower, it is the issue of coverage, and it can be solved through the RF optimization; ② if it due to the small district re- select , then adjust the parameters of re-election to speed up the cell selection and re-election rate, which can be solved with cell parameters unreasonable re-election caused by the establishment of RRC connection failure; ③ if there were no re-election, it could be brought about by the low FACH power, then an appropriate should be increase.

(4) UE received the RRC Connection but Setup Message Setup Complete message is not issued.
Analysis and solutions: If the downlink signal quality is normal, then the phone could be abnormal. Otherwise, a dedicated channel may be the problem of synchronization, the initial downlink power of the downlink dedicated channel leading to be too low that can not be synchronized, you need to adjust the initial downlink transmit power.

(5) UE issued RRC Setup Complete message and did not receive RNC.
Analysis and solutions: ① adjust the open-loop Power Control parameter of uplink dedicated channel; ② it has something to do with whether the setting, with the initial uplink SIR, is reasonable or not, and has great impaction the initial production chain at the time of the initial uplink synchronization. If this parameter is set too large, may be the initial production chain allows users brought up when the interference is too large; If it is set too small, it will make uplink synchronization for longer hours and even lead to the failure of the initial synchronization.

3.3 Authentication Problem

When authentication fails, according to the network authentication failure message, which is returned by the UE, the causes of value should be given an analysis. Common reasons include the value of MAC Failure and Synch failure.

(1) MAC Failure
When the Mobile phone terminal in the authentication network, the AUTN parameter in the network side of the request of authentication message should be checked, if one of the "MAC" messages is error, the terminal authentication failure message will be reported and the cause of value is MAC failure.
The main reason causing the problems include: ①illegal users. ②Users'USIM card and HLR in different settings.

(2) Sync failure
Another kind of authentication failure is detected mobile terminal message of SQN serial number errors from AUTN, caused by the failure of authentication, the reasons for a value of: Synch failure (synchronization failure).
The main reason causing the problem are illegal users and equipment issue.

3.4 Safe Mode Issues
Safe Mode Safe Mode usually refuse to vary, the main reasons include:

(1) Cell phone capability, the phone does not support the configuration of the encryption algorithm.
Ability to check the phone as follows: in the RRC Connect Setup CMP news reported the ability of cell phones. Ability of mobile phones for the encryption mode rejection can only be resolved by replacing the phone.

(2) RNC and core network configuration does not match the encryption mode.
As to the mismatch of configuration mode for the encryption, you can view whether the MSC (Mobile services Switching Center) or SGSN (Service GPRS Supporting node) match with the encryption mode the RNC choose to determine that the MSC or the SGSN supports the encryption mode and the RNC to support the encryption mode must be the same.

If the encryption's configuration mode does not match, MSC and SGSN can be set to select all of the encryption mode, RNC chooses according to the actual situation of UEA0 or UEA1.

3.5 The Establishment of the Question RAB
When the RAB or RB fails to establish, RNC in the RAB Assignment Response Signaling in the back to establish a RAB assignment fails. Through the relevant cell of the causes of the failure we can bring the value of the specific reasons for failure.
Common failure of RAB / RB to establishment includes:

(1) Parameter configuration error RNC directly refused the request of RAB establishment.
RNC parameters result in a direct response to the illegal RAB failed to establish business networks in a smaller probability of occurrence, which is normally provided by a special user created by the special operation. The main scenario is: PS business users open an account and activate the application information over the phone's capabilities, resulting in a direct response to RNC refused. For example: a particular user's account is the ability to line up and down to 384K, but the use of the maximum uplink capacity of the mobile phone is only 64K, the use of AT commands in the user terminal or mobile phone to activate the software settings of the QoS information in the PDP line from top to bottom are at the greatest rate of 384K, so when RAB assignment request received by RNC, if it was found that the maximum rate of requests over the uplink capacity of the UE, RAB will have a direct return to the establishment of the failure and the process is not initiated by RB.

When the Parameter setting error over capacity UE caused RAB failed to establish, SGSN will re-consultate of the RAB to launch a new assignment, until the UE can support the ability to complete the RAB assignment, for the user, the activation of PDP (Packet Data Protocol) can still be successful, the instructions to obtain the maximum capacity rate of UE can support the maximum speed. However, if the UE in the PDP activation request QoS settings of the minimum requirements to ensure that the rate exceeded the capacity of UE, then the consensus although the network to accept a lower rate of UE's PDP activation request, but when the UE receive information that found in the PDP activation network consultations the rate is less than the minimum guarantee rate, the UE will take a new initiative to activate the PDP request, and ultimately unable to complete the PDP activation.

(2) UE in response to the failure of RB resulted in the establishment of the RAB establishment failure.
RB failed to respond to the establishment of UE and it was mainly due to the user's wrong action. First, users who have the data downlink 128K was received the VP of the RB business establishment request (VP caller or called party), as most terminals do not support the downlink PS business VP at the same time and high-speed, the failure of UE's direct establishment of RB because of unsupported configuration.

Another situation is that the caller 3G terminals called VP business presence in the GSM network side, do not support the VP business. Received in the RNC this RAB assignment request, the Core Network Call Proceeding issued immediately after the Disconnect command, the reason is that the Bearer capability are not authorized. At that time UE just received the RB_SETUP order, no enough time to complete the establishment of RB, after receipt of the Disconnect immediately initiated the establishment of the failure of RB in response, RNC return and RAB failed to establish, the reason is failure in radio interface procedure.

(3) Air interface to establish the failure of RB resulted in the failure of the establishment of the RAB.
Another failure of RB to establish is that RB did not respond to the establishment of order, therefore, RNC considered the establishment of RB fails. RB to establish a command performance for the ACK is not received or did not receive the order to complete the establishment of RB. Such a weak signal was seen in the area of weak signals, there are two reasons why the signals are weak, one is the absence of UE in the best area to launch access, and the other is the bad covering.

If UE does not reside in the best area to launch access, UE in the process of the establishment of RB will ensure that our programs are set to update by adding the best area (at the same time signal the presence of rapid changes in cell signaling leading to the rapid decline), but as a result of processes' caring out can not be nested (network and terminals do not support), activity update only after the completion of waiting for the establishment of RB, RB-building process resulted in weak signal area, and be prone to fail. In view of this problem, we need to increase the frequency of re-election with the start-up threshold and speed, making the presence of UE in the best area as soon as possible, in the best area to launch access.

The poor coverage which resulted in the failure of the establishment of RB can be divided into two situations: the upstream and downstream quality dissatisfaction. Downlink coverage leads to the failure of UE not receiving the order of establishment from RB. Downstream quality dissatisfaction is partly because poor UE demodulation performance, and partly because the need for RF optimization to solve the problem. The problem arising from Uplink coverage is that: UE received the order of establishment from RB, but RAN can not get the ACK established by RB or the order of the completion of RB-building, which may be caused by uplink interference, it can be confirmed by examining the RTWP.

4. THE SOLUTIONS OF TD-SCDMA NETWORK ACCESS

Through analyzing the physical layer related with access process and the wireless parameters, we can sum up the solution idea and the process about the issue of optimizing TD-SCDMA network access, As shown in Figure 1:

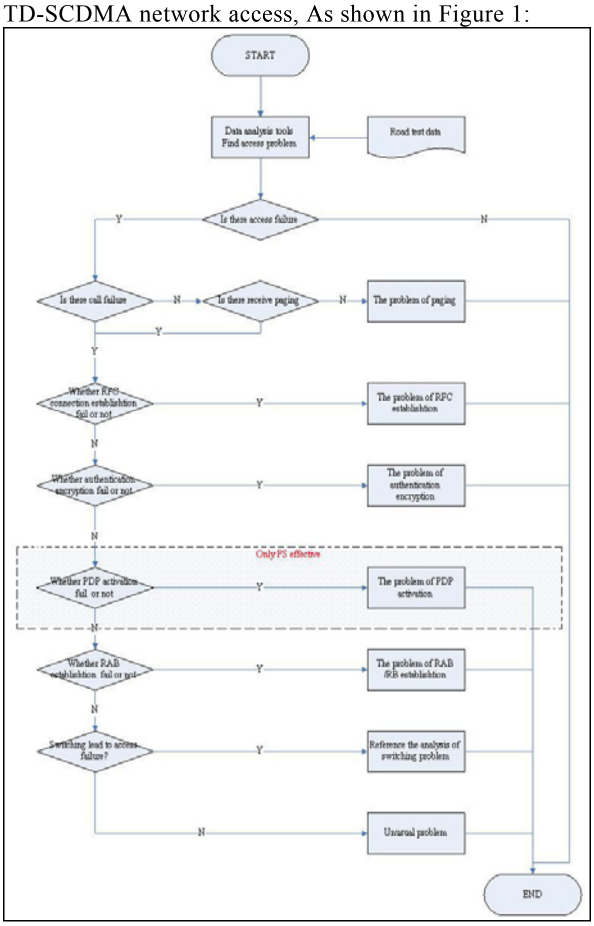


Figure 1

5. CONCLUSIONS

Under the full support of the Government and Operators, TD-SCDMA Industry Alliance and the industry chain has been basically established, product development has been further driven, more and more equipment manufacturers have invested in the camp of the TD-SCDMA product development. With the equipment developing, field test large-scale carrying out, the TD-SCDMA standard will also be further validation and strengthened. Whether the terminal successfully access network and get services or not, is a key indicator about measuring maturity of TD-SCDMA network. The paper summarizes the access problems that were encountered in the process of optimizing TD-SCDMA 3G network, and gives out the ideas of solution from the physical layer of TD-SCDMA network. However, the TD-SCDMA industry chain where there are immature will miss some unknown problem when the network running, so further optimizing the TD-SCDMA system is still the important research work which I am going to focus on.

REFERENCES

[1] Jian-Feng Liu, Bi-Xin Huang. Analysis of CDMA Network Planning and Optimization [J]. Mobile Communications, 2007, (10)

[2] Bo-Yu Liang,Xiang-Xu Chen. CDMA Network Planning and Optimization [J]. Computer Knowledge And Technology, 2008, (29)

[3] Jing-Shuai Fu. Analysis and Investigation about Common Interference of CDMA Network [J]. Shandong Communication Technology, 2008, (02)

[4] Tong-Xu Zhang. TD-SCDMA Network and Its Future Development Prospects [J]. Telecommunications Engineering Technology And Standardization, 2008, (10)

[5] Tong Zhan. The Research of Key Technologies about Optimization of Mobile Communication Network System Under the Background of 3G [J]. Silicon Valley, 2008, (06)

[6] Yi-Ming Chen. TD-SCDMA Technology Standard and Its Development Prospects [J]. Information Technology And Informationization, 2007, (04)

[7] Hong-Wei Wang. Network Optimization of the Mobile Communication Network [J]. Information Technology, 2006, (04)

[8] Tong-Xu Zhang. On the TD-SCDMA Network Construction and Future Prospects [J]. Communication World, 2008, (27)

STUDY OF PORTABLE ECG MONITOR BASE ON S3C244B0X *

Tan Yue, Shen Limin, Wang Honghu, Tan Yunfu
School of Information Science and Engineering, YanShan University
Qin Huang Dao, Hei Bei, China
Email: tanyunfu@163.com

ABSTRACT

In this paper, combined with a portable ECG monitor research work, introduced the design process of the ECG monitor system's hardware and software systems. First introduced the structural design of hardware systems, S3C44B0X embedded microprocessor interface and ECG acquisition circuit, and then introduced the transplant of the μC/OS-II real-time operating system, and finally the design process of the ECG features and the software. The ECG monitor which could detect and analysis the patients' ECG signals in real time could record or reproduced the changes in ECG for some time past. Therefore, it may assist in the treatment of heart disease effectively.

Keywords: ECG, S3C44B0X, μC/OS-II, Portable

1. INTRODUCTION

Cardiovascular disease poses a serious threat to human health and life and the ECG which has become the main basis of the diagnosis of cardiovascular disease could accurately reflect the situation of heart disease. At present, the widespread used of heart disease monitoring system which was a retrospective analysis system can hardly diagnosis and alarm the temporary and paroxysmal arrhythmia timely. In this paper, studied the issue and proposed an ECG Monitoring Program based on embedded micro-controller.

2. THE SYSTEM OPERATION PRINCIPLE

The ECG signals which was acquired form the patient's chest by Special electrodes was send to the signal shaping circuit, and then into the band-pass filter to filtered the interference signals outside the ECG signal frequency range after the amplification of the preamplifier, and then after the main amplifier's further enlarge of the scope to fit through the 50Hz and 35Hz notch filter to filter out EMG-frequency interference, and be in line with the requirements of ECG analog signal into S3C44BOX after the on-chip ADC for A / D conversion. A detection circuit was added in order to prevent the leads' loose and fall off. The system uses Samsung's 32-bit high-speed microprocessor control S3C44BOX devices as a core, S3C44BOX contains LCD controller, an external 3.5-inch LCD to display the ECG waveform and ECG diagnosis in real-time. It uses the Ethernet interface to transmit the ECG data to the hospitals through Internet for further diagnosis and the system connects the S3C44B0X'S GPIO interface with micro-printer which could print the electrocardiogram and data. The embedded real-time operation system μC/OS-II which could ensure the system operate in real-time, high speed and steady was transplanted to the system. The structure of the Portable ECG Monitor's hardware system is show as figure 1.

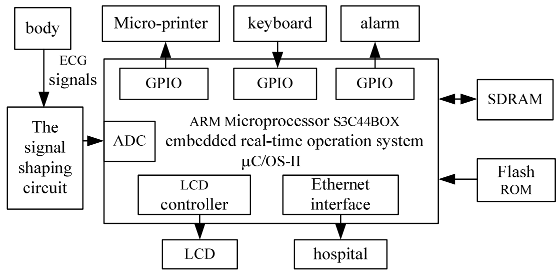


Figure 1. The Structure of The Portable ECG Monitor's Hardware System

3. THE DESIGN OF HARDWARE SYSTEM

3.1 The Signal Shaping Circuit

The ECG of the body normally in the 0.05-5mV, its frequency range from 0.05Hz to 100Hz, and with the characteristics of weak and low-frequency and vulnerable interference. The signal shaping circuit includes pre-amplifier, band pass filter, main amplifier and notch filter modules, as shown in figure 2.

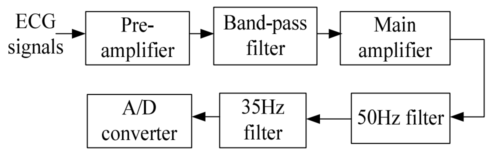


Figure 2. The Signal Shaping Circuit

The pre-amplifier[1] was the core of the amplifier module. The system choose the ANALOGDEVICES'S AD623 which has the characters of high input resistance, low input bias current, low noise and high common-mode rejection ratio as the core of the pre-amplifier. The AD623 could set its gain from 1 to 1000 with only one extern resistance through setting the impedance Rg between the pin 1 and pin 8. In order to enhance the capacity of DC polarization resistance and to avoid entering the cut-off or saturation, the pre-amplifier's gain should not be too big and the gain was designed as: G=100k/Rg+1=10， Rg=11kΩ. The main amplifier which contained the OP1177(U7)、R12、R13 modules enlarged the ECG signals from the pre-amplifier to 100 times to achieve the requirements of the A/D converter. The band-pass filter was constituted by OP2177 from which the High-pass filter was constituted by U6A、C6、C7、R8、R9. In order not to lose the low-frequency components of the ECG, the Cut-off frequency was designed as: $f = 1/\left(2\pi\sqrt{C6 \times C7 \times R8 \times R9}\right) \approx 0.03Hz$, the low-pass filter was constituted by U6B、R10、R11、C8、C9 modules which Cut-off frequency was : $f = 1/\left(2\pi\sqrt{C8 \times C9 \times R10 \times R11}\right) \approx 110Hz$, and from which the Low-frequency and high-frequency noise could be suppressed effectively. The band-pass filter and the main amplifier was shown as figure 3.

*The project is supported by the National Natural Science Foundation of China(serial number is 60873008)

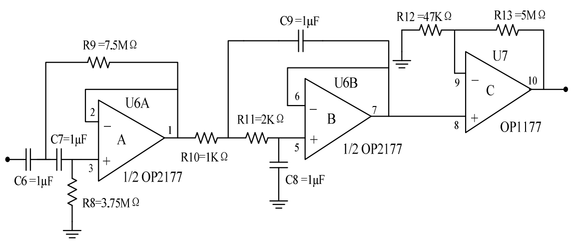


Figure 3. The Band-pass Filter and Main Amplifier Circuit

The 50Hz frequency noise is a major interference of the ECG signals and the EMG interference which frequency was mainly about 35Hz also have varying degrees of impact to the ECG. For these two interferences were all in the coverage of ECG signals, the 50Hz and 35Hz filter were designed. The detection circuit of preventing the leads' loose and fall off which could output an pulse from high-level to low-level would produce an alarm signal to remind the patient to replace the leads.

3.2 S3C44BOX and Peripheral Interface

S3C44BOX is 16/32-bit RISC processor, and it was developed by Samsung company. It designed specially microcontroller solution of high-performance and cost-effective for portable devices. CPU core use 16/32-bit ARM7TDMI RISC processor by ARM designed, the working dominant frequency is 66MHz, the working voltage is only 2.5V. It has 8KB cache, and external memory controller has eight banks, a total of 256MB, 8-channel and 10-bit ADC, and LCD controller that support color/gray/ black-and-white display. It was provided with the power control mode of ordinary, slow, idle and stop. The all-static design was low power consumption, rightsizing and excellent, which was especially suitable for low-cost and power-sensitive applications [2].

3.2.1 A/D Conversion

Before data analysis and processing and storage, first of all, the collection analog ECG signals were converted into digital signal by the A /D conversion. The consideration of for the ECG came from mainly the requirements of abnormal analysis and processing for the ST segment. ST segment level variation was 0.05mV, which was recognized. Therefore the sampling precision was 0.025mV at least.

When using 10-bit A / D converter that it was working positive polarity and full-scale voltage +2.5V, the minimum of input voltage changes that it could distinguish from was 2.5V/210≈2.5mV. The amplification of the ECG acquisition circuit was about 1000 times. So the minimum resolution of the input terminal was 2.5mV/1000=0.0025mV. Therefore, the precision of the 10-bit A / D converter had completely meet the needs of ST segment analysis and processing of the system.

3.2.2 Ethernet interface

S3C44B0X microprocessor has not integrated network interface controller, so expand a PTL8019AS controller of REALTECK Company for the system, it has good software migration and simple interface circuit and other characteristics. When CPU send some data, the data want to send will be written to RTL8019AS by RTL8019AS drivers according to specified data format and start to sending orders; On the contrary, RTL8019AS receive signal from the physical channel, then they would be converted into corresponding data format directly from Ethernet protocol, and send out interrupt, which request CPU to read data [3].

3.2.3 Extended Memory Flash ROM and SDRAM

S3C44B0X microprocessor did not have its own ROM, so using a SST39VF160 as Flash ROM of the system external and used to store system boot code and automatic diagnostic programs of ECG. SST39VF160 by SST introduces and a capacity of 1MBX16 is the CMOS Multi-Purpose Flash ROM. It has good reliability, low power consumption, fast read, etc. After the system power on, first of all, it begins to perform from the address of 0X00000000 where the boot code stored in this address. Therefore, the SST39VF160 mapped Bank0 region of S3C44BOX microprocessor.

HY57V641620 chip of Hynix Company is an 8MB SDRAM, it has the characteristics of large storage density and high-bandwidth, and it is very suitable for running space, data area and stack area of the system program. Moreover, directly connect extended storage manager of S3C44BOX with the S3C44BOX, no software drivers, the data transmission can be realized after setting BANKCON register. The HY57V641620 mapped Bank6 0XC000000H-0XCFFFFFFH region of S3C44BOX microprocessor.

3.2.4 LCD and Keystroke.

LCD could display the ECG waveforms, ECG parameters and the system menu in real-time. At the same time, keyboard was used to control and display of instrument, and to provide a friendly human-computer interface for doctors. The LCD was liquid crystal display HS240128-1 and it was made by Shenzhen Han Sheng Industrial Co., Ltd. HS240128-1 was using STN (a kind of display materials) blue-grey display mode, which integrated within the LCD controller chip T6963CFG and the row-column driving chip T6A39, T6A40 of Toshiba. It can carry out the characters, graphics and graphic mixed display easily.

The dynamic ECG monitor achieved the display mode of cascading menu through the parameter setting, ECG detection, alarm processing, ECG printing, network transmission and ECG intervals ,the six keys, and to provide the choice of function for doctors; the six keys adopt interrupt input mode, scanning and positioning keys in the software when the key interrupt occur, and write corresponding the functions and menu interface under different keys' states; just only writing the corresponding character code on the display buffer of the corresponding location on the LCD screen when each function menu was displayed .

4. SOFTWARE DESIGN ON(S)

System software include mainly three parts, they were embedded real-time operating system μC/OS-II, ECG filtering and detection program, automatic arrhythmia diagnosis program. The follow would mainly introduce the first part and the third part.

4.1 Embedded Real-time Operating System μC/OS-II

μC/OS-II was a complete, portable, curing, clipping multi-tasking preemptive real-time embedded operating system; support the 56 tasks of user, support semaphores, message mailboxes, message queues and other inter-process communications used commonly [4]; when μC/OS-II was transplanted, need to rewrite code associated with the processor, these code constitute three files, the files is respectively OS_CPU.H,OS_CPU_A.S and OS_CPU_C.C.

4.1.1 System Initiation Process

The system initiation code was:

```
int Main(int argc,char **argv)
{
TargetInit(); // Initialize S3C44BOX
GUI_Initialize(); // Initialize LCD
OSInit(); //μC/OS-II Initialization
OSTaskCreate(Task_Main,&task_main,&TaskStK[TASK_STK_SIZE-1],5); // Founds the main task
OSTaskCreate(Task_Keyboard,&task_keyboard,&TaskStK[TASK_STK_SIZE-1],3); // Founds the keyboard task
OSTaskCreate(Task_Reset,&task_reset,&TaskStK[TASK_STK_SIZE-1],8); // Founds the reset task
keyscanMbox=OSMboxCreate(0); // Founds keyscanMbox Message Box
keyMbox=OSMboxCreate(0); // Founds keyMbox Message Box
driverresetMbox=OSMboxCreate(0); // Founds driverresetMbox Message Box
beginResetMbox=OSMboxCreate(0); // Founds beginResetMbox Message Box
OSStart();    // Start μC/OS-II;Start task scheduling
return 0;
}.
```

After system initiation, we first initialize S3C44BOX and LCD, and initialize μC/OS-II. Then founds the main task, the keyboard task and the reset task, their priority respectively is 5, 3, 8. Finally founds the Message Box, between the duty cut realizes through the News Mechanism.

4.1.2 System Main Task

It founds main task which named Task_Main in the initialization code, when after the keyboard task gives main task sending a message, main task start; In the main task, first permits the keyboard interrupt, then display system menu, finally according to pressed key value selective calling module. Main task flow chart was shown in Figure 4.

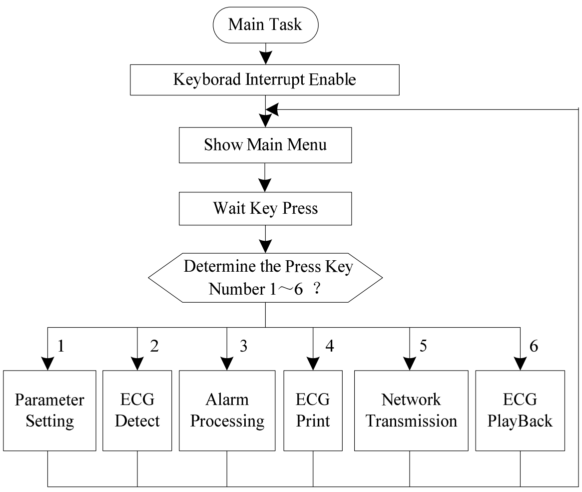


Figure 4. Main Task Flow Chart

4.2 Arrhythmia Automatic Diagnosis

Heart electrical signal after filter and examination procedure obtains the RR time and the QRS wave group width two parameters, the automatic diagnosis has provided the correct basis for the arrhythmia.

4.2.1 Heart Rates

Heart rate was the number of the heart beat per minute, it including instantaneous heart rate and average heart rate. The heart rate can be calculated using the R wave: The instantaneous heart rate was referred to the reciprocal of two neighboring the R wave time-gap T of the heart electrical signal, namely: $f=1/T=60/T$ (times /minute). The average heart rate was the ratio of, in certain counting interval, the number of R wave integer n and counting interval T, namely: $F=n/T$ (times/minute).

4.2.2 Arrhythmia Diagnosis

The arrhythmia auto-analyzer procedure was recognition and classification through the rhythm analysis and the profile morphological analysis to the heart electricity profile which obtains, then the basis determined in advance the diagnosis standard comes to the arrhythmia to make showing off to earn the diagnosis. According to the clinical need, we have studied 13 kind of important arrhythmia. To enable the procedure to automate analysis and diagnose each kind of arrhythmia, we needs to use the mathematical method to define them. The definitions are following [5]:

Bradycardia: $RRt>1.5s$(Heart Rates <40 times/minute)or $ARt>1.2s$;

Tachycardia: $ARt<0.5s$(Heart Rates >120 times/minute);

Asystole: $RRt \geq 3s$;

Leaky-wave: $RRt > 1.5RR^*$;

atrial premature beat: $RRt<0.75RR^*$ and $WWt \leq 120ms$ or $RRt+ RRt+1 < 2RR^*$;

ventricular premature beat: $RRt< 0.75RR^*$ and $WWt > 120ms$ or $RRt+ RRt+1 \geq 2RR^*$;

Bigeminy coupled trigeminy: $RRt\text{-}3<0.9ARt\text{-}4$ and $RRt\text{-}1<0.9ARt\text{-}4$ and $RRt\text{-}3+RRt\text{-}2 \approx 2ARt\text{-}4$ and $RRt\text{-}1+RRt \approx 2ARt\text{-}4$;

Trigeminy trigeminy: $RRt\text{-}2<0.9ARt\text{-}3$ and $RRt\text{-}1<0.9ARt\text{-}3$ and $RRt\text{-}2+RRt\text{-}1+RRt \approx 2ARt\text{-}3$;

coupled ventricular premature beat: Presents ventricular premature beat two time continuously;

R-on-T: $0.2s < RRt < 0.33RR^*$;

Insertion of ventricular premature beats: $RRt < 0.67RR^*$ and $WWt > 120ms$ and $0.9RR^* \leq RRt+RRt+1 \leq 1.1RR^*$;

paroxysmal ventricular tachycardia: Presents ventricular premature beats three or more times continuously

Myocardial ischemia: The level of type, under the ramp type (slope ≤ 0) ST segment depression $\geq 0.05mV$ and duration $\geq 0.08s$;

In the above definitions, RR^* was the average RR time of four continual normal QRS wave group. It is founded by procedure when initialization produces. RR was the RR time; WW was the QRS wave group width; AR was the average value of the 5 RR time; and the subscript was the relative time.

5. CONCLUSIONS

Through transplants the μC/OS-II operating system on S3C44BOX, we has realized the system resources effective management, realized the functions of the heart electrical signal real time display, the arrhythmia automatic diagnosis, exception alarm, the heart electricity data net transmission and heart electrical signal playback and so on; This heart electricity guardianship instrument has the merit of the low power dissipation, the small volume, the light weight and the higher intellectualized degree. It has the high clinical practice value and the broad market prospect.

REFERENCES

[1] Zou Bo, Chai Jihong. Based on ARM processor's ECG Blood Pressure Monitor Development[J]. Medical Equipment,2005,1(18):7-10.

[2] Wang Tianmiao. Embedded Systems Design and

Examples of Development[M]. Beijing: Tsinghua University Press, 2002.
[3] Xu Jing. Embedded system design and examples development explains in detail[M]. Beijing: Beijing University of Aeronautics and Astronautics Press,2005.
[4] Shao Beibei. Embedded real-time operating system μC/OS-II [M]. Beijing: Beijing University of Aeronautics and Astronautics Press,2003.
[5] Qi Jianxin, Bian Zhengzhou, Yang Qiang. Based on microcomputer's heart electrical signal real-time automated analysis system [J]. Beijing Biomedical Engineering,1997,16(3):157-160.

Tan Yue, student of the Computer Science and Technology, School of Information Science and Engineering, Yanshan University, the leader of National Undergraduate Innovative Test Program (serial number: 081024612), she’s research is Embedded Systems.

Shen limin is a Full Professor of School of Information Science and Engineering, Yanshan University. He graduated from Yanshan University in 1984. He is a Visiting Professor of the IIT(Department of Computer Science, Illinois Institute of Technology,USA 2005), His research interests are in flexible software engineering, commence and application software, performance analysis of e-commence systems.

MOBILE OA SYSTEM RESEARCH AND DESIGN BASED ON 3G

Wu Wanping
Department of Computer Science , Wuhan University of Science and Engineering
Wuhan, Hubei Province,430073, China
E-mail: wwp@wuse.edu.cn

ABSTRACT

In recent years, as the level of information technology rises, enterprises' requirement for automation goes up. Enterprises want to build office system which is not limited by time and space, in order to increase efficiency and core competence. The paper introduces the building of a mobile ERP structure on the basis of 3G and how to realize it.

The paper first discusses the structure of enterprise mobile office system on the physical model and its parts. Then, on the basis of analytical system logic model, it designs the system software plane, proposes a solution to an enterprise mobile office system plane on the basis of WAP 2.0, and studies data synchronization between mobile server and enterprise server in wireless network. Afterwards, according to the detailed requirement of a decoration company's mobile office system, the paper shows how to build the plane and integrate data of the old system.

Keywords: 3G, WAP2.0, Mobile OA System, J2EE, Web Services.

1. INTRODUCTION

In recent years, as the level of information technology rises, enterprises' requirement for automation goes up. However, this kind of office system can only be used in offices or other places connected to the Internet. It's limited by time and space. If relevant employees go to field work, some emergencies can't be handles in time, so that the whole project would be delayed. Nowadays, 3G-based mobile communication system has been put to use. If we use 3G to build a mobile office platform, it can well fill up the deficiency, create a good applying environment, expand business to user's mobile phones by Internet and wireless communication, and truly realize 3A mobile business mode.

The general solution of mobile business is combining automation and communication technology. It not only provides all kinds of business needed function, but also guarantees that users can access and process information anywhere and anytime, which totally solves the hard knot of unable to do business out of the office. It greatly increases efficiency, and releases managers from traditional offices.

The paper takes a famous decoration company as example to illustrate the structure of mobile business system and its key technology.

2. ANALYSIS OF MOBILE OA SYSTEM

The building of the OA system is based on Java programming language, provides services in WAP 2.0 mode, and is based on 3G wireless technology.

2.1 JAVA technology – the main technology support of mobile business system

JSP(Java Server Pages) JSP is a technical standard for dynamic web page advocated by Sun Mcrosystems and built by many other companies. The aim of designing JSP is to make executive programs based Web faster. Those executive programs work with all kinds of Web servers, application servers, browser and develop tools.

2.2 3G technology – the support for raw data which mobile devices access

3G(The third generation mobile telephone technology) is short for the third generation mobile telephone technology. Compared to the first generation, simulating mode, and the second generation, GSM, TDMA , the digital mode, 3G combines multimedia communication such as wireless communication and the Internet. 3G technology enables a large quantity of information transmission, such as audio, video and multimedia files. It makes everything swifter for the user, either watching videos online or accessing the Internet. The system mainly takes advantage of the feature of 3G, large information transmission and support multimedia data, realizes the utilize of computer and mobile devices in OA system through WAP.

2.3 WAP2.0

WAP is short for Wireless Application Protocol. It's an open global standard. WAP 2.0 is the newest version of WAP. It's based on the old version but took on some newest protocol and standards. It strengthens the support for Internet communication protocol and security, better adapting to the changes in wireless situation and the expected market needs.

XHTML is the most basic language in WAP 2.0. XHTML is advanced by W3C. It can be used to develop normal Internet …, or mobile phones. It better adapts to the application of 2.5G/3G broadband wireless network. The developers can write programs that apply to both PC and mobile phone in order to reduce the expenses.

WAP 2.0 also consist of CSS. On applying CSS, developers can separate style property of XML file from actual contents, thus reduce the quantity of marker language in the memory of browsers. XHTML and CSS provide developers stronger format function. With CSS, one set of XHTML page can apply to different need according to various size of the screen.

3. THE MAIN STRUCTURE OF MOBILE OA SYSTEM

As an expanded application in wireless, mobile OA perfectly combines modern information system. It enables managers to know timely information and check important projects when absent to the office. Anywhere, absent employees can keep a high efficiency as in an office and get access to company's inside data. Important notices can reach relevant employees by text messages. It solves the difficulty while they are outside. Mobile OA has the following advantages:

(1) It broadens work space. Mobile OA breaks the physical border of work. It enables clerks to break out of the limit of the offices and realizes non-difference of working environment.

(2) Work is done anywhere and any time. It enables leaders and clerks to work anywhere and any time, including project examination, checking report forms, checking financial states, the staff statistics, consultation of address list and automatic text message notices. It brings great convenience to relevant departments in managing work in time and avoids blocks in procedures.

(3)It increases efficiency by doing work in time.

3.1 The structure of mobile OA

Figure 1 illustrates the structure of mobile OA platform. Mobile OA servers link to the wireless network via service provider. Mobile terminals visit mobile OA servers via wireless service providers. Mobile OA servers link with original information servers and exchange data.

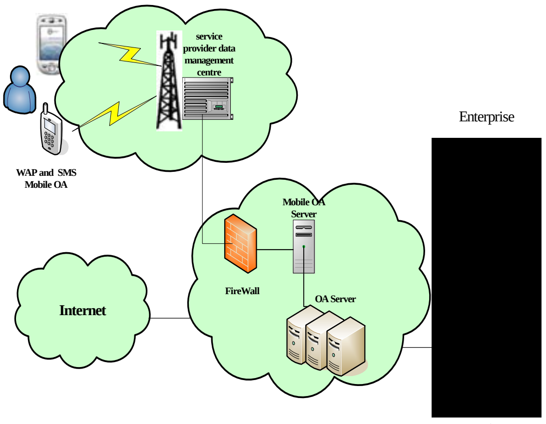


Figure 1. The Structure of Mobile OA.

The idea of mobile OA platform is to build a reasonable, open and standard-based office platform. It gradually covers the integration of applied system in all business fields. The management system can be chosen according to various need and users can build their own application system within the network in order to form a whole wireless platform.

Wireless application platform is based on wireless network. On the basis of interface with MAS, it perfectly combines with present allocation system and integrates Internet technology, wireless communication technology, MAS, ADC ect and enables field workers to exchange data through intellective mobile terminals. The combination of wireless application and traditional Web mode not only satisfies users' using preference, but also applies a new working mode and greatly increases efficiency, especially on the platform of 3G wireless network.

3.2 The Logic Frame of Mobile OA

After the expand of mobile OA, the office system becomes a whole, thus realizes working anywhere and any time. Figure 2 illustrates the system frame.
The mobile OA consist of three components as follows:

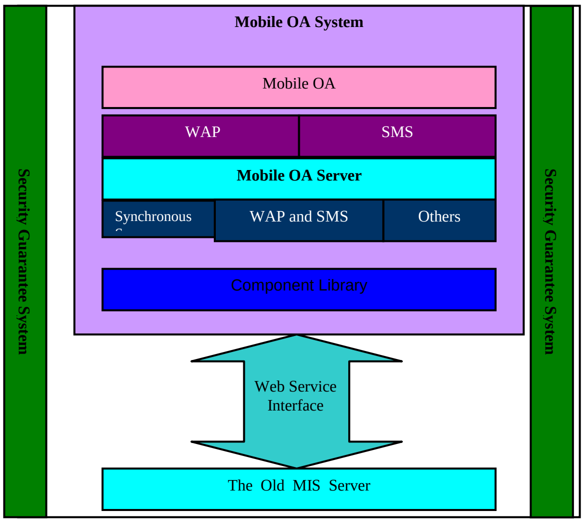


Figure 2. The Logic Frame of Mobile OA.

(1) Application of mobile terminals
Aimed at the diversity of working needs, we combine WAP and text messages. Mobile terminals are not fit for much sophisticated operation. They are mainly for browsing and a little feedback, so visits via WAP doesn't rely on mode or brand of the mobile phone. It's convenient and current. As with some urgent approval and notices, WAP PUSH technology sends the messages to terminals of relevant employees.

(2) Mobile service provider and special mobile network
Mobile terminals transmit data via network provided by mobile providers. Some data need secrecy, so terminals need to apply for special line in wireless APN and open APN visiting rights to certain SIM cards. Thus mobile terminals visit wireless OA server via APN software.

(3) Wireless OA server
Wireless OA servers function as mobile application and process logic and exchange data with traditional office system.

4. THE REALIZATION OF FUNCTIONS OF MOBILE OA PLATFORM

With OA technology mentioned above, the paper sets forth an actual example, i.e. the realization of mobile OA of a famous decoration company. This company has more than 10 branches in China. An integrated information system within the company satisfies most needs in various affairs and has hugely increased efficiency and has reduced expenses to the lowest. Employees are able to break out of trivia and participate in more creative work. But limited by features of the industry, leaders are often away. Under traditional OA system, many urgent affairs can't be handled in time, therefore delay the whole process of work. So this company needs a office system that can be used anywhere and any time. Mobile OA system is an expansion to the original information system using wireless communication network. Work can be done anywhere mobile phone signals can reach. It's timely and convenient.

Mobile OA system consist of two main components: function of wireless terminal user and background application server.

4.1 Function of Wireless Terminal Users

Login to system:
All users must login to access to the system. Users input user ID and password while logging in, and server affirms user rights guide users into different pages according to different rights. The affirmation of users can be bonded to mobile phone numbers or SIM card numbers by requirement.

Urgent matters:
Urgent matters are an important component of the whole OA system. They're the windows which the system faces users. When the system runs, projects to be approved, such as contract low discount application form, contract changing protocol, contract cancellation application, contract ending report, contract ending application, client balance report and project management report, all belong to leaders' awaiting matters. Leaders only need to check and deal with matters in urgent affairs. Easy operations can finish the checking and approving and brings great convenience.

Consultation of report forms:
Leaders can check data of daily and monthly report of the company and also of branches through mobile terminals.
Consultation of important information:
Via mobile OA, leaders are able to consult important information of the company such as circulating capital, finance state, expenses and faculty statistics.

Meeting notices:
Besides sending messages to notice meeting participants, participant list can be consulted through mobile OA. Clicking on the list shows the contents of meeting.

Consultation of inside directory:
Mobile terminals' storage is limited, therefore it's impossible to store address of everybody. So the mobile OA has a inside directory which enables field workers to consult.

Text messages:
When there're awaiting matters, urgent notices, form created or meeting notices, the system automatically send short messages to relevant leaders and employees.

4.2 The Function of Mobile OA System Application Server

Data sorting:
The mobile OA system doesn't run alone. A lot of data combine various business figures and become a united information system. New systems need interface to exchange data with present information system. They must keep updated in order to synchronize with other systems.

Integrated system:
The mobile OA system integrates with present information. In the course of information, there'll be other data sources and application systems being built and new integrated devices. Mobile OA information systems serve in daily affairs, so then they'll need to integrate with the new information and devices.

Group short message platform:
Via this platform, system can send short messages to inform relevant staff of urgent matters. Replies can be feedback to the system to check if the message is read. So to integrate short message platform into the system, the system runs the platform interface to send text messages.

5. CONCLUSION

Mobile OA platform is based on 3G IMS network, WAP and J2EE. It acts as a base to swiftly realizing 3G mobile business. It definitely will guide mobile business to be wireless, personal, intellective, cooperative, portal, network-based and general. It's hopeful that in the coming years, mobile OA system will develop and utilized in the trend of combing network and business.

REFERENCES

[1] Zhang Si-quan ,and Zhang Zhan-song, "WAP 2.0 And Push Application In Wireless Network",Electronics Quality,Guangzhou, 2008.1, p124~127.
[2] Richard Monson-Haefel. J2EE Web Services-The Ultimate Guide[M]. Tsinghua University Press. 2008,4.
[3] Michael Taylor， Strategies for J2ME MIDP/J2EE Integration over HTTP, http://microjava.com/articles/J2ME_J2EE_integration_tartegies.Pdf.
[4] Li Fudong , and Wu Weiming, "The Architecture of 3G Mobile Office Platform And Technology Related"[J], Office Automation, Beijing, 2008.2.
[5] Jonathan Knudsen, Parsing XML in J2ME, httP://develoPers.sun.com/techtopics/mobility/midP/articles/parsingxml/.
[6] Walsh AaronE. J2EE 1.4 Web Sevrices.Dr.Dobb's Journal，v28，n4, April，2003.
[7] Li Qing, "An Architecture Design and Services Implement for Mobile Best Tone WAP Portal" [J], Telecommunications Science, Beijing , 2009.5.
[8] Zhang Yong, "Rea liza tion Plan forMobile OA System and Its Practice"[J], Hubei Electric Power, Wuhan, 2007.2.
[9] Wang Ying, "OA System Research and Design Based on Workf low"[J], MECHANICAL ENGINEERING & AUTOMATION, Taiyuan, 009.2.
[10] Jin Lingchun and Wang Xin, "Study on mobile work system based on short message service(SMS)"[J], Journal of Harbin University of Commerce, Haerbin,2004.4.

Wanping Wu, an associate professor of Department of Computer Science, Wuhan University of Science and Engineering. Her research interests mainly include: Database application development, Computing Grid, network technology.

AN EFFICIENT MEMORYMANAGEMENT SCHEME IN SOFTWARE TRANSACTIONAL MEMORY SYSTEM *

Huang GuoRui, Zhang Ping, Liu XiaoXian, Xiao JinGe, Wei GuangBo
Department of Computer Science and Technology, Zhengzhou Information Technology University
Zhengzhou, Henan 450002, China
Email: howie-mgw@263.net

ABSTRACT

Software transactional memory is the most promising method of parallel programming on multi-core processor, but in some non-blocking software transactional memory systems, non-blocking synchronization supported by memory management is weak and the performance of it still needs improvement. This paper discusses an improved reclamation scheme based on epoch, and implements this scheme in software transactional memory system. The epoch of each shared operation is set on a transaction. Transactions update epoches when start and finish. The reclamation time is determined by comparing epoches of objects with current epochs of transactions. In addition, singularity detection mechanism of threads is designed to avoid blocking, which satisfies the requirement of non-blocking synchronization. The results of the est show that this new memory manager is high performance.

Keywords: multi-core processor, software transactional memory, non-blocking synchronization, memory management, garbage collection.

1. INTRODUCTION

Software transaction memory is the realization of transactional memory model by software. The main idea of transactional memory is using transaction mechanism to ensure the consistency of data sharing among threads. Its final goal is that system provides a simple programming abstraction to the programmer for convenient programming. The programmer does not need to consider excessively the synchronization among threads, and does not need to worry about deadlocks, priority inversion, or preemption, brought by locks [0].

The main concerns in the realization of transaction memory are how to track transaction's write/read operation on shared data, and how to implement transaction's commit or rollback as well as how to control threads concurrently. Transaction memory can be divided into three categaries by implementation: software transactional memory (STM), hardware transactional memory (HTM) and hybrid transactional memory (HyTM). HTM uses the hardware mechanism to implement transactional memory model, so it is constrained by the hardware structure, for instance, the buffer. HyTM, like HTM, also needs the support of hardware, but there are more software characteristics in it. STM is implemented by software, and it uses datastructures to identify the transaction and the memory units, which allows all kinds of existing algorithms to execute. STM is not limited by the hardware structure, so it is flexible and is easier to be developed and improved. Nevertheless, the insufficient is that the performance of STM system is lower than the performance of HTM system. Currently, STM gets people's much more attention for all reasons mentioned above.

A significant issue in STMs is concurrency control, which can eliminate the cost of locking and enhance concurrency. There are two alternative approaches：blocking synchronization and non-blocking synchronization. Some STM systems are built with blocking synchronization to support several transactions to access an object concurrently, and use timeouts to detect and abort deadlocked or blocked transactions [0]. In most STM systems, however, the non-blocking synchronization is used to control concurrent transactions accessing the shared objects. This mechanism offers a stronger guarantee of forward progress than blocking mechanism.

However, one important attribute affecting the non-blocking property and performance promotion of the system is memory management in non-blocking STMs. First of all, most non-blocking STM systems use read-copy-update technique to guarantee that parallel transactions which operate on the shared data could go ahead[0], then, if there are large numbers of operations on memory in the system, the efficiency and performance of memory management will directly influence system performance. Moreover, the non-blocking property of memory management, which lies at the bottom of the system, also directly influences upper transactions' execution. Therefore, many STM systems use the running environment with fully automatic memory management schemes to avoid the problem, e.g. memory management used in Java or C#. However, not all the languages provide this kind of support such as C/C++ language, and the cost of using automatic memory management really exists. In addition, blocking property of memory management in STMs affects threads' forward progress guarantee. Because aborted threads or delayed threads can affect other threads' memory allocation or reclamation, other threads may be blocked. Therefore, memory management needs to support non-blocking synchronization. Consequently, the memory management method is of great importance in STM system.

The rest of the paper is organized as follows. In section 2, we discuss the non-blocking memory management by describing non-blocking synchronization. Section 3 describes and analyzes four memory management approaches. In section 4, we present a memory management approach in STMs. Section 5 reveals the performance of the new approach by test. We summarize our conclusions in Section 6.

2. NON-BLOCKING SYNCHRONIZATION

The key problem of memory management in some STMs is the non-blocking memory management supporting the transaction mechanism, and normally divided into the sub-problems of memory allocation and garbage collection. This paper focus on the notion of garbage collection, including memory reclamation schemes in STM system. To discuss the non-blocking memory management, we illuminate the non-blocking synchronization firstly. The non-blocking mechanism offers a guarantee of forward progress: a stalled

* The financial support was from the National High Technology Research and Development Program of China (No. 2006AA01Z408).

thread will not cause all other threads to stall indefinitely. It includes three common cases.
(i) Wait-free is the strongest guarantee that all threads from the set of contending for the shared resource make forward progress in a finite number of their individual time steps.
(ii) Lock-free assures that at least one thread from the set of threads contending for the same shared resource makes forward progress in a finite number of time steps of any of the concurrent threads.
(iii) Obstruction-free is the weakest guarantee that a thread will make forward progress in a finite number of its own time steps in the absence of contention over the shared resource.

By contraries, blocking synchronization does not guarantee the forward progress of a system, and bring in problems, such as the deadlock, convoying and priority inversion.

3. MEMORY MANAGEMENT SCHEMES

This section provides an overview of each scheme we consider. There are four memory management schemes: lock-free reference counting, lock-free copying garbage collection, hazard pointer based reclamation and epoch based reclamation. Further details about these schemes are discussed in the following text.

3.1. Lock-Free Reference Counting

Lock-free reference counting (LFRC) is a well-known non-blocking garbage-collection technique, by tracking the number of references to objects, reclaiming any object whose count is zero. Counting the number of references is the key of this approach. This garbage collection method protects the safety of the local reference and the global reference [5]. Valois early presented LFRC scheme retaining the object type after reclamation. Michael and Scott corrected it by using CAS and fetch-and-add (FAA). Fraser also used the LFRC methods in FSTM [7]. However, the primary disadvantage of reference counting scheme is the prohibitive performance cost due to the update of the reference counters of referenced objects, even for read-only access. The document [2] and [3] show that LFRC introduces much more overhead by some experiments.

3.2. Lock-Free Copying Garbage Collection

Herlihy and Moss propose a lock-free copying garbage collector. The process can create a new copy within a per-process to region when it updates an object. Periodically each process will create a new to region; the old region is added to a set of from regions to be reclaimed when it is safe to do so. A scan routine in each process will occasionally scan for objects in a from region, copy them to its own region, and update references to that object to point at the new copy. A from region can safely be reclaimed, when it does not contain live objects and each process has subsequently passed a safe point. The major drawback of this scheme is a number of the copying overhead and the memory space cost.

3.3. Hazard Pointer Based Reclamation

Michael presented a hazard pointer based reclamation scheme (HPBR) in the document [3]. It is an existence locking mechanism. Each thread, which performs lockless operations, has K hazard pointers that protect objects from reclamation by other threads, as Figure 1. Herlihy et al. presented a very similar algorithm that applied their repeat offender problem (ROP) mechanism—Pass the Buck (PTB) [9]. PTB uses PostGuard operation to prevent freeing a shared object and use Liberate operation to free a shared object safely, avoiding that other threads access a free object. Since HPBR and PTB only protects the safety of the local reference but not the safety of the global reference, so these schemes are limited in non-blocking algorithm implementation. Moreover, they have high operation costs.

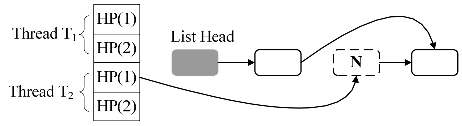


Figure 1 Illustration of HPBR

3.4. Epoch Based Reclamation

We will illustrate the concept of grace period and quiescent state before discussing epoch based reclamation (EBR) [4] [7]. A grace period is a time interval [a, b], which the removed objects before time a may safely be reclaimed after time b. Nevertheless, the safety of reclamation objects is not guaranteed between time a and time b. A quiescent state for thread T is a state in which T holds no references to shared objects. Hence, in multi-thread environment, a grace period is any interval of time during which all threads pass through at least one quiescent state. As Figure 2 shows, any interval of time includes [t_1, t_3] or [t_3, t_5].

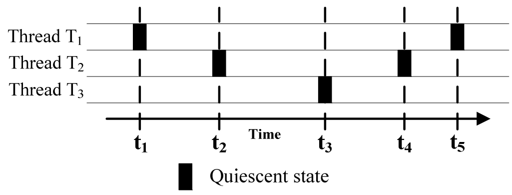


Figure 2 Grace periods and Quiescent States

Kraser's FSTM applies epoch based reclamation scheme, and its epochs are quiescent states [7]. A quiescent state is a time when a thread updates the global epoch. Nevertheless, a thread attempts to do it only if it has not changed for some pre-determined number of critical region entries. Figure 3 shows how EBR allows safe memory reclamation. A grace period of EBR is any interval of time when any concurrent thread occurs at least one time that it atomically updated the global epoch. Upon entering a critical region, a thread updates its local epoch to match the global epoch. Upon exit, the thread atomically clears its flag. The epoch protects each shared operation. In addition, the local epoch of each thread may lag at most one epoch behind the global epoch. The global epoch is one of three logical epoches (value1, 2 or 3). Threads in critical regions can be e or e-1, if the global epoch is e, but not e+1. Removed objects whose epochs are e can be put into a linkedlist whose epoch tag is also e.

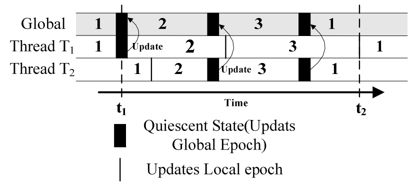


Figure 3 Illustration of EBR

EBR avoids the requirement for the maintenance thread, tries to distribute the reclamation costs among threads, and supports the non-blocking synchronization during the grace period. In addition, the reclamation process by EBR scheme is invisible to the application programmer, making it easy for a programmer to use. However, there is a crucial issue. These threads may wait for a quiescent of other thread in order to reclamate memory, so it may cause threads to block, as Figure 4 shows.

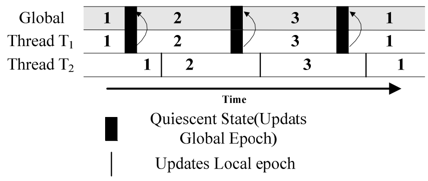


Figure 4 Blocking Threads of EBR

4. MG MEMORY MANAGER

By researching those schemes mentioned above non-blocking schemes spend much more costs than blocking schemes. Hence, we consider an EPR scheme in STM system and design the MG (Malloc & Garbage collection) memory manager. In this section, we describe an overview of this scheme, the data structure, the design of memory allocator and nursery, the design of garbage cleaner and non-blocking implementation technologies.

4.1. Overview

MG memory manager is based on EBR and modifies it. Memory manager contains allocator, nursery and garbage cleaner. Memory allocator carries out the allocation of memory blocks. Memory nursery manages the used memory. Garbage cleaner is responsible for cleaning the garbage memory blocks. **Figure 5** shows the structure of Memory manager.

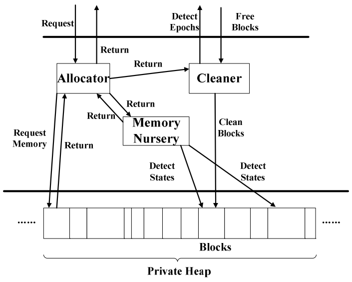


Figure 5 The Structure of Memory Manager

4.2. Data Structure

Block structure contains a next pointer, an object size region and a payload data region. The lowest bit in object size region is a clean bit (or a deleted bit) which indicates that this block can be reclaimed to reuse. The payload data region is practically operated by transactions on threads.

4.3. Design of Allocator and Nursery

The allocator, which contains several size class linked lists, manages all different size memory blocks. Memory allocator transforms the request size into a standard size, when transactions request memory allocation. Then, it allocates a standard memory block to transactions from size class lists. At one time, the memory nursery also obtains the address of the returned block and occasionally detects these used blocks. If the cleaner has cleaned blocks (the clean bit is 1), the nursery can return blocks into the allocation lists.

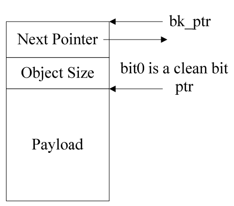


Figure 6 Memory Block

4.4. Garbage Cleaner

Our garbage clean scheme is epoch-based reclamation, but epoches are set on transactions. Each thread has its own epoch. The thread updates its epoch when the transaction is about to start or finish. A running transaction cannot attempt to update the epoch. The cleaner scans delete_committed list or delete_aborted list. It will clean the garbage objects in lists if finds that those epoches of objects are smaller than epochs of concurrent threads. The nursery can reclaim the blocks after the garbage cleaner deletes blocks (the clean bit is set to 1).

(i) **Garbage Objects**

In STMs, the garbage is relatively with the end state of the transaction, since STM updates the copy of the data object. The garbage object is the old version object when the transaction is committed. In another case, the garbage object is the new copy of the data object when the transaction is aborted. Hence, we make two lists manage the different objects: delete_committed list and delete_aborted.

(ii) **Clean of Garbage Objects**

In MG memory manager, the cleaner cannot return the garbage blocks to the running system, and not return the allocator. The cleaner only sets the clean bit. The reclamation of the objects is performed by the nursery.

(iii) **Design of Epoch**

We set the epoch on the transaction since all shared operations are contained in a transaction. At the start of a thread, the value of the epoch is zero. When the transaction is about to start, the epoch is updated to 1. When the transaction will finish, the value is updated (the increment value).

(ⅳ) **Detection of Garbage objects**

The time when the garbage objects are safely cleaned is that threads do not hold the reference to objects. There are three types of references on a transaction: private reference, reference of other transactions and reference of memory manager. Because the memory blocks are not returned to the runtime system, so reference of memory manager is legal. For private reference of the transaction, the transaction cannot hold the reference to objects when the garbage cleaner starts. Then, private references affect the clean and the reclamation of objects. However, references of other transactions cannot be assured when a transaction begins to run the garbage clean. Nevertheless, we can attempt to clean the older epoch garbage objects by comparing the epoch of garbage objects and epochs of all current concurrent threads.

As Figure 7 shows, a transaction of the thread T_1 firstly enters the cleaner. The garbage cleaner records epochs of all current threads in garbage structure (each value is 1). Then, it compares epoches of the structure to each current thread, but the values are equivalent, so do not clean the objects. The transaction of T_1 will end and the epoch is updated (the value is 2). T_1 starts a transaction again, and updates the epoch. When this transaction finishes, the cleaner detects epochs of the garbage structure, finds out that the epoch of T_2 has been changed but the epoch of T_3 is still 1. Therefore, the T_1 cannot begin to clean objects. T_1 can clean objects in the garbage structure until T_1 arrives t_4, because the epoch of T_3 has been updated, (the value is 3). At this time, T_1 can clean objects whose epoch is one or three, but not delete objects whose epoch is five. In Figure 7, [t_1, t_4] is a grace period of T_1 and [t_2, t_5] is a grace period of T_2, but the grace period of T_3 is not detected.

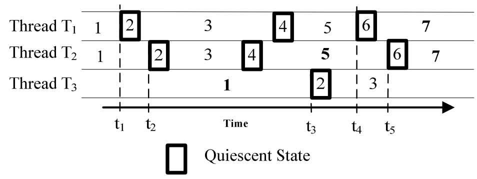


Figure 7 Reclamation of MG Memory Manager

4.5. Blocking Problem

Another problem is that a thread delayed or failed may cause other threads not to clean and reclaim any garbage object, eventually blocking threads. We consider singularity detection and response mechanism. Firstly, we make use of try/catch exception handling to catch some errors, since we implement MG memory manager in C++ environment. We also creates a custom exception type —ABORTED class, as shows. In addition, the cleaner can detect delayed threads by epoches of threads, as Figure 8. When transactions are committed or aborted，threads will record these epoches. If transactions find that the epoch of a thread is updated a long time ago, other transactions can attempt to abort delayed threads until transactions can clean garbage objects. These methods avoid delayed threads to block other threads.

```
if(current_Epoch[id]-Last_Cl_Epoch[id]>=S_thresh)
{
  if(Try_S_ter[id]=1)
  {
          if(current_Epoch[id]-
Last_Cl_Epoch[id]>=H_thresh)
      Cancel_Thread(thread_id);
      Cas(&Try_H_ter[id],0,1);
  } else
  {
      STM_abort(tx);
          cas(&Try_S_ter[id],0,1)
  }
}
```

Figure 8 Pseudocode on detecting delayed threads

5. PERFORMANCE EVALUATION

Our memory management scheme is implemented in RSTM [6]. To reveal the performance and drawback of MG manager, we run some typical test programs. In addition, MG memory manager’s performance is compared with the MallocHeap memory allocator in the experiments.

Our experiments are conducted on a 1.6GHz 2-core Intel E2140 with 1MB of L2 cache and 32KB of L1 I-cache and 32KB of L1 D-cache. Runtime environment is Microsoft Visual C++ 6.0 with POSIX multithread library and RSTM Win32 library, and operation system is Windows XP Professional (SP2). We vary the number of worker threads from 1 to 32. All results are averaged over a set of 3 test runs.

In the LinkedList program, transactions traverse a sorted list to locate an insertion/deletion point, opening list nodes in read-only mode. Once found, the target node is reopened for read-write access. There are more look-up operations before an insertion or deletion to get the target node. Because every active thread performs a 1:1:1 mix of insert, delete, and lookup operations, the count of lookup operation is the most in the program. The RandomGraph program is similar to the linkedList, performing more look-up operations. The MG memory manager works flatly, as **Figure 9** and **Figure 11** shows.

The HashTable program is different from the LinkedList program, since there are not too many lookup operations. Then, memory manager performs roughly equal numbers of read and write operations. Therefore, MG memory manager is slightly faster than MallocHeap manager is, as Figure 10 shows.

MG memory manager reveals the advantage of memory allocation and reclamation in the case of LFUCache and Counter, as Figure 12 and Figure 13 shows. The reason is that these two programs perform significantly more write operations than other programs.

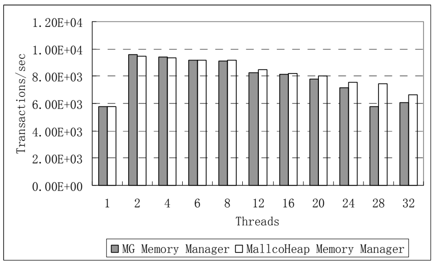


Figure 9 LinkedList Test

Figure 10 HashTable Test

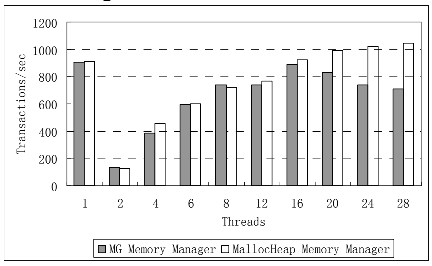


Figure 11 RandomGraph Test

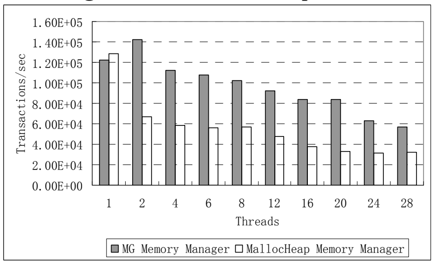


Figure 12 LFUCache Test

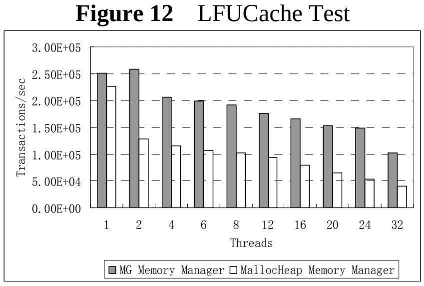


Figure 13 Counter Test

6. CONCLUSION

The memory allocator, nursery and cleaner support a memory management scheme in STMs. The modified EBR garbage collection simplifies the design and improves the performance. The mechanism detecting and handling errors assures the non-blocking synchronization of MG memory manager. Test results indicate the performance of MG memory manager have advantages in memory allocation and reclamation. In the future, we sequentially plan to investigate concurrent control on STM.

REFERENCES

[1] J.R.Larus, R.Rajwar. Transactional Memory[M], New York: Morgan & Claypool, 2007.

[2] H.Sundell. Wait-free reference counting and memory management[C]. In Proc. 19th Intl. Parallel and Distributed Processing Symp., Apr. 2005.

[3] M.M.Michael. Hazard pointers: Safe memory reclamation for lock-free objects[C]. IEEE Trans. on Parallel and Distributed Syst., 15(6):491–504, June 2004.

[4] T.E.Hart, P.E.McKenney, A.D.Brown. Making Lockless Synchronization Fast: Performance Implications of Memory Reclamation[J]. IEEE. 2006.

[5] A.Gidenstam, M.Papatriantafilou, H.Sundell et al. Efficient and Reliable Lock-Free Memory Reclamation Based on Reference Counting[C]. In Proc. 8th I-SPAN, 2005.

[6] V.J.Marathe, M.F.Spear, C.Heriot et al. Lowering the overhead of nonblocking software transactional memory[C]. TRANSACT 06, Ottawa, Canada, 2006.

[7] K.Fraser. Practical Lock Freedom[D]. Cambridge: Computer Laboratory, University of Cambridge. 2004.

[8] M.Herlihy, V.Luchangco, P.Martin et al. Nonblocking Memory Management Support for Dynamic-Sized Data Structures[R]. ACM Transactions on Computer Systems, Vol. 23, No. 2, May 2005:146–196.

[9] M.Herlihy, V.Luchangco, M.Moir. The Repeat Offender Problem: A Mechanism for Supporting Dynamic-sized Lock-free Data Structures[R]. SMLI TR-2002-112, July 2002.

[10] Memory Management in a Massively Parallel Shared-Memory Environment[D]. Notre Dame, Indiana: University of Notre Dame, April 2007.

[11] A.Sarkar. Design and Analysis of Lock-free Data Structures[D]. North Carolina, USA: North Carolina State University. 2007.

Huang GuoRui, born in 1980, is a postgraduate at Zhengzhou Information Technology University. His interests include practical and theoretical aspects of parallel processing. Zhang Ping, born in 1969, is an associate professor at Zhengzhou Information Technology University. Her research interests include parallel computing and compiler optimization. Liu XiaoXian, born in 1985, is a postgraduate at Zhengzhou Information Technology University. Her research interest includes practical of parallel computing. Xiao JinGe, born in 1982, is also a postgraduate at Zhengzhou Information Technology University. His research interests concern practical and theoretical aspects of parallel processing in network system.Wei GuangBo, born in 1977, is a postgraduate at Zhengzhou Information Technology University. His research interests concern practical and theoretical aspects of parallel programming in multi-core processor system.

THE ASIP DESIGN SPECIFICATION BASED ON ADL *

Zhu Yong
College of Computer Science, Wuhan University of Science and Engineering
Wuhan, Hubei, 430073, CHINA
Email: zy@wuse.edu.cn

ABSTRACT

One important challenge of ASIP (application-specific instruction set processors) design is the huge design space, The ASIP design specification based on ADL (Architecture Description Language) be put forward in the paper, specially xpADL and LISA.

Keywords: Specification, ASIP, ADL. Modeling

1. INTRODUCTION

In the last decade, the embedded processor have emerged as the computational core of many digital devices in everyday life. Examples are mobile phones, organizers, personal computers, networking devices and embedded systems for automotives and industrial automation. The capability and complexity of digital devices will continue to grow. However, the growth in design productivity for processor cannot keep up with the technological growth . This gap represents a serious bottleneck for the implementation of new competitive devices.

Traditional system design methods can be divided into three parts: the system specification stage, the hardware / software design phase and integration testing phase. As the misunderstanding of model specification and the definition of hardware and software interface often leads to design process back, that is, fix-in-later problem, a lot of time will be wasted. IP(Intellectual Property)-based design reuse has been proven through repeated use of the IC module, the IP module, which simplifies system design, shorten the design time. IP module that is usually divided into hard IP, soft IP and firm IP. After the design of hardware and software collaboration specification, co-synthesis and co-simulation and evaluation of three phases to complete the system design.

With the complexity of system design and integration, as well as time to market reduce, the design method is faced with enormous challenges. With platform-based design methods functional (application) and implementation (platform) will be separated using structured semantic model to describe. And platform is independent, with use of application software development and configuration platform combination by the application experts rather than experts to complete system hardware design. Platform-based design process include module generation, IP reuse modules, chip integration and software development and so on[1].

Model-driven design has been put forward to address the complexity of modern systems, which is based on an executable model. In the model-driven approach, the design of the first high-level abstraction layer will be captured from the system, designed system can be proved through formal, a lot of automated tools to support the achievement of the code generated.

* This work has been funded by the HUBEI Nature Science Fund under No: 2008CDB332.

The embedded processors provide a significantly better cost/performance ratio than general purpose processors for desktop applications. The so-called application-specific instruction set processors (ASIPs) are able to fill the energy-flexibility gap between dedicated hardware and programmable DSPs for a given application. ASIPs take advantage of user-defined instructions and a user-defined data path optimized for a certain target application.

Application-specific instruction set processors has the potential to become a key building block of future integrated circuits for digital signal processing. ASIPs combine the flexibility and competitive time-to-market of embedded processors with the hould computational performance and energy-efficiency of dedicated VLSI hardware implementations. Furthermore, ASIPs can easily be integrated into existing semicustom design flows: the ASIP designer has full control of the implementation and verification[2].

In the rest of this paper, we first introduce related work about ADLs, then in Section 3 we show ASIP architecture specification, case study by xpADL are presented in Section 4, and finally Section 5 concludes the paper.

2. RELATED WORK

Standard language: the advantages of this method is that the basis of the existing tools can be extended, making the model easy to be combined into a new or meet some new features of the new framework; At the same time, very easy to learn the language itself, and the use of its tools will be easy. Its biggest drawback is that if the model can not be supported by the design characteristics or have been forced into a language structure, it is likely to lose the advantages of the model and design, thereby reducing the effectiveness of description, as well as the readability of documents.

Front-end language: It can be converted into a standard language of the conceptual model of the special. In this process, you can choose the model with the need to match the language, so that to meet the functional language to describe and document the needs, and then into a standard language, which make use of existing tools. Its main drawback is the front-end often need to learn, and the language need to be considered as a result of front-end model and design needs, as well as consider the case of standard language, the middle of the transformation mechanism has a certain degree of difficulty.

Application-specific language: To choose the corresponding language of using the model with the need to direct and design. The advantage of this approach is to ensure that model and the language was a very good match, the system in order to achieve functional expression and documents support. Corresponding to its disadvantage is the need for new tools and learning.

UML as a third-generation modeling language, strictly defined the semantic element model of the object and the object structure and behavior of capture and provide a symbol of communication. Its symbol, using a simple and strict semantic framework for the development of high-performance and high accuracy of the system requirements provides a consistent views. As the UML is the executable, which can be automatically generated executable UML model system[3].

Unified Modeling Language (UML) is widely accepted by the software engineering community. For abstract system conception and definition, it is usual to start with UML use case diagrams to identify the actors, and UML state machine diagrams are used to capture the behavior in each use case. [4]presents design patterns to easily go from the UML system functional description to a SystemC executable model.

The use of XML to describe the structure and parameters of the system architecture model can be a very good match, a valid expression of the realization of the model; developers for their own needs can define your own markup structure to make the form simple, easy to understand description of examples, thus to achieve universal; XML has a clear semantics, analysis can be defined and implemented, to restore the XML document information in a variety environment in order to achieve its clarity; At the same time, opening up many applications in support of XML and schema in order to achieve its simplicity.

SystemCTM has gained much interest in academia and industry all over the world. The main reason for its popularity is its ease of use and flexibility. OSCI organizations supported by the SystemC become the actual executable system design standards, which SystemC TLM (Transaction Level Modelling) Working Group has developed an abstract standard of the Internet, but did not explicitly put forward the function description of acts of abstraction layer. With the use of non-invasive SystemC a PSM model be set up, as well as examples of JPEG coding[5]. A static scheduling have been introduced using par, pipe,fsm and seq constructs that help the designer during architecture exploration and refinement. today. Over the time, SystemC has gained more high-level constructs that make it possible to use SystemC as a “true” SLDL (System Level Design Language).the component language (COLA) is a synchronous data flow language with formal semantics, which is well-adapted to both data flow description and control automata, the generated VHDL code can be synthesized to very efficient FPGA circuits[6]. It has the following five features: ①Designs are modeled at a higher level of abstraction compared to those in traditional HDLs. ②COLA is well-adapted to describing both control automata and data flow. ③COLA offers a graphical representation to specify both functional behavior and structure of the modeled application. ④The system’s correctness can be formally reasoned about. ⑤Since COLA is already used for efficient embedded software development.

The meta-language xHDL was conceived, providing flexible and friendly mechanisms for component parameterization, customization, instantiation and interconnection [7]. It is desirable to provide the highest degree of parameterization of a design, hardware designers appreciate languages and tools that help them when implementing this kind of designs. The design cycle has been significantly reduced even though the parameterization degree has been incremented. xHDL is easy to learn by hardware designers because most constructs of the meta-language have been defined taking into account the VHDL structures. Finally, given a set of parameters related to the xHDL description of a core, the xHDL set of tools has been extended to generate a library of similar components.

3. ASIP ARCHITECTURE SPECIFICATION

The design space for ASIPs has been characterized with the goal to provide well-defined degrees of freedom for the ASIP designer. This characterization enables both the ASIP designers together with the compiler designers to decide, whether specific architectural features of an ASIP are needed and if compiler support for these features is possible. This design process is an important aspect for the design efficiency of ASIPs and is commonly referred to as compiler / processor co-design.

In contrast to previous ASIP design approaches, the ASIP hard and software is in the main design iteration loop of the proposed design methodology. This implies that many of the tedious ASIP design tasks have to be automated to a large extent in order to obtain a short time-to market.

ASIP-specific extensions to LISA tool suite have been introduced covering the optimization of the instruction encoding and the generation of the hardware decoder. In order to transform the LISA ASIP design environment into a design platform for ASIPs a comprehensive library of processor templates and example processor designs for various application domains has to be implemented. These library elements should include LISA descriptions and optimized HLL compilers as well as fully verified hardware descriptions.

4. CASE STUDY BY XPADL

Behavioral ASIP design methodology suit the SoC design and development of the early stages of the entire design. [8]presents a behavioral ASIP design methodology, and a detailed description of a visual architecture description language based on the ASIP design platform xptools behavioral function, structure and working mechanisms. Behavioral ADL design goal is to be accurate, complete, easy to describe the behavior of ASIP-level design all the information needed. In order to act on the instruction set level for a description of behavioral ADL need for instruction set architecture instruction formats, instruction encoding, instruction functions as well as the necessary hardware information (such as memory cells) in the definition.

xpADL using C language to describe the conduct of instruction, will conduct mapping instructions for the corresponding function. With the use of data binding (Java Class Lib by castor project) it complete the analytical description[9].

The "description - search - refinement" of the ASIP architecture orthogonal design method be put forward, design problems will be divided into structure-domain and behavior-domain. The direction of the architecture is divided into processor, pipeline and micro-structure of three levels, the direction of the behavior are classified into behavior instruction set, pipeline and register transfer level of three types behavior. xpADL on the target architecture described in the description of the operation, command description, operation description mapping, structural components of a description of command lines and data transmission path, as well as store-level description of six parts, which constitute the target SoC of the system behavior and structure abstract. Respectively, of the architecture description, architecture design and architecture of the three aspects of optimization technology research environment through the design of the prototype architecture A^2IDE of PISA, ARM and Vcore conducted experiments to verify its effectiveness.

5. CONCLUSIONS

SoC designing concept, using the executable model description Design Methodology and putting the advanced technology of ADL and ASIP into the System Implementation, all that can make a qualitative leap of high-speed channel in performance. ASIPs are instruction set oriented processors with user-defined instructions, a user-defined data path and, optionally, a more dedicated user-defined accelerator. A major obstacle of ASIP design is the larger design space compared to pure hardware or pure software implementations. For this purpose, automatic hardware description generation is needed in order to reduce the time for one design iteration. Architectural ASIP design optimizations are of paramount importance in order to satisfy the constraints of an application.

REFERENCES

[1] Yang Jun，Research on Application Specific Instruction Set Processor (ASIP) Architecture Design，Doctor of Philosophy of UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA，May 2006，pp19-23

[2] Tilman Glökler and Heinrich Meyr，《Design of Energy-Efficient Application-Specific Instruction Set Processors》，KLUWER ACADEMIC PUBLISHERS，2004，ppxi

[3] Bruce Powel Douglass,《Developing Real-Time System with UML，Objects，Frameworks，and Patterns》，Addsion Wesley Longman Inc.，1999，pp3

[4] Joaquín Pérez etal.，System Behaviour Capture: from UML to SystemC，Forum on Specification and Design Languages 2008，pp235

[5] Kim Grŭttner，Wolfgang Nebel，Modelling Program-State Machines in SystemC™，Forum on Specification and Design Languages 2008，pp7-12

[6] Zhonglei Wang etal，A Model Driven Development Approach for Implementing Reactive Systems in Hardware, Forum on Specification and Design Languages 2008，pp197-198

[7] Miguel Angel S´anchez etal，Designing Highly Parameterized Hardware using xHDL，Forum on Specification and Design Languages 2008，pp78-83

[8] Chen Ai etal，Study on Application Specific Instruction Set Processor (ASIP) Behavior Level Design Methodology，Computer Engineering and Applications，2004,29，pp44-47

[9] Su Henian etal，Research on ASIP system-level synthesis approach Based on Language，Microcomputer and Application，2005,9，pp55

[10] Zhu Yong，Design and Research on High-speed Channel of Infrared Image Based on SoC，Doctor of Philosophy of Huazhong University of Science and Technology，2007，pp1

SOFTWARE PRODUCING CHAIN FOR PROCESS MANAGEMENT *

Ji Bian, Haiyang Wang, Yanming Xu, Shaolong Ma, Meng Xu
School of Computer Science and Technology, Shandong University
Jinan, Shandong Province 250101, P.R.China
Email: sdubj@yahoo.com.cn; why@sdu.edu.cn; xuhonk@tom.com; shaolongma@gmail.com; xumeng_wf@163.com

ABSTRACT

To improve efficiency and quality of software production is a serious subject (project). High reliable development tool and integrated environment is the key part of software. It is the highlight point and key base for development of software. Therefore, it has meanings in reality and strategy of long term development. This article includes the aspects of software production chain of process management and combined environment of development, implementation, operation, configuration and administration supporting the process administration application system, to realize the business process dynamic administration, provide convenience for adjustment and creation for business process, increases the business administration vitality, to fully implement the creation and power of decision making. The article is based on software producing chain technology of process administration and based on requirement of business, focus in process management, combines the requirements engineering, domain engineering, domain structure, core frame of process management, engine, tools and various services, forming a software producing chain with the features of united, integrated and visible. Evaluation of reliability and efficiency of the software producing chain is proved, reliability analysis, attestation technology, forming an application system of process management.

Keywords: Process Management, Software Producing Chain, Software Design.

1. INTRODUCTION

At present, business process management has become a basic management requirement of modern enterprises. According to the analysis and prediction of Gartner Group, till 2005, there are 90% big enterprises need operation process management. Those who use hard-code methods to control process or insist manual management will lose the competition to the enterprises using business process management [1].

The traditional software producing chain mainly uses data and functions, neglecting the existence of business process, the process logics are totally fixed and hidden in logics of application. These hidden operation process language has two main obstacles, first the process logics have been listed (coded) in business reality, therefore, it can not separate the process logics from the application, the operation process management can not be realized; once the business process has changed, it must change information system and operation structure; second, because operation semantic is hidden in the information system, it is difficult to prove whether the system designed by the IT experts can meet the operation process requirement defined by operation experts; it also can not change quickly and accurately to locate the operation logics which is required to change according to the changed process[2].

The software producing chain for process management in this article will meet the requirement of operation process of enterprises, integrate the process into the core frame and engine, tools and various service parts, forming a software platform organically. Using the process management producing chain, enterprises can separate the process logics and process logics, to realize dynamic management in business process. It is convenient for adjustment and creation in business process and increase the management vitality, to exert and improve the creation and decision-making ability of the organizations [3].

2. THE OVERALL STRUCTURE OF SOFTWARE PRODUCING CHAIN AND PROCESS MANAGEMENT

The whole software producing environment applies (plug-in) board and modular system [4]. Every part communicates, coordinates, and supports the whole integrated process through abstract service mode. The whole structure chart is as following(Figure 1)：

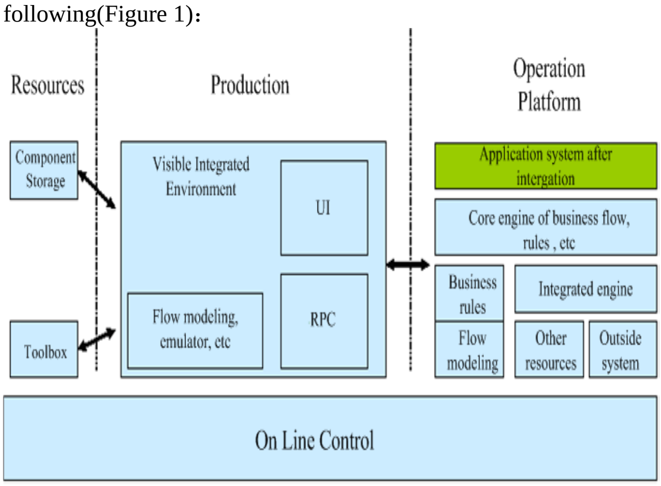


Figure 1. Software Producing Chain of Process Management

Software Producing Chain of Operation Flow Management is divided into four parts in logics [5] [6]:

1) Data (Resource) Base: including Module Base and Tool Collection (Base). Module Base provides various modules, every engine; integrated modules (may) used by producing chain through expending way.
2) Visible Integrated Environment: It provides a visible environment for integration people about data integration tools. The user's interface, data, process, service, and other integration as modules can be finished in this environment.
3) Operation Platform: Operation platform includes bus of service integration, adaptor for application, data module, operation logics module, process module and service module.
4) On Chain Control (Monition): On chain control

* The Development and Research of a Semantic-based Intelligent Virtual Platform for Travel Agency(60673130)

management includes two conditions (time tense) process control for the system while integration development of system and operation.

3. THE KEY TECHNOLOGIES REALIZATION OF THE STRUCTURE

3.1. Visible Integration Environment

The visible integration environment provides an integrated environment of development for user. The Web Application Development for process management can be finished in this environment. It also can integrate the process, report form (statement), data exchange, service pack, portals (gateway) etc, to finish integrated customizing and developing in data interface and process interface used by enterprises. The integration development platform is based on subject 1 Producing Chain Integration Structure, applying Eclipse Plug-in Unit System to realize integration of process modeling, forms setting, statement form development, data exchange, duty setting onto one platform based on application and development functions of J2EE. Figure 2 has shown a whole picture for visible integrated environment of service administration.

Figure 2. Visible Integrated Environment of Service Management

The basic development functions supported by visible integrated environment include:

1) Basic Web Application Development: it provides standard based tool platform through which software developer is able to create professional J2EE and Web application development and allocation debugging.
2) Process Modeling: it provides the function of graphic workflow definition. User can easily realize business flow definition through moving the mouse. Process design saves the process definition under the name of XML in accordance with the XPDL standard, and imports XML defined documents into data base for engine use.
3) Report Form Development: "WYSIWYG" visible report form design tool can complete various report form design tasks, preview the report form effect in the designing process at any time and finish the report form's release, allocation and management.

Application Development and service pack: the development platform reflects SOA concept. It realizes the zero-programming mapping from integrated service model to integrated service operation system. When the coordinated relations between integrated services change caused by the change of demand or environment, simple, fast and high-efficient application reconfiguration can be realized through changing or adjusting the integrated service model.

3.2. The Tool System of Process Management

In term of process management, the tools and methods of workflow management and business flow management, such as process management tools and basic service modules can be applied to realize dynamic and real time monitoring and management towards process integrating process and various activities.

Organizational Model Management Tool: It is used to establish association or mapping between the process participants and the real application operators to describe

1) people's liability, competence and privilege. Organizational rights management model realizes the maintainer of RBAC model based user, organization, role, right, resource and operation, mainly including five components of resource operation management, rights management, role management, user management and organization management.
2) Process Designer (Definition Tool): The major functions of process designer are of describing the business process in graphic form, saving the defined message into the model file of business process, defining the process and related targets, and executing the version control and effective inspection to the process definition.
3) Management and Monitoring Tool: It is a tool to enable the administrator to manage and control each process case in operation by displaying the operational condition of the process case in graphic form and allowing the administrator to change the operational arguments of the process case and control the process.
4) Message Alarm Tool: It allows users to self define message alarm for any processing situation and trigger condition for sending message. There are a variety of message alarming ways including web mail notification (JavaMail), instant information notification (a small tool based on the Windows clients), SMS notification etc.
5) Process Analysis/ Report Tool: It carries out analysis to the business process according to the existing historical data of the process operation in order to provide evidence for management decision-making and reference for process improvement. It deals with statistical analysis to the average completing time of some process definition, the average waiting and processing time of some process definition, the percentage of finishing the process instances and the utilization ratio of statistical resources.
6) Task Management Tool: It is a standard business operational frame to provide business personnel with task processing entrance and realize calls of the business system/ service through enterprise interconnection technology such as Web Service、SCA/SDO、JCA and so on.
7) Unified Authentication and Authorization: It adopts unified authentication and authorization service（CAAS，Central Authentication and Authorization Service）to realize the transparent switchover between many application systems and handle one business processing through the functions supported by different application systems.
8) Workflow Execution Server: it receives the outside service requirement sent by clients and transfers the

requirement to workflow engine for processing. The workflow execution service supplies various basic services of affair management, naming and security test etc for workflow engine.

3.3. The Core Engine for Process Management

The core dispatching engine of software producing chain for process management is formed based on the optimization and improvement of the already existing operation flow engine which supports the Web service（BPMS4WS） and business regulation engine, on the basis of which to design a unified coordinated agreement and combining the core engine and core basic service modules to form a basic operation supporting environment of software producing chain for process management. The specific engines are as follows:

1) Workflow Engine: It is responsible for explaining process definition, controlling process instances, arranging the execution order of activities and adding work programs to the user's worksheet. Workflow engine maintains internal control data which includes internal status message related to various processes or the executing instances. The process definition data coordinates with the workflow related data (during operation period) to control the active navigation in the process, supply the transferring conditions of the activities, the collateral execution of different activities, the sequential execution option, user task, and task related application entrance arguments. If the process definition includes organization model/ role entity type, the organization/ actor model data shall be visited to achieve the above tasks.
2) Task Coordinative Engine: It mainly solves the following issues: Distributing the right thing (business) to the right person or system to deal with at the right time, and providing suitable tool (business system) for the settlers. The task coordinated engine does not only solve the problem of coordinating and distributing tasks between many people for business personnel, but also the problem of coordinating the work between the business system and tool solving business system in the form of process. The task coordinative engine provides the function of managing tasks in both the process and non-process forms.
3) Data Exchange Engine: It is a middleware platform driven by events and based on information agency. It adopts component system structure with each part connecting through standard CORBA IDL interfaces. Its core level is a message agency based on release and subscription which integrates abundant public services such as message routing, permanent storage, event service and notification service and so on. The data exchange engine encapsulates all the message exchanges between the application systems in the XML message form and realizes reliable and unified message transmission and control between the application systems.
4) Real Time Message Service Engine: It is an adapter that joins various real time message tools to the system integrated bus. The adapter should handle the two-way message transmission between the IM tool and other service engines.
5) Report Form Engine: It is a report form service program running on the service party which is applied to analyze report form model, receive report form data, generate report form and supply strong support for the operation, allocation and maintains of the report form. It provides many service entrances such as JSP, ASP, PHP and .NET and so on, supporting both the J2EE application and other various WEB applications.
6) Service Integrated Bus: Based on Geronimo and OSGi (Open Services Gateway initiative）microkernel frame, it supplies the enterprise integration bus for service（ESB：Enterprise Service Bus）. The service integration bus provides intermediary service, transmission service and event service, in which the intermediary service is mainly responsible for the routing between services, conversion between the SOAP standard serving message and ESB standard serving message; the transmission service supplies synchronous and asynchronous transmission service to realize the integration of various communication models; event service is mainly used to make quick reaction to the outside event message, provide the processing model of the event driven system structure to realize the loose coupling interaction of the service interacting parties.

4. CONCLUSIONS

BPM Product enjoys a wide range of categories, but most of them still stay on the first generation workflow management stage. Concerning the functional support of the product system structure and business process management, they are relatively far way off the software producing chain about the process management mentioned in the thesis. Based on the SOA structure, the software producing chain for the process management provides abundant management tools and service modules. It boosts the advantages of platform independence, visibility, flexibility and universal property etc, which make it enjoy relatively robust competition in the market.

BPM pays attention to the application integration of the business process rank, while the combination of process and service is one large-scale developing object of BPM at present. After experiencing the two revolutions of structuring and object orientation, the software development is welcoming the developing upsurge of service orientation, and the tendency several years later will be the procedure oriented software development.

REFERENCES

[1] S. Smith, D. Neal, L. Ferrara, F. Hauden. The emergence of business process management. CSC's Research Services. 2002.
[2] Yuhong Yan, Zakaria Maamar, Weiming Shen. Integration of Workflow and Agent Technology for Business Process Management. The Sixth International Conference on CSCW in Design. 2001.
[3] Chen Minder, Zhang Dongsong, Zhou Lina, Empowering collaborative commerce with Web services enabled business process management systems. Decision Support Systems, v 43, n 2, March, 2007, p 530-546
[4] Business Process Execution Language for Web Services (BPEL), Version 1.1. http://www.ibm.com/developerworks/library/ws-bpel, May 2003.
[5] D. Shuiguang, W. Zhaohui. Managing Serviceflow in a Flexible Way. The Fifth International Conference on Web Information Systems Engineering. 2004
[6] Bart Orriens, Jing Yang, Mike P. Papazoglou: A Framework for Business Rule Driven Service Composition. TES 2003:14-27

RESEARCH ON THE TRANSFORMATION FROM ENTITY DATA MODEL TO RELATIONAL DATA MODEL *

Liu Yang , Liu Hongxing , Simi Neli
School of Computer Science & Technology, Wuhan University of Technology
Wuhan, Hubei 430063, P.R.China
Email: ayang@whut.edu.cn; liuhongxing@whut.edu.cn ; simi_nelly82@yahoo.com

ABSTRACT

Entity Data Model (EDM) is a new kind of executable conceptual model developed by Microsoft. 'Executable' means that the data in a relational database can be manipulated directly based on EDM, for which Microsoft has provided tools to support the transformation from relational data model (RDM) to EDM. As a conceptual model, EDM is plat-independent and with rich semantics, so it is also suitable for conceptual data modeling. To make use of EDM in database design, the transformation from EDM to RDM is important. In this paper, the mapping rules from EDM to RDM are studied based on the correspondences between EDM and RDM. Then a transformation algorithm for transforming EDM to RDM is presented. The research in this paper would make EDM more useful in the design and integration of information systems.

Keywords: Entity Data Model, Relational Data Model, Model Transformation, Mapping Rules, Algorithm

1. INTRODUCTION

In the process of traditional software development, the Entity-Relationship (ER) model [1] introduced by P.P. Chen is widely used for database conceptual modeling. But an ER conceptual model is only a design artifact, which could not be used as an executable runtime artifact. In order to make the conceptual model play a more important role, “Executable” conceptual model is presented, and becomes an important research direction in the field of the database and conceptual modeling [2]. In recent years, some new data models, such as Executable UML [3] and Visual SQL [4], have been developed; these make Executable Conceptual Model getting practical.

Entity Data Model (EDM) is a new kind of executable conceptual data model, which is developed by Microsoft and has been supported in Visual Studio .NET 2008. Currently most of the data in enterprises are managed by using relational database systems; so the Relational Data Model (RDM) is the most important logical data model. A logical data model is platform-dependent, for they may be supported by different Database Management Systems. To integrate information and to eliminate the impedance mismatch for both applications and data services, it is necessary to raise the level of data abstraction from the logical (relational) level to the conceptual level [5]. Microsoft provide tools to support the transformation from RDM to EDM, as shown in Figure 1, this makes the EDM ‘Executable’, i.e., the data in a relational database can be queried and manipulated directly based on EDM.

On the other hand, EDM as a conceptual model is platform-independent and with rich semantics, so it is also suitable for conceptual data modeling. In order to make use of EDM in database design, it is necessary to perform the transformation from EDM to RDM. So we study the transformation in this paper. The mapping rules from EDM to RDM are studied based on the correspondences between EDM and RDM. Then a transformation algorithm for transforming EDM to RDM is presented. We believe the research in this paper would make EDM more useful in the design and integration of information systems.

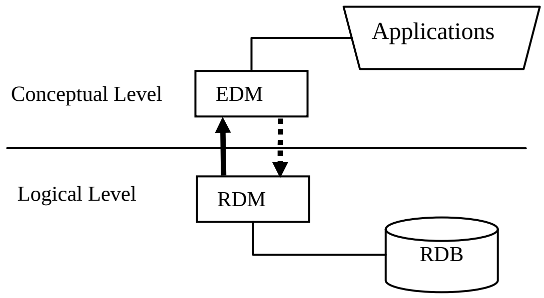


Figure 1. EDM and RDM

The rest of the paper is organized as follows. Section 2 analyses the structure and the representation of EDM and RDM. Section 3 introduces the mapping rules between EDM and EDM; this is the foundation for model transformation. Section 4 presents a transformation algorithm which can transform an EDM to the corresponding RDM. Section 5 concludes the paper.

2. ENTITY DATA MODEL AND RELATIONAL DATA MODEL

2.1. Entity Data Model

EDM is a conceptual model and it is based on the ER model. It extends the classic relational model with concepts from ER model. The central concepts in EDM are entities and associations. Entities represent top-level objects with independent existence and identity, while associations are used to relate two or more entities. EDM models can be represented in two ways; one is by EDM diagrams which are visual and are very similar to the ER diagrams; the other is by using Conceptual Schema Definition Language (CSDL) which is introduced by Microsoft and is based on XML.

An EntityType in CSDL describes the definition of an entity. An entity is a top-level object with independent existence, it has a payload: Zero or more properties that describe the structure of the entity. Additionally, an EntityType must define a key, i.e., a set of properties whose values uniquely identify the entity instance within its container. An EntityType may derive from other EntityTypes. EDM only supports a single inheritance model.

Entity instances (or just entities) are logically contained within an EntitySet. An EntitySet is a homogenous collection of

* This work is supported by the Science& Technology Foundation of Wuhan University of Technology (No.471-38650243)

entities. All entities in an EntitySet must be of the same (or derived) EntityType. An entity instance must belong to exactly one entity set. Figure 2 is an example of EDM diagram.

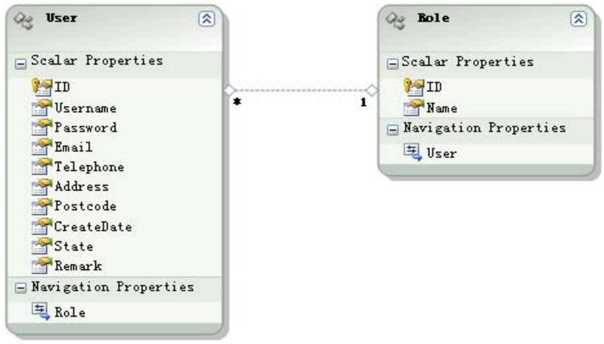


Figure 2. A Simple Example of EDM Diagram

The following is the CSDL definition of the EntitySet User and Role:

```
<EntityContainer Name="Model1Container" >
        <EntitySet Name="User” EntityType="Model1.User" />
        <EntitySet Name="Role" EntityType="Model1.Role" />
</EntityContainer>
```

In this example two EntitySet, User and Role, are defined, the corresponding entity types are User and Role. The EntitySet element is used to define the EntitySet, it is contained in the EntityContainer element.

Property and primary key are mainly defined in EntitySet element. In EDM, the definition of each entity contains two different types of members, one is scalar property which is similar to column in database; and the other is navigation property which is similar to foreign key in database. Moreover, every entity must have an identifier or primary key. The following is the CSDL definition of Role entity in figure 2:

```
<EntityType Name="Role">
   <Key>
        <PropertyRef Name="ID" />
   </Key>
   <Property Name="ID" Type="Int32" Nullable="false" />
   <Property Name="Name" Type="String" Nullable="true" />
   <NavigationProperty Name="User"
     Relationship="Model1.RoleUser" FromRole="Role"
     ToRole="User" />
</EntityType>
```

The definition of scalar property usually includes Name, Type, Nullable, etc. The Property element is used to define the scalar property; it is contained in the EntityType element. For Example, there is a scalar property named ID in Role entity, its type is Int32 and it can’t be null. The following is the CSDL definition of this scalar property:

```
<Property Name="ID" Type="Int32" Nullable="false" />
```

The definition of navigation property is according to the definition of association. The navigation properties are automatically generated in each entity after associations are defined. The NavigationProperty element is used to define the navigation property; it is also contained in the EntityType element. The definition of navigation properties usually includes Name, Relationship, From Role, To Role, etc. For example, there is a navigation property User in Role entity, the following is the CSDL definition of this navigation property:

```
<NavigationProperty Name="User"
Relationship="Model1.RoleUser” FromRole="Role”
ToRole="User" />
```

The Key element is used to define the primary key. The primary key references one or more scalar properties. For example, the property of ID in Role entity is the primary key, the following is the CSDL definition of this primary key:

```
<Key>
     <PropertyRef Name="ID" />
</Key>
```

In EDM, associations are used for modeling the logical connections between entities. EDM supports the AssociationType of associations, which models the peer-to-peer associations between entities.

In an association, each participating entity is called an End. Each End has a Role attribute that is used to name and describe the logic of each End of the association. Type attribute of an End element identifies the entity type of the End element. Associations have a Multiplicity attribute, which specifies the number of instances of each End that can participate in association.

An association has the following characteristics [6]:
(1)Multiplicity: Association End elements can be 0..n, where the cardinality of either End can vary.
(2)Non-Exclusive Membership: An entity type can participate in multiple associations.
(3)Direction: Associations are bi-directional between entities and can be navigated from either End.

All the homogeneous associations’ collections construct an AssociationSet. The following are the CSDL definition of an association between User and Role in Figure 2:

```
<EntityContainer Name="Model1Container" >
   <AssociationSet Name="RoleUser"
Association="Model1.RoleUser">
       <End Role="Role" EntitySet="Role" />
       <End Role="User" EntitySet="User" />
   </AssociationSet>
</EntityContainer>
<Association Name="RoleUser">
   <End Type="Model1.Role" Role="Role" Multiplicity="1" />
   <End Type="Model1.User" Role="User" Multiplicity="*" />
</Association>
```

2.2. Relational data model

Relational Data Model (RDM) is composed of three parts: relational data structure (relation), relational operation and relational integrity constraint. In the process of transformation, relational data structure and relational integrity constraint are mainly considered. The logical structure of the relation is a two-dimensional table. The relational integrity constraint includes entity integrity, referential integrity and user-defined integrity. The entity integrity and referential integrity must be met in RDM, so entity integrity and referential integrity are mainly considered in the transformation from EDM to RDM.

3. MAPPING RULES

3.1. The mapping rules of entities

Entities in EDM are corresponding to the tables in EDM, so every entity in EDM can map to a table in RDM.

<Rule 1>: According to the definition of EntityType and the definition of element Name in EntityType, an entity in EDM is mapped to a table in RDM, while the table name is assigned by the attribute value of Name.

As an example shown in figure 2, according to <rule 1>, Role entity is mapped to a table named Role.

3.2. The mapping rules of attributes

The attributes of entities in EDM are corresponding to the columns of tables in RDM. So the attributes of every entity can map to columns of corresponding table in RDM.

<Rule 2>: According to the definition of Property of each EntityType, each attribute of entity in EDM is mapped to a column of corresponding table in RDM while the column name is assigned by the attribute value of Name in Property element, the column's data type is assigned by the attribute value of Type, and the column's mandatory is assigned by the attribute value of Nullable.

As an example shown in Figure 2, according to <Rule 2>, the Role table will be composed of two columns: ID and Name. The data type of ID should be Int32 and the value of ID cannot be null; the data type of Name should be Varchar and the value of Name can be null.

3.3. The mapping rules of Primary key

The primary key of entities in EDM is corresponding to the key of tables in RDM.

<Rule3>: According the value of Name in PropertyRef element which is in the Key element of EntityType, the primary key of table is assigned by the attribute value of Name in PropertyRef element which is defined in the corresponding entity.

As an example shown in Figure 2, according to <Rule 3>, the column ID is set as the primary key of Role table.

3.4. The mapping rules of associations

Association in EDM is corresponding to the foreign key constraint in RDM. So each association can map to a foreign key constraint in RDM.

An association connects two entities, so the association could be mapped to a foreign key constrain only after the two entities have been mapped to the tables. This is the precondition for using the following rules.

<Rule 4>: According to the content of an Association element: if the values of Multiplicity in both entity in the association are 1 or 0..1 (i.e., 1: 1 Association), then the primary key of one table should be added to the other table as the foreign key. The data type and mandatory of the foreign key are as same as that of the primary key.

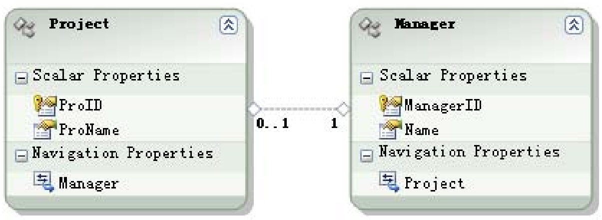


Figure 3. 1:1 association

In the example shown in Figure 3, the entity Project and Manager should have been mapped respectively to the table Project and Manager according to <rule 1>. Next, according to <Rule 4>, the primary key ManagerID in table Manager should be added to the table Project as the foreign key.

<Rule 5>: According to the content of an Association element: If the value of Multiplicity in one entity is 1 or 0..1 and the other is * (i.e., 1: N Association), then the primary key of the table on 1/0..1 side should be added to the table on * side as the foreign key. The data type and mandatory of the foreign key are as same as that of the primary key.

In the example shown in Figure 2, the entity User and Role should have been mapped respectively to table User and Role according to <Rule 1>. Next, according to <Rule 5>, the primary key ID in table Role should be added to the table User as the foreign key.

<Rule 6>: According to the content of an Association element: if the values of Multiplicity in both entities are * (i.e., M: N Association), then a new table should be added to the RDM, and the primary keys in both entities should be combined as the primary key of the new table. The two primary keys are also foreign keys, and the data type and mandatory of each foreign key are as same as that of the primary key.

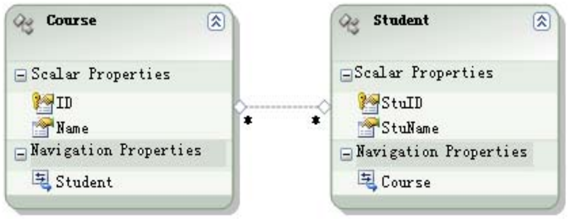


Figure 4 M: N Association

In the example shown in Figure 4, the entity Course and Student should have been mapped respectively to table Course and Student according to <Rule 1>. Next, according to <Rule 6>, a new table is added to EDM, the Primary key ID in table Course and the Primary key StuID in table Student are combined and added to the new table as the Primary key. The column ID in the new table references the Primary key ID in table Course, and the column StuID in the new table references the Primary key StuID in table Student.

4. TRANSFORMATION ALGORITHM

To transform EDM to RDM, an algorithm is presented as follows, which has three steps. The mapping rules defined above will be used in the algorithm. The input of the algorithm is an EDM model defined in CSDL, and the output is a RDM model.

STEP 1: Read the content of EntityType element; transform each entity to a table in relational schema according to <Rule1>. Then read the content of Property element in each EntityType, transform each property to column in corresponding table according to <Rule 2>.

STEP 2: Read the content of Key element in each EntityType, define the primary key constraints in each table according to <Rule3>.

STEP 3: Read the content of Association element, according to the definition of Multiplicity in each entity in association, transform EDM to RDM as follows:

(1)If the values of Multiplicity in each entity in an association are both 1 or 0..1, define a foreign key constraints according to <Rule 4>.
(2) If the value of Multiplicity in one entity in association is 1 or 0..1 and the other is *, define a foreign key constraints according to <Rule 5>.
(3) If both values of Multiplicity in each entity in association are *, define an extra table and corresponding foreign key constraints according to <Rule 6>.

Base on the three steps above, most of the EDM models can be transformed to the corresponding RDM models. The RDM model can be described by using SQL. EDM models are described in CSDL which is a XML language, so we can use XML API, such as DOM, to process the EDM models. There are some limitations in our transformation solution, for example, the user-defined integrity is not considered yet.

5. CONCLUSIONS

EDM is an executable conceptual model which raises the level of data abstraction from the logical level to the conceptual level; this is useful to information integration and application development. However EDM as a conceptual model could also be useful in database design, for this purpose the transformation from EDM to RDM is very important. In this paper, the mapping rules and a transformation algorithm are presented which would support the automatic transformation from EDM to RDM and make EDM more widely accepted as a conceptual model in database design.

We intend to extend the research in two directions. Firstly, we are planning to improve and refine the mapping rules and the transformation algorithm to make the solution more elaborate. Secondly, we will develop a visual tool which is able to implement the transformation automatically.

REFERENCES

[1] Peter P. Chen. The Entity-Relationship Model: Toward a Unified View of Data. ACM Transaction on Database System, 1(1):pp.9-36, 1976
[2] Peter P. Chen, Bernhard Thalheim, Leah Y. Wong. Future Directions of Conceptual Modeling. Lecture Notes in Computer Science 1565, pp.287-301, Springer, 1999
[3] S. Mellor, M. Balcer. Executable UML: A Foundation for Model-Driven Architectures. Addison-Wesley, 2002
[4] Berbhard Thalheim. Visual SQL: Toward ER-Based Object-Relational Database Querying. Lecture Notes in Computer Science 5231, pp.520-521, Springer, 2008
[5] Jose A.Blakeley, S.Muralidhar, Anil Nori. The ADO.NET Entity Framework: Making the Conceptual Level Real. Lecture Notes in Computer Science 4215, pp.552-565. Springer, 2006
[6] Microsoft. Association (EDM). http://msdn.microsoft.com/en-us/library/bb399198.aspx, 2009.6

Liu Yang is a graduated student in School of Computer Science and Technology, Wuhan University of Technology. His research interests are in conceptual modeling and model transformation.

Liu Hongxing is a Professor and the vice dean at School of Computer Science and Technology, Wuhan University of Technology. He received a PhD in Computer Software and Theory from Huazhong University of Science and Technology (2006). His research interests are in database, conceptual modeling and information system.

AN EXTENSIBLE STATIC ANALYSIS METHOD USING DEFECTS DESCRIPTION LANGUAGE

ZHU TIAN QING
Department of Computer Information Engineering, Wuhan Polytechnic University
Wuhan, 430040, China
E-mail: zhutq@126.com

ABSTRACT

As an important software defects analysis method, static analysis is wildly used in software coding and test. But this traditional method is inextensible and inefficient that users have to find defects artificially. A framework is implemented to locate defects called DDS(Defects description system) and an annotation Language was designed to describe defects which named DDL(Defects Description Language). The rules written in DDL will be saved in rule set and can be used by DDS to locate the defects automatically. The experimental results demonstrate that the method is available to identify defects in programs.

Keywords: Static Analysis, Defects Description, Syntax Tree Matching

1. INTRODUCTION

For a safe software, each input should bring a precise output. Any defect should be eliminated because it will cause an unexpected result[1].

Many types of security defects can be detected in extended static checking at compile phase[2]. Given semantic rules, software attributes in each executing status can be found in static analysis rapidly in automatic. Meta-Compilation, a static analysis tool developed by Prof. Dawson Engler of Stanford University, has identified hundreds of defects in Linux kernel[3]. Hao Chen and David Wagner built a steady framework for static analysis and developed an analysis software named MOPS[4]. An extensible system was presented in [5] that users could define test rules themselves in annotation language to find more defects. Crispin Cowan indicated the methods to find out the defects of buffer overflow in programs[6,7]. David Wagner gave a new method to test dynamic memory fault[8], and alarm optimizing method was proposed for the sake of reducing false positives[9].

In [10], a checking method is presented to detect the out-of-bounds accesses with conditional range constraint.

However, defects analysis becomes more complicated the ever before because that software is always multifunctional with great capability nowadays. Traditional analysis methods are inextensible and inefficient that users have to find defects artificially. We need implements of superior quality to find the defects in programs with higher executing speed and extra precision.

To solve the problems, a rule-driven system named DDS(Defects Description Systems) is described in this paper. In this extensible system, user can define rules by describing defects in DDL(Defects Description Language) at first. With the rules, the defects in programs could be located by syntax tree matching.

The subsequent sections are organized as follows. Section2 presents the defects description system; Section 3 describes the defects description language with an lock_checker example; The functions of DDS Automaton are discussed in section 4; Section 5 describes the experimental results of defects analysis by DDS; Conclusions are given in the last section.

2. DEFECTS DESCRIPTION SYSTEM

2.1 Basic Modules of DDS

DDS is composed of four modules including the information extraction module, the rules module, the analysis module and the results module, as shown in figure 1. The information extraction module processes the source code with lexical analysis and syntax analysis. The open source software such as Flex and Bison can be used in the information extraction module. In the rules module, detect rules are defined by users. The analysis module is the most significant part of the system which locates defects in the input data from the information extraction module and the rules module. Finally, detect results are output in the results module.

We developed an experimental system of DDS with C language. The information extraction module is composed by GCC4 and DSS Back-End parser. GCC4 transforms source code to intermediate code without any additional syntax analysis and semantic analysis. Back-End parser of DSS produces symbol pair for analysis module.

Rules Module is composed of rule file and DSS Front-End parser. Rules can be written by users according to their requirements. DSS Front-End parser can read rules and create DSS intermediate code for analysis module.

Analysis module is an automaton which performs states transition. Automaton includes a series of internal functions which accomplish defects matching and location.

Result Module can output the defects located in source code in text files or XML files.

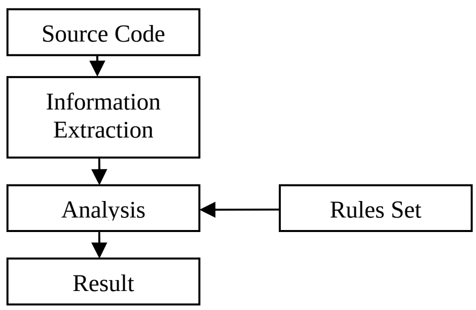


Figure 1. Components of DDS

2.2 Workflow of DDS

Before reading source codes, Front-End parser reads defects rules from the rule module and generates DDS rule codes which will be registered in automaton. The rules stored as syntax tree and matching functions are defined by the registration.

Then GCC4 reads the source code and creates the intermediate codes after lexical analysis and syntax analysis.
DSS Back-End parser reads the intermediate code and creates pointer pairs of the object variables and the object statements. According to the pointer pairs, the syntax tree will be established and put into the automaton.

In the next step, the syntax tree from DSS Back-End parser will be matched with the one generated by Front-End parser. The matching functions which has been registered in the automaton will complete the matching process. Once the trees are perfectly matched, the automaton will transit its state and output its results.

When each step is finished and the automaton reaches the end state, the halt function will be called to terminate the performance.

The whole workflow of DDS is shown in figure 2.

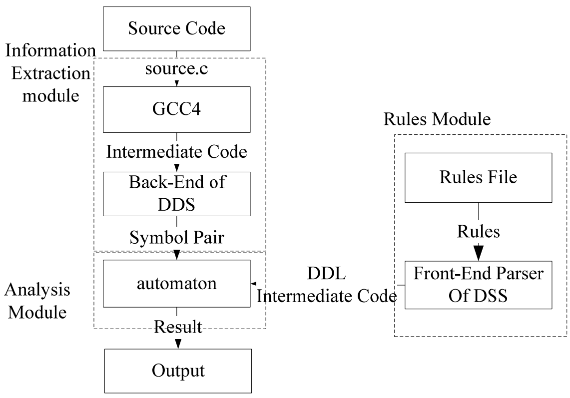


Figure 2. Architecture and Flow of DDS

3. DEFECTS DESCRIPTION LANGUAGE

3.1 Syntax and Grammar

In DDS, an annotation Language is designed to describe defects which named DDL. Annotation keywords are defined by the extension writer. They must be declared before use, typically they reside in a header file that is included by the checked code. Undefined annotations are flagged as errors to catch misspelling and mismatches between annotations and extensions.

The following are some examples of keywords used in DDL:
st: keyword of automaton status transition rules.
Option: keyword of automation option which followed by a specific string. For example, the string "sensitive" denote to generate direction information.

Declare: declare placeholder variables which only care about location and type.

declare_sv: declare automaton variable which not only care about location and type but also status.

DDL is a kind of similar C language. We can use DDL to write rules and store them in the rules module. DDS Front-end parser will read certain rules from the file when users need to check corresponsive defects.

Here is an lock_checker example written in DDL:

```
st lock_checker{              // rule name
option sensitive;             // generate direction information
declare_sv { int }     lk;    // declare automaton variable
                                 named lk
lk.init: { try_lock(lk) != 0 }  => lk.locked
         { } // initial state, locked successfully, transit to
              locked status
lk.init: { try_lock(lk) = = 0 }      => lk.unlocked
         { } // initial state, locked faulty, transit to
              unlocked status
lk.init: { lock(lk) }                      => lk.locked
         { } // initial state, locked, transit to locked status
l.init: { unlock(lk) }              => lk.error
     { err("unlock error for %s\n", getid(); backtrack(); )
         // initial state, try to unlock，show error and
           backtrack information
l.locked: { lock(lk) }              => lk.error
     { err("double lock for %s\n",getid(); backtrack(); )
          // locked state, keep on locked, show error
            and backtrack information
lk.locked: { trylock(lk) }          => lk.error
     { err("double lock for %s\n",getid(); backtrack(); )
              // locked state, try to lock, show error
              and back trace information
lk.locked: { unlock(lk) }           => lk.init
         {} // from locked state to unlocked initial status
lk.locked: { END }                  => lk.error
         { err(path ends with lock held for %s\n", getid();
backtrack(); )}
         // halt in locked state, lock has not released, show
          error and backtrack information
```

This rule can be used to check the lock status of critical resource and locate the locked defects by reading the code.

3.2 Rules Parse

Syntax analysis and lexical analysis will be performed during rules parse. We can use open source tools such as Flex and Bison.

Flex can create the annotator list of the rules for DDL intermediate code.

Bison can be used to create a syntax parser as long as users support its regulation. The goal of syntax analysis is describing variable state transition map, thus the rules generated by Bison should meet some basic requirements as follows:

- The name of each state should be defined.
- The conditions of state transition should be given.
- The additional operation after state transition should be designed.
- The additional operation of halt state must be defined.

Applying the rules generated by Bison, familiar C statements(such as arithmetic operations, function calling, pointer citation, assignment statement and etc.) could be expressed.

The rules will be converted to DDL intermediate code after parsing.

3.3 DDL Intermediate Code

DDL intermediate codes include the annotator lists and the syntax trees. Respectively, the annotator lists are generated by lexical analysis and syntax trees are generated by syntax analysis. They are compared by matching functions and executed by other operating functions.

For example, the locked rule can be convert to syntax trees as follows:

```
l.locked: { lock(lk) }                => lk.error
{ err("double lock for %s\n",    getid());    backtrack(); }
```

The matching syntax tree and the operating syntax tree are shown in figure 3 and figure 4 respectively.

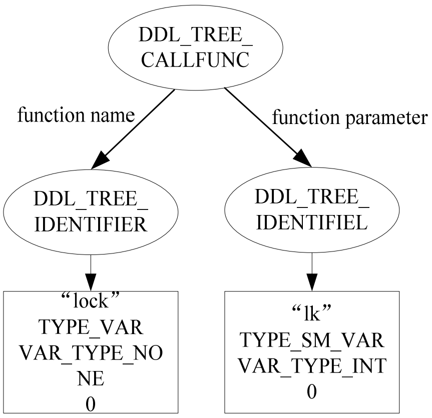


Figure 3. Matching Syntax Tree

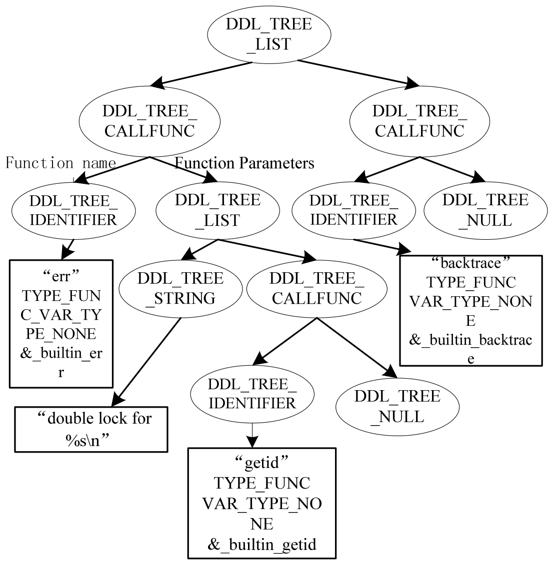


Figure 4. Operating Syntax Tree

The trees will be put into the analysis module to match the trees generated by source code.

4. DDS AUTOMATON

Automaton is composed of series of internal functions such as matching functions, halt functions, backtrack functions and etc. These functions are used to perform the operations including pattern matching, execute operation, defects backtrack and so on.

4.1 Matching Function

Matching functions are the kernel of the automaton. They are used to judge weather the codes being detected is matched with the mode of DDS. And the automaton will switch its state according to the matching results.

Because that the two syntax trees are in different format, the matching functions must access them at the same time in recursive way.

Let Gtree represent the syntax tree generated by GCC and Dtree represent the one generated by DDS, the matching flow is described as follows:

1. If the types of Gtree and Dtree are the same:

A)if the both trees are not function trees, compare the subtrees of Gtree with the ones of Dtree respectively, return success if they are all matched exactly.

B)if the both trees are function tree, not only the subtrees but also the relationship between the parameters of Gtree and Dtree should be compared. For example, foo(p,p,any) is matched with foo(pTree,pTree,NULL) but not matched with foo(pTree,pChild,NULL).

2. If the types of Gtree and Dtree are not the same, compare the subtrees of Gtree with the ones of Dtree and return success if any subtree is matched.

Return failure if match failed in step 1 or step 2.

As a case, the matching process between *p and ret=func(t,*t,0) is shown in figure 5.

4.2 Operation Functions

Operation functions will be called if the matching function returns a success result. The prototype of an operation function is defined as:

void trans_hook(const Automata *atmt, void *param), where param is a pointer which points to operation trees.

Backtrack functions are important in operation functions. It can locate the state conversion in source code.

For example, the following is a running result returned by the backtrack function.

```
======= Backtrack start =======
No.1: var p(file:/tmp/malloc.c, line 105) init => unknown
No.2: var p(file:/tmp/malloc.c, line 126) unknown => init
No.3: var p(file:/tmp/malloc.c, line 117) init => error
======= Backtrack stop ========
```

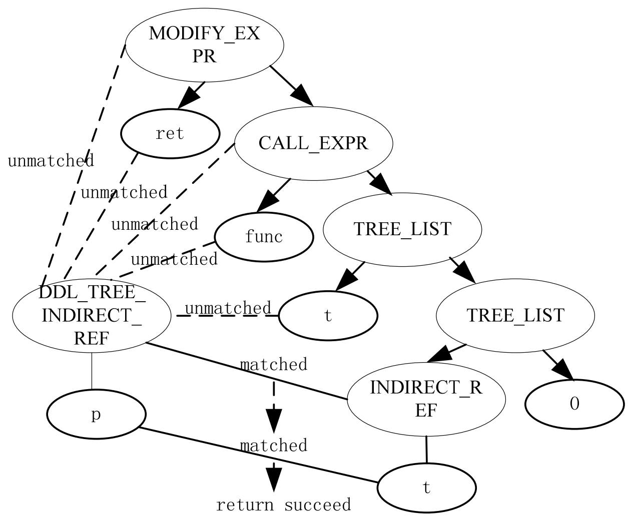


Figure 5. Matching Function

5. EXPERIMENTAL RESULTS

To test the performance of DDS, an actual source codes named slip.C comes from Linux-2.3.5 /driver/net is used in our experiments.

After analysis by DDS, two spin lock defects are detected in the source codes.

The result is shown as follows:

>>>>>>>>> function : sl_ioctl <<<<<<<<<<<<<<<<

```
init==>locked
lock hold when leave.
======= Backtrack start =======
No.1: var sl->lock(file:slip.c, line 1255) Init => Locked
=======Backtrack stop ========

init==>locked
locked==>init
lock unlocked var.
======= Backtrack start =======
No.1: var sl->lock(file:slip.c, line 1255) Init => Locked
No.2: var sl->lock(file:slip.c, line 1275) Locked => Init
No.3: var sl->lock(file:slip.c, line 1313) Init => Error
======= Backtrack stop ========
```

Defects could be located by backtrack information:

1．Halt after locked

```
1248  static int sl_ioctl(struct net_device *dev,struct ifreq *rq,int cmd){
……
1255  spin_lock_bh(&sl->lock);        //  locked
……
1262  switch(cmd){
1263        case SIOCSKEEPALIVE:
1264        /* max for unchar */
1265        if (((unsigned int)((unsigned long)rq->ifr_data)) > 255)
1266            return -EINVAL;        //  forget unlocked
……
1282     case SIOCSOUTFILL:
1283        if (((unsigned)((unsigned long)rq->ifr_data)) > 255) /* max for unchar */
1284        return -EINVAL;
                                   //  forget unlocked
```

2．Unlocked twice

```
1248  static int sl_ioctl(struct net_device *dev,struct ifreq *rq,int cmd){
……
1255  spin_lock_bh(&sl->lock);        //  locked
……
1262  switch(cmd){
1263        case SIOCSKEEPALIVE:
……
1275            spin_unlock_bh(&sl->lock);
                                   //  unlocked
1276            break;
……
1312  };
1313  spin_unlock_bh(&sl->lock);        //unlocked
1314  return 0;
1315 }
```

DDS located the spin lock defects according to the rules defined this kind of defects. Similarly, other defects could be located as long as they are defined in relevant rules.

6. CONCLUSIONS

DDS can check defects in actual codes by defining in rules more than a prototype. It has the same description faculty using variables automaton as other analysis tools like MOPS and MC. MOPS performs its checking using language stipulation with more efficient but high fault positive. MC can execute fleetly. But DDS can localized defects more accurately by using backtracked functions in automation. DDS has the flaws just like other analysis systems which cannot differentiate the defects determined on running[11]. Therefore, using one front-end to parse rules and code can lessen the cost of the syntax trees matching.

In addition, optimizing the pointers handle and improving the global analysis ability will consummate the DDS for more efficient checking.

REFERENCES

[1] M Wood, M Roper, A Brooks, J Miller. Comparing and combining software defect detection techniques: a replicated empirical study [J]. ACM SIGSOFT Software Engineering Notes,1997,22(6):262-277

[2] Ken Ashcraft, Dawson Engler.Using Programmer-Written Compiler Extensions to Catch Security Holes[C]//Proceedings of the 2002 IEEE Symposium on Security and Privacy, IEEE Computer Society Washington, DC, USA:IEEE, 2002:143

[3] ED Willink.Meta-Compilation for C++ [D]. UK:Computer Science Research Group, University of Surrey, 2001

[4] Hao Chen, David Wagner. MOPS:an Infrastructure for Examining Security Properties of Software [R].Berkeley, CA, USA:University of California at Berkeley,2002

[5] Junfeng Yang, Ted Kremenek, Yichen Xie et al. MECA: an Extensible, Expressive System and Language for Statically Checking Security Properties [C]//Conference on Computer and Communications Security Proceedings of the 10th ACM conference on Computer and communications security, Washington D.C., USA:ACM, 2003:321-334

[6] Crispin Cowan, Perry Wagle, Calton Pu, et al. Buffer Overflows: Attacks and Defenses for the Vulnerability of the Decade [C]//DARPA Information Survivability Conference and Exposition,SC,USA:2000:119-129

[7] David Larochelle, David Evans. Statically Detecting Likely Buffer Overflow Vulnerabilities [C]//Proceedings of the 10th USENIX Security Symposium, Berkeley, CA, USA:USENIX,2001:14-14

[8] Wei Le, Mary Lou Soffa. Marple: a demand-driven path-sensitive buffer overflow detector [C]//Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering, NY:ACM, 2008:272-282

[9] Sarah Heckman, Laurie Williams. On establishing a benchmark for evaluating static analysis alert prioritization and classification techniques [C]//Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement, NY:ACM2008:41-50

[10] Xia Yimin, Luo Jun, Zhang Minxuan. Detecting Out-of-Bounds Accesses with Conditional Range Constraint [J].Journal of Computer Research and Development,2006,10:1760-1766

[11] M. Pistoia, S. Chandra, S. J. Fink, et al. A survey of static analysis methods for identifying security vulnerabilities in software systems [J].IBM Systems Journal,2007, 46(2):265-288

APPLICATION OF WIRELESS TERMINALS IN CONTAINER TERMINAL MACHINE DISPATCH SYSTEM

Xing Guo, Guoxian Wang, Hanbing Xiao
School of Logistics Engineering
Wuhan University of Technology, Wuhan, Hubei Province, China
Email: guoxing1967@sohu.com

ABSTRACT

Since container terminal is the interchange hub between the water and land transportations, its operation efficiency directly affects the goods flow rates and port operation costs. Application of wireless terminals to machine dispatch systems in container terminals will make port operation efficiency enhanced and costing reduced. This article depicts a case study of design and development of a container terminal machine dispatch system using wireless terminals and a range of significant issues involved are addressed.

Keywords Wireless terminal, Container terminal, Machine dispatch system

1. INTRODUCTION

In traditional container terminals, task assignment for yard cranes, quayside container cranes and yard trailers relies on information exchange by the help of the talk-back system, which is not only of low-efficiency, but also prone to bring about wrong operations due to exchange and transmission of wrong information[1]. The customers cannot be promptly served and the port operational yields cannot be maximized. Without corresponding hardware systems, the relevant information cannot be dynamically transmitted among the yard cranes, the quayside container cranes, the yard trailers and the control centre, even if highly-advanced managerial software systems are deployed. In order to improve the port operation efficiency, a new trend can be observed that the yard cranes, the quayside container cranes and the yard trailers tend to be equipped with wireless terminals with relevant accessories so that a real-time scheduling and management system is built up using the wireless network technologies[8-9, 2-3].

2. SYSTEM COMPOSITION AND FUNCTIONAL ANALYSIS

In this study, the system adopts C/S-based terminal simulation mode[6,7]. Data rendering and browsing through terminals are supported while the response speed is fast even if multiple users are online. Strong data handling and analyzing ability is provided on the server side. As shown in Fig. 1, the full application system is composed of a Teklogix Wireless Terminal Series (including vehicular 8560, vehicular 8255 and portable 7035), A Wireless Base Station (Base Station 9150), a Server (with database modules included).

The operation command sequences consisting of all port operations, including loading, unloading, container gate-in and container relocation, etc., are able to be sent out from the control centre and then rendered on the mobile computing terminals of vehicular or portable wireless terminals. Crane and trailer operators/drivers or tally clerks will receive real-time operation commands which are then strictly executed. The empty driving and waiting time will be shortened and the operation efficiency will be improved.

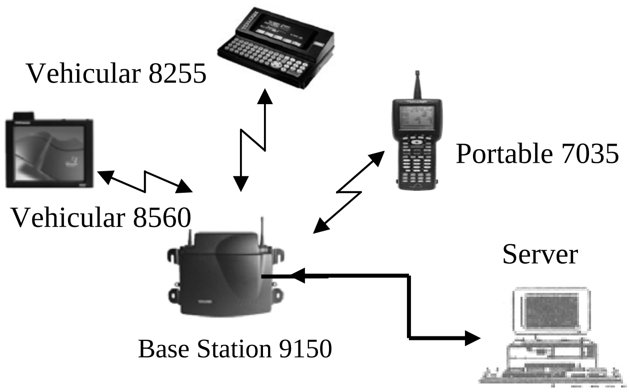


Figure1. Physical ComponentsInvolved in The Scheduling System

The complete wireless scheduling system consists of software and hardware systems[4,5]. The software system involves a database system deployed at the network center and the container information management system based on the database system, as well as some corresponding middleware components. Additionally, corresponding terminal application programs are required at the wireless terminals. The hardware system mainly comprises two sections, the wireless network and a series of wireless terminals. The wireless terminal refers to the conventional vehicular or portable terminals. Figure. 2 shows how hardware components of vehicular terminals and portable terminals compose the scheduling system.

2.1 Mobile Computing Terminals

Mobile computing terminals include all sorts of vehicular terminals installed on yard cranes, quayside container cranes and yard trailers as well as portable terminals used by tally clerks. The developed system adopts Teklogix series terminals, specifically, model 8580 used in cranes, model 8255 used in container trailers and model 7035 used by tally clerks. Multiple options are possible both for vehicular and portable terminals. For example, the input devices can be key-based or touch screen-based. The operating system can be DOS, WIN CE or Windows.

2.2. Wireless Network Communication Platform

The wireless local area network products conforming to standard 802.11b are utilized. Vehicular terminals communicate with wireless base station while moving. Teklogix 9150 series base station is adopted to achieve seamless connection with terminals.

2.3 Data Interchange Middleware

The data interchange middleware components are deployed within the communication server. Conventional network interface card is installed which is able to realize data interchange protocol and data format conversion between the container information management system and the vehicular terminals. In this way, seamless integration is formed between the wireless scheduling system and the container information management system.

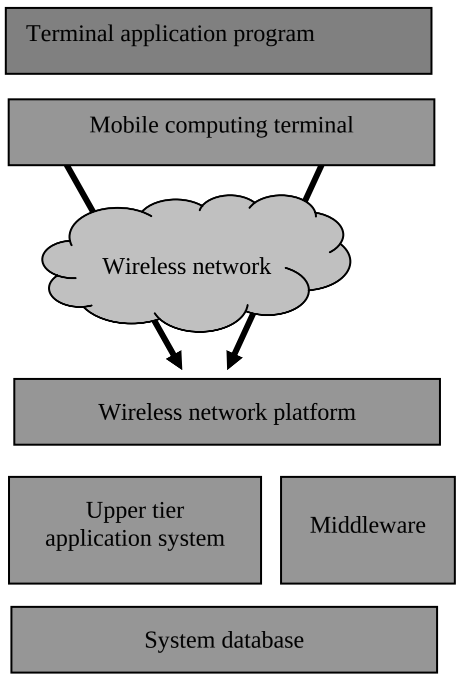


Figure 2. Configuration Architecture of The Scheduling System

2.4 Terminal Application Programs

Terminal simulation mode is adopted in this scheduling system. The terminals are devices able to communicate with the host computer. The working procedures are:

a. The data flow sent to the terminals by the host computer consists of a sequence of control orders. For each control order, the terminal parses it to a sequence of actions, for example, rendering the letters on the screen or locating the cursor on the right position.

b. What is typed on the keyboard is sent to the host computer with a special data sequence. These control sequences can be depicted by an automaton with limited status.

3. KEY TECHNOLOGIES

When a wireless local area network is constructed, the terminals should be logically banded with the master computer while spatially detached from there, the data collected from all wireless terminals should be effectively managed, and furthermore the source data should be encoded/decoded and modulated/demodulated to guarantee high communication quality. Key technologies involved include terminal simulation technology, SQL Server database technology, channel coding and error control technology and modulation technology, etc.

3.1 Terminal Simulation Technology

Terminal simulation refers to the practice using a full-fledged computer to simulate a computer terminal to allow it communicate with the host computer interactively. The practice occurs in two kinds of environments. In the first one, an asynchronous communication interface is used as the communication protocol with terminals. In the second one, communication is made possible through the computer network gateway (maybe also through asynchronous communication interfaces), or the remote terminal protocols running on the network, e.g. the DECnet's CTERM protocol and LAT protocol.

The terminal simulation mode allows users logging into the back-end computer system from the wireless terminal devices as if they are directly operating the back-end computer that is in front of them. It uses the Telnet protocol to provide bidirectional character-oriented communication connection. Client programs are running on the terminals while the Telnet server programs are running on the back-end host computer. The terminals accept user inputs, all of which are sent to the back-end host computer to parse and handle. The results after handling return in turn to wireless devices to render.

Fig. 3 shows the software framework under the terminal simulation mode.

The terminal simulation mode only requires developing and customizing application system on the host computer. For example, application developer can customize screen size to fit certain wireless devices of different specifications to make convenient for users to operate on the small screens of the wireless devices.

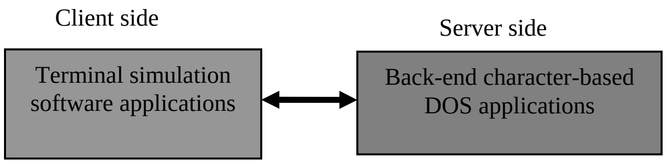


Figure 3. The Software Framework Following The Terminal Simulation Mode

3.2 SQL Server Database Technology

Microsoft SQL server 2000 is a high-performance relational database management system. It is dedicatedly designed for handling huge bulk of data and managing data storage. SQL Server provides for the users several interfaces supporting developing environments on the client side, such as ODBC, Remote Data Object (RDO), OLE DB and ActiveX Data Object (ADO).

ODBC is an Application Programming Interface (API) which complies with industry standards and is used for moving data in or out of a database. As a popular database API, it has been approved as formal interfaces by ANSI and ISO. SQL server provides high performance ODBC interfaces for all Windows programming environments.

3.3 Channel Coding and Error Control Technology

In port mobile wireless communications, signal transmissions tend to be disturbed by the noises emitted from the base station nearby as well as the noisy sources with identical frequency but belonging to other transmission routes. As such, in the digital and data transmission course, irregular coding errors might occur, which may further cause other random errors. In order to ensure communication quality, the signals should be channel-encoded while error codes should be controlled and remedied instantly.

Usually, error detection and message resending mode or information feedback mode is used in the international digital communication systems. However, for communications through complex shortwave channels or dispersive channels, hybrid error detection mode is always used. Since in container terminal wireless communications, shortwave channel propagation and dispersive channel propagation coexists simultaneously, it is more rational to adopt hybrid error detection mode to control any possible errors.

3.4 Spread SpectrumTechnology

The most significant advantage to adopt the spread spectrum technology is its strong counter-interference capability. There are two types of spread spectrum systems, i.e., Direct Sequence Spread Spectrum (DSSS) and Frequency Hopping Spread Spectrum (FHSS). Different spread spectrum systems have different physical anti-jamming mechanisms and therefore, react differently in industrial settings. The choice between radio techniques is dependent on application and environment[10].

In general, FHSS scheme is less likely to collide with other transmissions since changing of transmitting frequency makes colliding in certain time frame more difficult than in DSSS scheme. DSSS, for its part, can remove the interference completely, if the interfering signal power is within the jamming margin, especially in case of low two medium narrowband interference. To summarize, DSSS is ideal for systems that require high speed and are applied to environments with low or medium narrowband interference, whereas FHSS is suitable for applications that require low cost, lower data rate [10].

The container terminal operation environments are extremely harsh. Interferences may not only come from the internal of the system and the interactions between external systems, but also from thunders, windstorms, hard lighting and other noises. On the other hand, the information interchanged between the operational sites and the control center is exceptionally important and cannot be rejected because of abrupt state transitions or to avoid interferences. Therefore, it might be more suitable to select DSSS technique as the terminal wireless signal transmission scheme.

4. ADVANTAGES OF THE SYSTEM

A range of advantages can be observed in the developed system compared with others, especially those without wireless terminals.

- Tight integration is applied with the container operation management system to allow real-time data exchange. The operation efficiency of the yard cranes, quayside container cranes and yard container trailers is thus significantly enhanced while incorrect manipulations are greatly reduced.
- The cranes and the yard container trailers can exchange information with the scheduling center while moving without spatial and environmental restrictions. Truly mobile computing is realized.
- Vehicular terminal products with strong functionality are adopted on the client side. This not only meets the strict requirements from the complicated application programs, but also meets the requirements due to the awful working environments.
- Following the technical advancement direction, the advanced IEEE802.11b-standard- consistent wireless network products are used which not only provide strong mobile communication capability, but also ensure revenues for the customers' investments.
- The design involves standard network environments and transparent protocols which support platforms on the mainstream. This not only satisfies the host/terminals communication mode, but also enables construction of Client/Server mode.

5. APPLICATION Of THE SYSTEM

This scheduling system adopts asynchronous communication interfaces and terminal communication protocols. The client programs running on the client side is WinTekTerm.exe, similar to the web page browsers, to create connections with the servers. The server programs running on the host server is an application developed using VB6.0 to satisfy scheduling communication requirements. The working process flow chart is shown in Figure. 4.

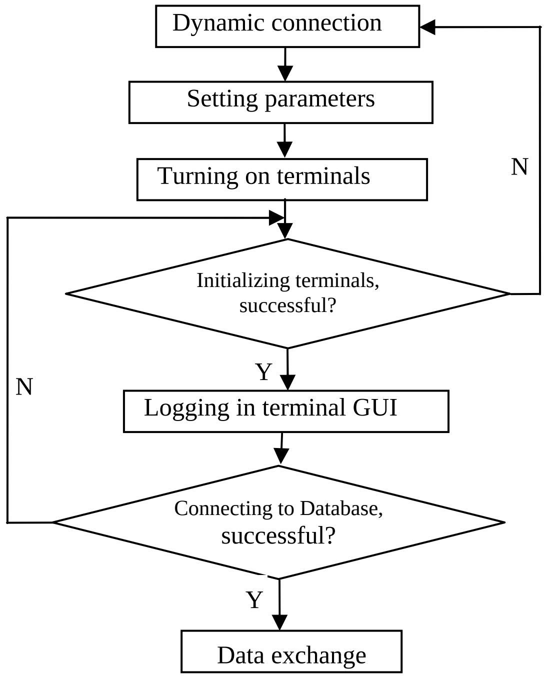


Figure 4. Working Process Flow Chart

Before the system is ready for handling requests, a TCP protocol address object should be created at first. The corresponding procedures include:

(1) Create SQL connections with the database:
Dim db As New ADODB.Connection
Dim rs As New ADODB.Recordset
db.ConnectionString = "DRIVER=SQL Server; SERVER=yanzheng; DATABASE=Port Management Information System; Persist Security Info=False; UID=sa; PWD=yilijun"

(2) Cerate record set
db.Open
rs.Open "select ∗from user ", db, adOpenStatic, adLockReadOnly
rs.Open "insert into user values ('" & str1 & "','" & str2 & "')", db, adOpenStatic,

(3) Update database
adLockReadOnly ／Insert data into the record set
rs.Open "update user set passwd='01'where name='" & user01 & "'", db1, adOpenStatic, adLockReadOnly
／Update data encapsulated in the record set

(4) Activate terminal operation

TsxGetLastFunctionKey(xd(0), fkey&) ／Enable terminal function keys
TsxSelectPage(xd(0), "logo") ／Log in through the GUI
TsxGetFieldData(xd(0), "username.e", size, str) ／ Capture terminal inputs
TsxUpdateFieldData (xd(0), "namedisp.f", TwStrLen (prodname), prodname) ／Update terminal appearance

6. CONCLUSIONS

This study has shown that there are a wide range of advantages for a wireless terminal-enabled container terminal machine dispatch system. It can provide real-time machine state monitoring, scheduling and information query functions for all the mechanical equipments. Manipulation efficiency can be significantly improved while operation cost can be dramatically reduced. It is believed that wireless terminals will be widely applied at container terminal along with ever-increasing advancement of the information and network technologies.

REFERENCES

[1] Y. Liu, “Information management of tianjin port container terminal (in Chinese),” Port Engineering Technology, 2008, no. 1, pp. 22-23.
[2] Z.R. LI, “2.4G wireless network system of container terminals (in Chinese),” Port & Waterway Engineering, 2005, no. 5, pp. 96-98.
[3] H.L. Xie, “The application of SIEMENS s wireless network on international container terminals (in Chinese),” China Instrumentation, 2008, no. s1, pp. 149-152.
[4] C.B. Zhang, D.H. Liu, S.J. Lin, et al, “Design and implementation of vehicular dispatch system (in Chinese),” Computer and Communication, 2002, no. 3, pp. 17—19.
[5] F. Lu and Y. Zhang. “Method of wireless communication with MSComm control of VisualBasic6.0 (in Chinese),” Electronic Engineering & Product World, 2004, no. Z1, pp. 102-103, 105.
[6] J.L Yao, Z.M. Chen, L.Y. Chen, et al, “Data transmission technology using simulation terminal mode (in Chinese).,” Applications of the Computer Systems, 1995, no. 11, pp. 29-31.
[7] X.Q. Zhao, F. Wang and Y. Tang, “Implementation of PC simulate terminal (in Chinese),” Measurement and Control Technology, 1996, no. 6, pp. 59-61.
[8] J.Y. Kim, “A TCP/IP-based remote control system for yard cranes in a port container terminal,” Robotica, vol. 24, no. 5, pp. 613-620, September 2006.
[9] A. E. Coronado, C. S. Lalwani, E. S. Coronado and S. Cherkaoui, “Wireless vehicular networks to support road haulage and port operations in a multimodal logistics environment,” Proceedings of 2008 IEEE International Conference on Service Operations and Logistics, and Informatics, IEEE/SOLI 2008, vol. 1, pp. 550-555.
[10] K.S. Low, W.N.N Win and J.E. Meng, “Wireless sensor networks for industrial environments,” Proceedings of In: International Conference on Computational Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies, and Internet Commerce, Vienna, Nov. 2005 , Vol. 2, pp. 271-276.

STUDY FOR INTELLIGENT COOPERATIVE CONTROL SYSTEM OF TRAFFIC SAFETY ON FREEWAY

Jun Tang[1], Gongliang Jiang[2], Ke Wu[3]
[1]School of Continuing Education
[2]School of Traffic and Transportation
[3]School of Foreign Language
Chongqing Jiaotong University, Chongqing, China 400074
Email: [1]JGL6388@sina.com, [2]jianggongliang@hotmail.com

ABSTRACT

The paper analyzed the influence factors of freeway safety driving speed and indicated the fatalness of speeding on freeway. In order to prevent and avoid traffic accidences, the paper investigated the architecture of freeway transportation safety intelligent cooperative control system based on safety driving speed. The main function of this system is introduced and the integrated model has been established for cooperative warning and control of traffic safety. At the same time, the key techniques, such as real time safety information collection, interactive communication, and safety coordinating control method, were studied. The mode of cooperative safety control is put forward by using ITS and GIS/GPS and information system of safety cooperative control. The integrated architecture of safety cooperative control is established, which realized real time inspection and monitoring of traffic speed.

Keywords: ITS; GIS; GPS; Cooperative Safety Control.

1. INTRODUCTION

Transportation is basic industry of country economy and society development. Along with the fast development of our country's economy, the freeway construction also grew swiftly. Until the end of 2005, the total mileage of our country's freeway has reached 41005 kilometers. At the same time, the freeway traffic safety becomes an important problem day by day. The mortality rate of traffic accidents demonstrates an obvious ascending trend from 1995 to 2005 [1].

At present the number of freeway traffic accident death per year and the mortality rate per ten thousand vehicle of our country occupy the first in the world. The significant accident cannot be under effective control. Therefore, the study for freeway traffic safety problem is an urgent need in transportation industry.

The research in recent years indicates that more than 90% of freeway traffic accidents in our country are caused by speeding, drunk-driving and driving fatigue, in which speeding is the primary reason. Under specific transportation environment, freeway traffic accident happens due to the maladjustment of human, vehicle, road and weather. The impact factors of freeway safety driving speed include vehicle structure, performance and service condition; path linearity, road surface quality, transportation organization and control mode; transportation environment and badly weather such as wind, rain, snow, fog and so on.

Therefore, according to reasons and characteristics of freeway traffic safety accident, the paper investigated the architecture of safety driving speed based transportation safety cooperative control system. The system is established on the basis of Intelligent Transportation System (ITS) and safety driving speed is main control goal to avoid and prevent accidents.

2. THE INFLUENCE FACTORS OF SAFETY DRIVING SPEED ON FREEWAY

The highway safety driving speed refers to the appropriate moving speed under the usual traffic condition. Usually, it is determined by design speed according to the automobile characteristic which goes on the road and the road geometric design request. Actually, safety driving speed on freeway is related to people, vehicle, road, circumstance, and so on. The main factors should be as follows:

People: including driving skill, experience and mentality diathesis of driver.

Vehicle: including vehicle type, drive capability of vehicle, brake capability, etc.

Road: including road surface condition of highway, line type, width, and so on.

Circumstance: including weather condition such as rain, snow, and fog, vehicle flux, traffic accident.

The speed limit of current freeway safety driving speed is generally fixed, and cannot change according to the characteristics of different vehicles, environment situation of road, combination of road line type as well as climatic conditions. Therefore, if the weather situation and the traffic condition change, the speed limit of existing freeway safety driving speed is unreasonable. It is difficult to ensure driving safety of vehicle. Furthermore, at present our country lacks effective monitoring and active control method to the highway vehicles driving speed.

3. THE FATALNESS OF SPEEDING ON FREEWAY

Vehicle driving speed on freeway has a direct influence on possibility and severity of accidents happened. Since there is no intersection, great vehicle flux and high driving speed will have a very high accident mortality and causes significant economy loss when traffic accident happens in case at speeding. The fatalness of speeding on freeway includes [2]:

(1) The possibility of mechanical failure happening will increase greatly to vehicles at speeding. That will directly influence operating stability and cause mechanical accidents such as tire bursting and brake failure.

(2) Sometimes driver is unaware at urgency case due to speeding, so the accident such as collision and vehicle turning over will happen easily. In this case, severe accident will happen at many times because of great collision.

(3) Because of speeding, driver's eyesight falls, eyeshot becomes narrow, and response capability becomes slow. In

case of emergency, he does not have enough time to take action and the possibility and severity of accidents increases greatly.

(4) The driver's spirit is strained at speeding, so the energy in psychology and physiology will be consumed greatly, which results in fatigue.

(5) Vehicle braking distance increases under speeding. Generally, in case the speed increases one time, the distance for braking increases four times. Especially under over-loading case and wet road, the distance for braking will be longer. Thus if vehicle ahead decelerates suddenly, the accident of rear-end collision must be occurred.

(6) When vehicle is running on crooked road, the higher the speed is, the larger the centrifugal force. Thereby operation difficulty will increase and it is very easy for vehicle to enter other lanes and results in accidents.

Thus, there is important practical significance that makes vehicle safety driving speed to be under control so as to enhance safety as well as preventing and avoiding traffic accidents.

4. THE FUNCTION FOR THE INTELLIGENT COOPERATIVE CONTROL SYSTEM OF TRAFFIC SAFETY ON FREEWAY

The intelligent cooperative control system of traffic safety on freeway based on safety driving speed needs to consider factors such as road environment, vehicles type, and weather and so on, and determines the safety driving speed of the different vehicles on different road section in different condition. The driving speed of the vehicle is detected and controlled real-time. It is insured that the vehicle drives at the safety speed.

The intelligent cooperative control system of traffic safety on freeway based on safety driving speed is composed of the safety driving speed model based on vehicle-road coupling and dynamic information system, the real time collection system of vehicle and vehicle driving speed, as well as the intelligent control system for vehicle safety driving.

(1) Safety driving speed model based on vehicle-road coupling and dynamic information system

Through different vehicle models and vehicle type database, road dynamic information system based on GIS is built, information between road traffic safety dynamic information based on GIS(including road condition, weather condition ,traffic environment) and vehicle status(including vehicle type, safety status, driven kilometer, engine characteristic) is integrated and analyzed. Different conditions such as road environment, vehicles type, weather and so on are considered, and the speed value for safety driving of different vehicles on different road section in different conditions is confirmed by simulation in computer.

(2) Real time collection system of vehicle and vehicle driving speed

The vehicle and vehicle real time driving speed is collected exactly by GPS, speed inspection instrument at vehicle, speed monitoring device on road.

(3) Intelligent control system for vehicle safety driving

By comparing the vehicle real time driving speed with the safety driving speed limit, the vehicle safety status is judged automatically. Then vehicle driving speed is controlled by intelligent module at vehicle.

Figure 1 shows the function of the cooperative control system based on safety driving speed.

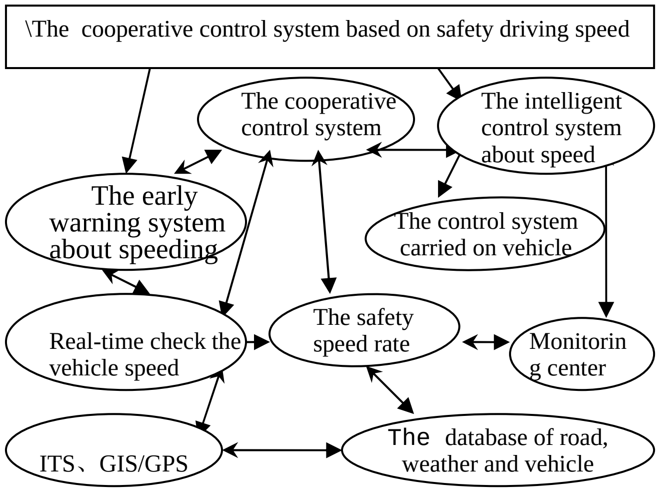


Figure 1. The function of the cooperative control system based on safety driving speed

5. THE ARCTECTURE OF THE COOPERATIVE CONTROL SYSTEM BASED ON SAFETY DRIVING SPEED

5.1 The Architecture of the Cooperative Control System Based on Safety Driving Speed

The architecture of the cooperative control system based on safety driving speed takes ITS and GIS/GPS as well as INTERNET as support.

ITS is established through effective fusion of advanced information technology, data communication technology, automatic control technology as well as the information processing technology, and is applied in the entire traffic control system. It has the characteristic of information, integration and intellectualization. It can display function in wide range, and is a real-time, accurate, highly effective transportation synthesis intelligence control and management system. Therefore, ITS system may provide the cooperative control system based on safety driving speed for technology support.

Using GIS technology, through configuring data network layout, establishing GIS space database of road environment and map essential factor attribute database, and integrating with road traffic environment real-time information system including the vehicles type database and real-time weather information system, it can dynamically input or output road traffic environment real-time information.

Through the digitized vehicles tracing and locating system including GPS, GSM and GIS technology and using computer network to track and real-time collecting the driving speed, it can confirm vehicle type and position as well as collecting vehicle driving speed real time, and achieve communications function with moving vehicle.

With the integration of GIS, GPS and GSM, the architecture of cooperative control system based on safety driving speed is constructed. Figure 2 shows the architecture.

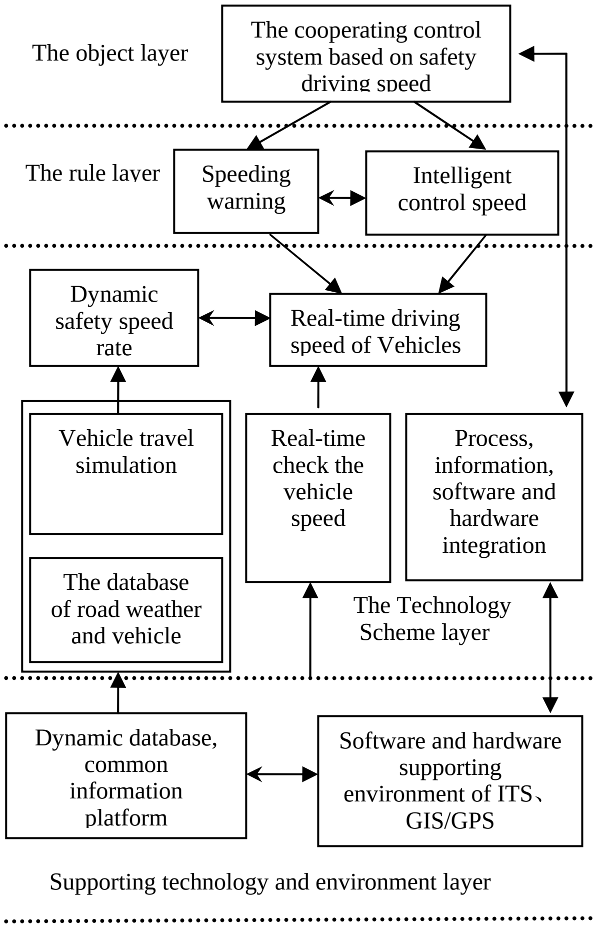


Figure 2. The architecture of the cooperative control system based on safety driving speed

5.2 The Cooperative Control Mechanism Based on GIS Sharing Information Platform

The cooperative control system of traffic safety on freeway takes the transportation geographic information system, the vehicles localization technology, the wireless communication as the main technical supports. GIS is the basic information system platform and it has the superiority including visualization, geography analysis and spatial analysis, database management and so on. The vehicles localization technology can gain real-time positional information of goal vehicles according to the application need. The wireless communication can achieve data transmission in wide range, and also achieve system direction, dispatch, monitoring, management, etc. The cooperative control system of traffic safety is built through integration of GIS based public information platform. The goal such as the real-time and dynamic transportation information, the monitoring, the analysis and the intelligent control can be effectively achieved.

5.2.1 The Information Sharing Platform Based on GIS

The cooperative control system based on safety driving speed mainly includes vehicle real-time driving speed gathering system, vehicles safety speed dynamic information issue system, overspeed early warning and intelligent control system and so on. There exist massive information exchange between subsystem, and information need of each subsystem is complex and diverse. However, some information can be shared. Through the use of information sharing platform, the cooperative control system can uniformly store, organize, and process the entire information, thus effectively guarantees the accuracy and intelligibility of data relation and avoids data redundancy, which enhances the utilization rate and transmission speed of system information.

The information is important fundamental element of the cooperative control system, and also link of each subsystem. The information of the cooperative control system based on safety driving speed includes static information and dynamic information. The static information includes road linetype, typical vehicles information, transportation attached facility information and so on, which usually changes slowly with the time. The dynamic information mainly refers to the information which is real-time gathered, such as real-time vehicle speed, real-time road environment condition, meteorological condition, vehicle location information and so on.

Because geographic information system (GIS) is taken as a technical system which synthetically processes and analyzes the spatial data, it can effectively carry on gathering, storing, searching, modelling, analyzing and output of the spatial data of road environment. Its unique merit lies in that it can organically unify geographical position and correlation attribute information. Such platform cannot only show the traffic information in the space, but also offer support for deep information mining and following information service as well as assistance of decision-making. Therefore, aiming at the information requirement characteristic of control system, it establishes the exclusive dynamic geography information database. By building the GIS platform through the network interconnection and the distribution database system, it realizes coordination and application in overall system.

5.2.2 The Technical Framework of Integration System

According to the requirement of GIS as the sharing platform of the cooperative control system, the system has three-level architecture:

(1) Client level. It is the user who uses the information platform, including the vehicles driver, the freeway safety manager, the public safety responsible department and so on.

(2) Application service level. It takes GIS as the information platform of the traffic safety cooperative control system, gathering data through each subsystem, and returning these primary data with prespecified form, then dealing with these datas such as classification, extraction, mining and fusion and so on. During data storage, the different information according to the standard agreement issued gives the corresponding application subsystem.

(3) Data management level. The level memories the foundation data, provides information connection between the platform and various subsystems.

5.2.3 The Corresponding Function of the GIS Information Sharing Platform

The GIS information sharing platform is the hinge of whole traffic safety cooperative control system,. It has the responsibility of information collectiong, fusion and transmission.

(1) Information collection function. Various subsystems read sharing data according to the prespecified form, and realize the reorganization of the static information and the dynamic information. In addition, it also guarantees the data accuracy and readability, and avoid data redundancy.

(2) Information fusion function. According to the function requirement and the inner link between subsystems, the system handles gathered data such as classification, statistic, connection under certain criterion, mining deep information, and integrates information in order to be used for the

cooperative control system.

(3) Information issue function. According to the requirements of each subsystem, the system transfers the needed information to the subsystem with the specified form; According to the requirement of service and query right, it transfers information such as data, graph, picture, to the client.

6. CONCLUSIONS

Through the analysis of the characteristics of freeway traffic safety accident and in order to prevent and avoid traffic accidence, this paper investigated the architecture of the intelligent cooperative control system of traffic safety on freeway based on safety driving speed. The main function of this system is described and the integrated model has been established for cooperative warning and control of traffic safety. The architecture of safety cooperative control system is built based on ITS and GIS/GPS. It realized real time monitoring and inspection of those factors that can affect vehicle safety driving speed.

REFERENCES

[1] Cai Guo. Analysis of traffic accident trend in our country freeway [J]. China safety science journal. 11(4):pp.11-17,2006.

[2] Ma Zhaoyou,Wang Jiangqiang A Study on the Prevention Countermeasures of Traffic Accidents Caused Speeding .2006 BeiJing International Intelligent Traffic Forum pp.211-215,2006.

[3] Huo Liang. Study for integration and application of 3G technology and modern physical distribution management technology [D]. Wuhan University doctor's degree paper, 2002.

[4] Ji Shanbiao. Study and apply for the GPS dynamic localization[D]. Chinese Academy of Science Shanghai astronomical observatory doctor's degree paper .1999.

[5] [Liu qing, Wu Yanzi. The establishment of freeway traffic safety early warning management system [J]. Wuhan University of Science and Technology journal (transportation science and project edition), 27(3):pp.421-424,2003

[6] Qin Liyan, Shao Chunfu, Xiao Kai.The study and design of road traffic security management system based on the GIS[J].Transportation systems engineering and information, 5(2):pp.104-107.2005.

[7] Tang Haibo,Jiang Gongliang,Hu Xinghua.The Study of the Freeway Vehicle Rear-end Collision Warning System Based on GIS/GPS[J].Traffic and Transportation.pp.83-85,2007.

[8] Wu Mingxian. The study of freeway tunnel integration intellectualization monitoring technology and the system exploitation [D]. Chang An University doctor's degree paper, 2004.

SHORT-TERM TRAFFIC FLOW FORECASTING MODELS AND COMBINATORIAL OPTIMIZATION ALGORITHM

Chen Song [1], Zhang Min [2], Peng Dewei [2]
[1]School of Energy and Power Engineering
[2] School of Computer Science and Technology,
Wuhan University of Technology, Wuhan 430063, China
Email：nemo863@126.com

ABSTRACT

Accurate and real-time short-term traffic flow prediction is the foundation for urban traffic control, and also an important part of intelligent transportation system. We made research and improvement for several major forecasting models based on the analysis of the traditional methods of traffic flow forecasting.

According to the idea of data fusion, the comprehensive results of the various models, and the dynamic weight distribution, we establish a prediction model combinatorial optimization.

The idea of data fusion in accordance with the comprehensive results of the various models, the dynamic weight distribution, the establishment of a prediction model combinatorial optimization. Finally, Validation and comparison have been taken through concrete examples. The result shows that the combinatorial optimization algorithm has the best forecasting for it has the smallest average relative error and mean square error.

Keywords: Short-term, Traffic Flow, Equal-dimension and New-info Model;Dynamic Grey model, BP Neutral Network, Combinatorial Optimization Algorithm

1. INTRODUCTION

Transport system is the infrastructure supporting socioeconomic development, and plays an important role in the socioeconomic system. With the continuous improvement of social development and the extent of urbanization, the traffic problem has become a "bottleneck" problem[1]. Restricting the stable development of economy and society to some extent.

Therefore, people make higher demands of the management and control on road traffic and transportation system. Many countries have invested a large amount of manpower and material resources to carry out road transport system management and control technology development and the establishment of Intelligent Transportation Systems ITS (Intelligent Transport System).The traffic control and guidance system is the core topic of ITS ,while the key problem of achieving traffic flow guidance system is how to carry out real-time, accurate traffic flow forecasts, that is how to predict the traffic condition effectively by using real-time traffic

data.[2]The prediction approaches of short-term traffic flow are mainly traditional regression analysis, time series forecasting, as well as the recent neural network based on artificial intelligence. However, the transport system is a complex system affected by various factors, not a single model can do predictions precisely. This paper does some research on the three major short-term traffic flow model, analyzing and discussing the characteristics of the models, and finally establish a multi-model-based combinatorial optimization prediction algorithm.

2. SHORT-TERM TRAFFIC FLOW PREDICTION MODELS

2.1Equal-dimension and New-info Model Based on Multiple Regressions

Regression analysis prediction method proceeds from causal relationship. Based on observational data, it can figure out the correlation coefficient between dependent variables and independent variables by mathematical statistics methods. The greater the correlation coefficient is, the closer the relationship between cause and effect is. The regression equation can be determined through correlation coefficient, thereby to forecast the trend of future development. However, this method requires a large amount of data, so it cannot be achieved when data is limited [1].Therefore there is a need to improve on the original model to improve the utilization rate of data, to enhance the range of application of the model. At the same time, in order to use multiple regression methods to predict continuously, the input data should be of the dimension. Therefore, we establish an Equal-dimension and New-info Model based on the multiple regression to predict changes in traffic flow.

Firstly, we build multiple regression model,

$$\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, Cov(\varepsilon,\varepsilon) = \sigma^2 I_n \end{cases}$$

Among which,

$$Y = \begin{bmatrix} y_1 \\ \cdots \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

Using experimental data (sample values) for points on the unknown parameter β *and* σ^2 estimation and hypothesis testing, in order to establish the numerical relationship between y and $x_1, x_2, \cdots, x_k$; make the prediction and control for the value of y at the place of

$x_1 = x_{01}, x_2 = x_{02}, ...x_k = x_{0k}$

That is to make interval estimation of y ,then to make estimation for β *and* σ^2 , and figure out the estimator for $\beta_0, ..., \beta_k$ by least square method. Then figure out the regression plane equation.

At last, put the latest observation data $\tilde{x}$ into the original data, and remove the furthest old data x_1 at the same time, composing Equal-dimension and New-info Data Column $\tilde{X} = (x_2, x_3, \cdots, x_k, \tilde{x})$.

Then put them into the model for the next forecasting.
Its schematic diagram is as follows:

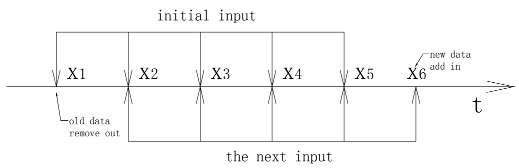


Figure 1. Equal-dimension and New-info Model Based on Multiple Regressions

2.2 Dynamic Gray Model

The system of road traffic law is a typical gray system for it has no clear regularity ,and the status, the relationships between boundaries of the system is difficult to describe accurately [3].so we can use Grey System Theory to establish GM(1,1)model based on the poor information and small sample to forecast the traffic flow. Its advantage is the algorithm is simple, the computing speed is fast, and it can give a better prediction for short-term forecasts.

However, in the GM (1,1) model, the meaningful data for forecasting are only one or two data after $x^{(0)}(n)$. Other further data beyond the data is planning data rather than forecasting data, that is data in future development under the totally same condition of current. As the time goes by, the old data has become increasingly unsuited to the new situation. That is to say, the information meaning of the old data declines as the time goes by[4].Therefore, we established a dynamic GM (1,1) model, which is able to accept information from the feedback system so as to update the data continuously to improve the prediction accuracy.

The steps are as follows:

1) Data Preprocessing:
Firstly,

Accumulate $x^{(0)}$ once to create (1-AGO) to get $x^{(1)}$.
Then, figure out contiguous average and get

$$z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1)$$

2) Establish GM（1，1）Differential Equations:

$$x^{(0)}(k) + az^{(1)}(k) = b$$,

3**) Parameter Identification:**

Do least squares estimation for parameter $\tilde{a} = [a,b]^T$

$$\tilde{a} = (B^T B)^{-1} B^T Y$$

Among which,

$$Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(2) \\ \vdots \\ x^{(0)}(2) \end{bmatrix}, \quad B= \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(2) & 1 \\ \vdots & 1 \\ -z^{(1)}(2) & 1 \end{bmatrix}$$

4) Time responding function:

Solve the shadow equation of GM(1,1)， $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$

We can get the time responding function:

$$x^{(1)}(t) = (x^{(1)}(1) - \frac{b}{a})e^{-at} + \frac{b}{a}$$

5) Forecasting function:
The forecasting function gotten by s-step extrapolation of GM(1,1)model is:

$$x^{(1)}(k+s) = (x^{(0)}(1) - \frac{b}{a})e^{-ak} + \frac{b}{a}, k = 1, 2, \cdots, n$$

6) Inverse Accumulation recovery and figure out $\hat{x}^{(0)}$

$$\hat{x}^{(0)}(k+s) = \hat{x}^{(1)}(k+s) - \hat{x}^{(1)}(k+s-1)$$

At last, update data to compose new time series, and put them into the dynamic gray model to re-forecast.
The flow chart as follows:

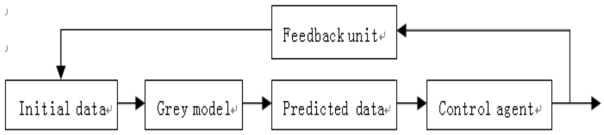


Figure 2 Dynamic Gray Model

2.3 Neural Network Prediction

Neural Networks is a mathematical model based on the results of modern neuroscience research, and it is a simple abstraction and simulation of brain function. Neural network has a strong self-organizing and self-adaptive ability to overcome the shortcomings of non-linear and uncertainties that general linear regression model can't reflect changes in traffic flow. Neural networks are generally adjustable, or can be trained, so that requiring output can be obtained by a particular input. As it is shown below. Here, the network adjusts itself according to the comparison of output and target until the network output and the target match.

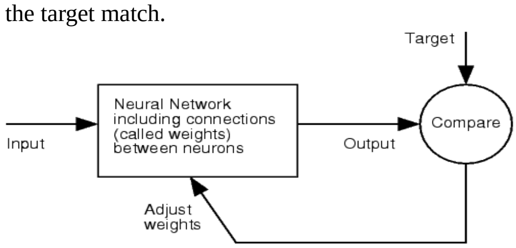


Figure 3. Neural Network Schematic

In this paper, we use MATLAB to establish back-propagation network (BP network) to do prediction. The process of neural network calculation is as follows:

1) Data Preprocessing:
Data preprocessing can improve operating efficiency of procedures, so as to accelerate the convergence of training network.

First of all, using mapminmax function to normalize data, then call function dividevec to divide data into training, validation and testing three parts in order to meet the requirements of network training.

2) Network creating
Calling the function newff to generate a BP network and initialization. Function newff will call the initialization function init automatically after determining the network structure, using default parameter to initialize the network weights and all thresholds, and result in a feedforward training network.

3) Network training
Calling the function train, and training the network according to default parameters net．tminFcn and nettrainParalT. Using batch mode to modify the weights and threshold value of the network to achieve the targets set by the requirements of network performance.

4) Network Simulation

Using the function sim to simulate the network, and figure out the output under the given input of network.

The main procedure is as follows:

```
[p2,ps]=mapminmax(p);
[t2,ts]=mapminmax(t);
[trainV,val,test] = dividevec(p2,t2,0.20,0.20); net = newff(minmax(p2),[40 1], 'tansig','purelin'},'traingd');
net.trainParam.epochs=1000;
net.trainParam.goal=0.01;
LP.lr=0.1;
[net,tr]=train(net,trainV.P,trainV.T,[],[],val,test);
a2 = sim(net,p2);
x = mapminmax('reverse',a2,ts);
```

2.4 Multiple Model Combinatorial Optimization Algorithm

2.4.1 Data Fusion

Since each model has its own limitations, for the realization of complementary advantages, we make combination forecasting according to the idea of data fusion.

Assumption :

$M_i(t+1)$ is the forecasting result model i at the moment of i ,and $\tilde{M}(t+1)$ is the combinational forecasting result.

w_i is the weight of model i, then:

$$\tilde{M}(t+1)=\sum_{k=1}^{3}\omega_i M_i(t+1)$$

Among the equation, $\omega_1+\omega_2+\omega_3=1$.

That is, making the weighted average is the final predicting result. The weight is decided by the error of the last stage.

2.4.2 Principles of Weight Distribution

The key of combination algorithm is how to distribute the weight of the model. As in different periods, each model has differences of the effect of forecasting. For a certain model, maybe it will coincide with the fact result of the earlier time, while have a great error of the latter time.

Therefore, we need to distribute weight dynamiclly according to the forecasting accuracy of the last moment.the standard measurement of accuracy usually using some predicting accuracy indicators.The less these indicators are, the better the performance of prediction is.the commenly used indicators are MAPE(the mean absolute percentage error)and RMSE (the root squared error). Relative error can reflect the degree of deviation from the two-side. It's effect is better than absolute error when doing comparision of different two-side prediction.Mean absolute percentage error can not only reflect the value of the error, but can also describe the degree of concentration and discretation of error distribution.the larger the mean absolute percentage error is, the discreter the error series is, so is the prediction effect.

$$MAPE=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t-y_t}{y_t}\right|$$

$$RMSE=\sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t-y_t)^2}{T}}$$

Taking average relative error e as measurement standard.Then the larger e is, the smaller the weight should be.So we use the multiplicative inverse to distribute weight for each model.

$$w_i=\frac{1/e_i}{\sum_{k=1}^{3}1/e_i}$$

According to the equation above, the smaller models have the larger weight, and it meet the demand of the Combinatorial forecasting models.

2.4.3 System Diagram

Its system diagram is as follows:

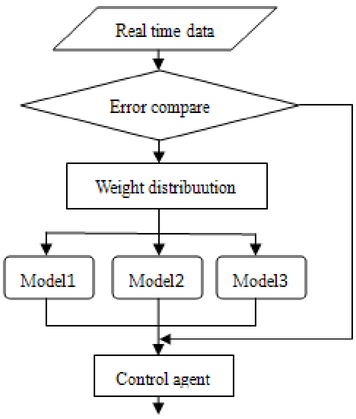


Figure 4. Combinational Algorithm Schematic

3. SAMPLE ANALYSIS

For further specific analysis of short-term traffic flow forecast, we make use of statistical data providing by traffic engineering [5] to verify the model. All of these approaches were used for traffic flow simulation and forecasting. We can compare each model's predictive validity through the error of the true value.

Using those three models and the combinatorial optimization algorithms to predict short-term traffic flow, we can get the results through the MATLAB programming.

The graphs are as follows:

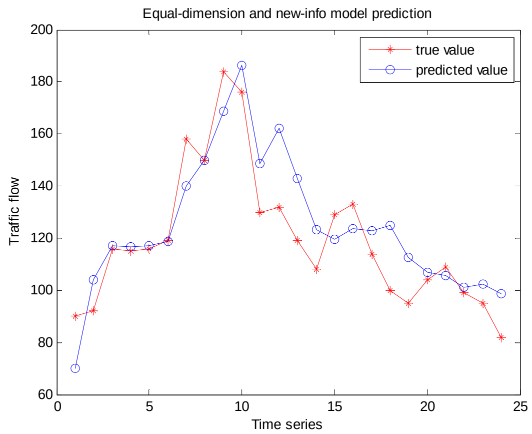


Figure 5. Equal-dimension and New-info Model Prediction

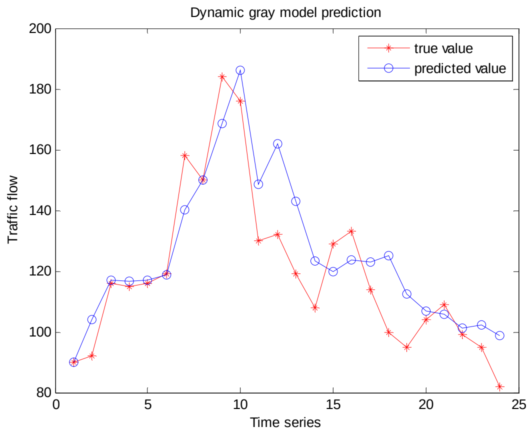


Figure 6. Dynamic Gray Model Predictions

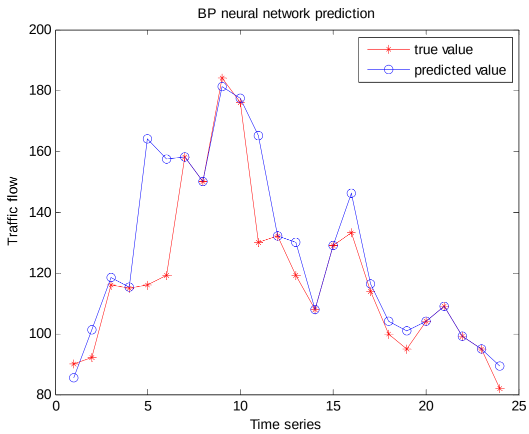


Figure 7. Dynamic Gray Model Predictions

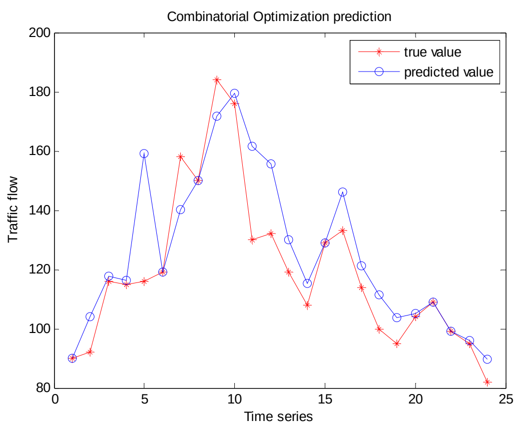


Figure 8. Combinatorial Optimization predictions

Evaluate the prediction effects by error evaluating index:

Table 1. Comparison of prediction error of different models

indicators / model	MAPE	RMSE
1.equal-dimension and new-info model	0.0989	14.1560
2.dynamic gray model	0.1292	20.6090
3.BP neural network	0.0798	19.2821
4.Combinatorial Optimization prediction	0.0750	14.0082

According to table1, the average relative error and mean square error of these models are small, so they can predict the short-term traffic flow relatively accurately.
From the two indicators, the four models prediction effects are:

MAPE : Model 4 > Model 3 > Model 1 > Model 2

RMSE : Model 4 > Model 1 > Model 3 > Model 2

In terms of a single model, BP neural network has the smallest average relative error ， while equal-dimension and new-info model has the smallest Mean Square Error。After combinatorial optimization, the forecasting results have the smallest average relative error and mean square error, that is, the forecasting result of combinatorial optimization is the best. Therefore, integrated different forecasting models, and do data fusion for the results can reduce error and improve accuracy of forecasting.

4. CONCLUSIONS

As a dynamic random process, each prediction model of short-term traffic flow has its own limitations, so it is very difficult to obtain satisfactory results of different conditions.
In this paper, firstly, we analyze and improve three of existing models. Then, we take the weighted fusion algorithm, and use these three models to establish combinatorial optimization prediction model.

Examples of analysis show that this combination algorithm can further improve the prediction accuracy.

Therefore, the short-term traffic flow forecasting the development trend will be to improve the existing model at the same time, the preparation of a new integrated portfolio of algorithms to predict a number of models.

Therefore, the development trend of short-term traffic flow forecasting should be improving the existing models, and drawing up new combinational algorithm at the same time.

REFERENCES

[1] Summary of Short-time Traffic Flow Forecasting Method [J] JOURNAL OFUNIVERSITY OF JINAN (Sc.i & Tech·)，2008，22(1):89-93.
[2] The Optimizing of Many Traffic Flow Forecasting Models [J] COMMUNICATIONS STANDARDIZATION，2007，(4):207-210.
[3] PREDICTION OF TRAFFIC FLOW BASED ON GREY THEORY AND BP NEURAL NETWORKS [D] Harbin：Harbin Institute of Technology，2006.
[4] PREDICTION OF THE SHORT TRAFFIC FLOW AND VISULIAZATION BASED MATLAB [D] Qingdao：Shandong University of Science and Technology，2007.
[5] Ren-Futian Liu-Xiaomin Rong-jian. Traffic Engineering [M] Beijing：China Communication Press,2003.
[6] A study on urban short-term traffic flow forecasting based on a nonlinear time series model [J] CHINA CIVIL ENGINEERING JOURNAL, 2008, 41(1): 105-109

LOCAL REGION RECOGNITION OF VEHICLE PLATE*

Dan Liu, Jiang Chang, Long Zhang
Department of Forensic Science & Technology, China Criminal Police University
Shenyang, Liaoning110035, P.R.China
Email: dliudlmu@gmail.com

ABSTRACT

The Local Region Recognition of Vehicle Image System is one kind of image processing tools. It can be applied as automatic license plate recognition system in the public security. The license plate recognition is done in the several operations: gray stretching, histogram equalization, image brightness enhancement. Gray stretching can deftly control the distribution of output histogram [3].

The most well-known method is the histogram equalization because of its automatic procedure. Image brightness enhancement, modifying the value of pixels, is a classical method for improving the visual perception. The input images such as blurred images or wry images are transferred by image processing algorithms to well-marked image. The processed image is more suitable and efficiency than the original image for some applications. Therefore, a system about digital image processing is carried out. It could acquire the local region information of the vehicle plate image by processing the local region of license plate, which may help the personnel of public security to investigate the criminal case.

Keywords: Local region recognition of vehicle image, Gray stretching, Histogram equalization, Image brightness enhancement, The public security.

1. INTRODUCTION

Follow the development of modern technology, vehicle image recognition is more and more complex, especially, the requirements of local region recognition of vehicle image technique is more and more advanced. The vehicle images are acquired by the camera. Because of the various working conditions, different outdoors scenes and no stationary, there are some factors to influence the quality of the acquired images, such as nature illumination intensity; cleanliness visual degree of vehicle license; vehicles' speeds [15].

The image may appear faintness or deflection so that its information can't be recognized. Therefore we need to process the blurred images by computer. This system is a new tool for automatic vehicle and traffic monitoring using digital image processing. It is using the local digital image processing technique to manipulate the license plates region of the entire image. Then the acquired vehicle license plates information could help policemen handle a case.

2. PRINCIPLES

2.1 Gray Stretching

An input image through pixels linearization or pixels non-linearization may be transferred to a new output image. Frequently, an image is scanned in such a way that the resulting brightness values do not make full use of the available dynamic range. It can be solved by stretching the range of the histogram over the available dynamic range. In other words, gray stretching is a method which changes the distribution of the pixel appearance frequencies over the entire width of the histogram. Thus, it is an operation that modifies the histogram to distribute the intensities on the scale of available values as well as possible [2]. It may increase the contrast of an image. For example, an image that is too dark will become more 'visible'. Gary stretching is subsection linearity function. It is given by,

$$g(x)=\begin{cases} \frac{g_1}{f_1}f(x) & 0\le f(x)<f_1 \\ \frac{g_2-g_1}{f_2-f_1}\left(f(x)-f_1\right)+g_1 & f_1\le f(x)\le f_2 \\ \frac{255-g_2}{255-f_2}\left(f(x)-f_2\right)+g_2 & f_2<f(x)\le 255 \end{cases}$$

Where, $(f_1, g_1),(f_2, g_2)$ are coordinates of the first two points for gray stretching.

The transform function graph is as follows:

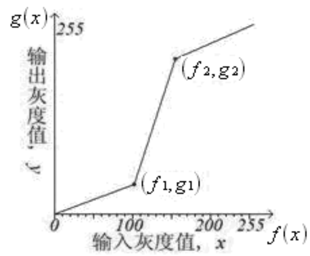


Figure 1. The transform function graph of gray stretching

In this way, if the points of the histogram are very close to each other, the stretching will make it possible to provide a better distribution in order to make light pixels even lighter and the values of dark pixels closer to zero, which is the value of black color.

2.2 Histogram Equalization

The histogram of an image can be composed of many bars. Different bars describe the number of pixels with different gray levels. The purpose of histogram equalization is to harmonize the image luminosity level distribution in such a way that each level of the histogram tend to contain the same number of pixels [3].

Firstly, it is common to normalize the histograms to a "standard" histogram. However, the effect of brightness saturation will be appeared in some quasi-homogeneous region. This causes the merging of adjacent gray levels to flatten the global histogram. Not only will some lower brightness be grouped together but also some higher brightness will uniform the histogram distribution. This operation aims at increasing the image nuances and making a constant for all brightness values.

* This paper is supported by China Criminal University Science Research Project (0209023).

The function is given by, $$D_k = \frac{255}{Width * Height}\sum_{i=0}^{k} N_i$$

Where, N_i is the number of pixels with the i th gray level [3]. Histogram equalization creates a brightness distribution where all values are probable equal. However, for any given image, this situation can only be approximate.

2.3 Image Brightness

Image enhancement is one kind of important image processing methods. The principal objective of this technique is to process the image so that the result is more suitable for specific application [6]. The brightness of the entire digital image is adjusted through increasing or decreasing each pixel value.

The formula is given by, $T(x) = kx + d$. When k =1, increasing or decreasing d will adjust the brightness of digital image [5].

3. SYSTEMIC ACHIEVEMENT

This system takes advantage of computer programming language of Microsoft VC++ to accomplish the partial vehicle license plate image processing and makes the technology of vehicle plate recognition be widely used in the public security. The feature of this system has the simple menu and the convenient operation, the more significant is that it can directly be used without any training. It is especially suitable to the front line of the policemen. When the input image is opened, the main functions of this system is as follows, firstly, it can modify the image's size with dragging the window so that the operator can conveniently identify the details of the image clearly. Secondly, it can mark the picture with the described oval to highlight the location of license plate; the most important functions to process the image of the local region of vehicle license plate are that gray stretching, histogram equalization and the image brightness enhancement or decreasing. These operations can get the distinct vehicle license plate, thus it is propitious to make tracks for the vehicles that commit a crime and accelerates the investigation of cases.

To achieve the local region processing of the image, the system describes the location of vehicle license plate with the oval at first, and then according to the chosen algorithm of gray processing, the number of pixels in the inner of the oval can be acquired. The system may get the drawing area of oval through calculating coordinates values. When the information of the inner oval is acquired, the local region image may be operated by the techniques, such as gray stretching, histogram equalization and the image brightness enhancement or decreasing. Consequently, we can accomplish the local region image processing. Especially, this system marks the vehicle image using the described oval not traditional rectangle. Because the oval not only describes the vehicle license plates but not lose parts information of the car body.

Let we see how the system works now, firstly, run this system program and open the well-preprocessed bitmap image, vehicle license plate 1.bmp. Secondly, make the image more clearly, the operator can drag the window to augment the image and click the paintbrush icon in the toolbar to draw an oval in the location of vehicle license plate with the mouse, finally, make vehicle license plate and part of vehicle come into the area which we depict as the figure 2,

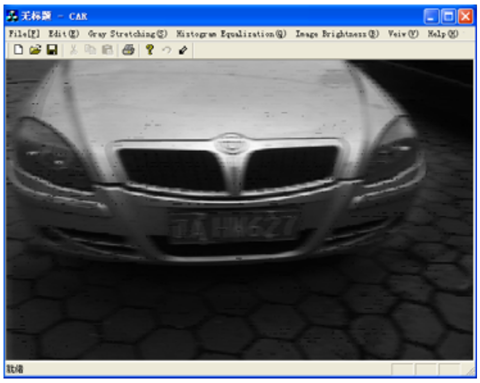


Figure 2. System interface

In the menu, we can click "Gray stretching", "Histogram equalization", or "Image brightness", these three buttons can process the digital image. The original image and the result images are as follows,

Figure 3. Original image

Figure 4. Image after gray stretching

Figure 5. Image after histogram equalization

Figure 6. Image enhancement

We can distinctly see the number of this vehicle license plate is that "辽A HM627" through this system processing.

4. CONCLUSIONS

Through the experiments, it is known that this developed system can do some simple gray level processing to part of the image, including gray stretching, histogram equalization and the image brightness enhancement or decreasing, and the accurate vehicle license plate's information can be acquired. Application of digital image processing could provide the higher efficiency in the public security, and relieve the police men from the arduously work [12]. In the research field of public security, the system is used to process simply vehicle license plate. This system is particular convenient and easy for people to operate, which can identify the vehicle at the first time and find the vehicle rapidly and in time.

REFERENCES

[1] Tiger Studio: Zhang Weihua, Liu Zheng, Zhao Zhigang, Visual C++ Programmer Actual Combat Training, Beijing: Posts & Telecom Press, Oct 2004 (in Chinese)

[2] Wang Zhanquan, Xu Hui, Familiarity Visual C++ Digital Image Processing Technique and Project Cases, Beijing: Posts & Telecom Press, Jan 2009 (in Chinese)

[3] He Bin, Visual C++ Digital Image Processing, Beijing: Posts & Telecom Press, April 2001 (in Chinese)

[4] Wang Guan Technology Studio, Visual C++6.0 Vogue Program Hundred Cases, Beijing: China Machine Press, Jan 2001 (in Chinese)

[5] Zuo Fei, Wan Jinsen, Liu Hang, Visual C++ Digital Image Processing Exploiture Introduction and Program Practice, Publishinghouse of Electronics Industry, Mar 2008 (in Chinese)

[6] A.Broumandia and M.Fathi, "Application of pattern recognition for Farsi license plate recognition", ICGST-GVIP Journal, Vol.5, Issue 2, Jan.2005

[7] Ching-Chung Yang, "Image enhancement by modified contrast-stretching manipulation", Optics & Laser Technology, Vol.38, Issue 3, April 2006

[8] Gu Sheng, "Causes and Solutions of Error Point in Digital Test Photographic", Criminal Technology, Issue 6, 2008 (in Chinese)

[9] Zhang Wenda, "Application of Digital Image Processing Technique in Criminal Identifying", Forest Police, Issue 6, 2008 (in Chinese)

[10] Feng Qingzhi, "Image Enhancement Technique Base upon Histogram Modification", Guangdong Police Technique，Issue 2, 2004 (in Chinese)

[11] Liu Yanting, Xu Hailan, Yang Lei, "The Development of Methods in the LPR Software", Journal of Beijing Broadcasting Institute(Science And Technology), Vol.12, No.4, Sept 2005 (in Chinese)

[12] Yang Yingcang, "Digital photo edit technique in the police field's application and its developing expectation", Journal of Guizhou Education Institute (Natural Science), Vol.18.No.2 Apr 2007 (in Chinese)

[13] Wang Shupeng, Fang Li, "A Technic of Improving Pictures Based upon Visual C++", Computer Knowledge and Technique（Science Communion）Issue 7, 2006 (in Chinese)

[14] Liao Ming, Zhang Jinlin, Zhen Shuxin, Fang Husheng, "Utilitarian Loacting Plates Algorithm Designation", Journal of System Simulation, Vol.17, No.10, Oct 2005 (in Chinese)

[15] Guo Jinquan, Wu Hongtao, "Vehicle Image Recognition Pretreatment Method Research", China Electronic Market Conditions: Communications Market, Issue 4, 2008 (in Chinese)

METHOD OF PASSWORD SECURITY EVALUATION

Miloslav Hub, Jan Capek
Faculty of Economics and Administration, University of Pardubice
Pardubice, Czech Republic
Email: miloslav.hub@upce.cz; capek@upce.cz

ABSTRACT

In this paper a method for the evaluation of password security against a dictionary attack and a brute force attack is suggested. The method is based on a mathematical model and a simulation of these types of attacks, and it can be used for future research in the data security field. The model of dictionary attack and brute force attack is based on some underlining assumptions that are discussed and justified. The suggested model is validated on a study of passwords.

Keywords: Authentication, Password, Dictionary attack, Brute force attack, Security.

1. INTRODUCTION

Data security is an actual issue that is being discussed, especially in the public administration domain. One of mechanisms used in achieving one's security goals is authentication based on some knowledge shared by the system and the user, a password[10]. Nowadays passwords are still commonly used for authentication purposes, although recently they are thought as not being secure as some of other forms of authentication mechanisms[8]. The reason behind this is probably because the implementing of passwords is easy and not so expensive[4]. There are many client/server applications that use passwords for authentication, for remotely logging onto a supplier's web database and/or bank systems. In these applications it is often possible for an attacker to intercept the password and then replay it to the server. This replay problem can be overcome by using a system called Onetime - Password System (OTP) [3], [9]. An OTP system has an advantage over a regular password system in that the former generates a different password for each time authentication is required. In the one-time-password system, the password entered by the user does not traverse the network. This enables the OTP systems to protect users against passive attacks [6], [9].

In this contribution the authors do not use the OTP way, they are more interested in the accuracy of old password systems and can prove that they still function effectively.

2. PROBLEM FORMULATION

Due to the fact that impostors are always looking for new ways on how to get access to protected accounts, many factors that influence password authentication security exist. Examples of these factors are: length of the password, encrypting methods, the mode in how the password is stored, and so on. As it is depicted in Figure 1, it is possible to divide these factors into two basic groups, human factors and technological factors.

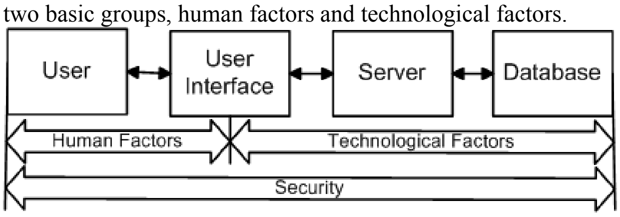


Figure 1. Factors of Password Authentication Security (Source: modified on the base of [8])

Human factors can be divided to two categories:

- Type of password (length, randomness, used characters, etc.)
- Mode the user guards a password (how often a user change his password, whether the user writes a password down, and so on)

Since users are thought to be the weakest link of every security solution, it is necessary to study their behavior. We are convinced of the need to study how users choose their passwords, because it evidently infers of security of this kind of authentication.

A lot of authors frequently discuss about the factors that influence password security, for example: length, randomness, and the period the password is used. Some authors are trying to make a distinction between a "weak" and a "strong" password, commonly by using an expert's opinion[1]. Other authors are trying to break passwords, and the results of their experiments are present as a proof of the passwords weakness[7],[2]. The authors of this paper are convinced about the need for an exact number that represents the security level of some password.

The characteristic of password security will serve for various purposes:

- Decisions on how password authentication will be implemented (password security evaluation as a part of risk analysis).
- Surveys on long-time term trends in password selection.
- Surveys in password selection by different types of users.
- Studies on the effect of different modes of trainings in password selection.

3. METHOD SUGGESTION

3.1 General Principle

Let us determine the security of some password as being *S(p)*, where p is a password selected from the set of all possible passwords P, i.e. $p \in P$. Theoretically, the set P is a non-bounded set (it is always possible to select a longer password), but in practice every system has a maximum capacity. Next, it is possible to give reasons for $S(p) \in \langle 0, \infty)$. The value zero represents the situation where the impostor knows the password and does not need to break the password. The right boundary of this interval results from the fact that it is always possible to construct a more secure password (again we assume the system has an unlimited capacity).

The concrete value of *S(p)* should represent the effort an attacker must invest when attempting to break a password. When the value of this indicator is high, it means that the impostor is experiencing higher expenses while trying to break a password. Intuitively two criterion emerge:

- Time to crack a password
- Cost of cracking a password

The time needed for cracking a password depends on technological factors. This time will differ depending on whether an attack is carried out via the internet or via a local network. The fact that in some systems an account is obstructed for a period of time after an unsuccessful attack, can impact the time needed to crack a password. The cost of password cracking is also closely related to technological factors. As previously discussed, the goal of this paper is to create a method based on human factors with elimination of technological factors. Next, these two factors can change as a result of new progresses in technology.

For these reasons, as criterion of an attacker's effort, the expected value of the number of attempts an attacker has to carry out to break the password, is selected. The advantage of this criterion is non-dependence on technology factors. Time and cost criteria, which were previously discussed, can derive from this genuine criterion if needed. For example, it is not difficult to determine how many attempts you are required to make per hour in order to successfully crack a password, at a network level.

The evaluation of passwords from a security point of view is composed of two phases:

1. Attack simulation model
2. Password security evaluation, on the base of attack simulation model

3.2 Attack Simulation Model

When constructing a model of dictionary attack and a brute force attack we formulate two assumptions:

1. Attackers are choosing the most effective way of attack.
2. Attackers know the types of passwords users are selecting.

The first assumption results from the principle of rational behavior of every individual. The second results from our opinion that an attacker can get information about user habits either from an expert's opinion and/or from a study. It is reasonable to assume that Czech users will most likely select a Czech word as their password than a Spanish word. Next, a lot of research exists that describe which types of passwords users select (e.g. length of typical password, its structure, and so on). Moreover, this paper can also be used as a reference guide for attackers as well.

For simplicity but without loosing accuracy, we can think a brute force attack is like a special kind of a dictionary attack. For example, when an attacker is trying to know whether a password is two characters long, he is actually testing whether a password is a word from the virtual dictionary that consists of all two-character words. The size of this virtual dictionary can be calculated by eq. (1).

$$N_{VD}(NPC, L) = NPC^{L} \tag{1}$$

Where:

N_{VD} .... The size of a virtual dictionary.
NPC....... The number of possible characters.
L .... The length of a password.

Now we can consider a dictionary attack and a brute force attack to be a well-considered sequence of tests performed when trying to know whether a password is a word from a given dictionary. The question is "What dictionary does an attacker use?" on the first attempt, the second, and so on. Based on the assumptions previously discussed, the attacker prefers dictionaries that maximize the probability of his success and minimize the number of attempts to break the password. This criterion can by expressed by eq. (2).

$$SDA(d) = \frac{NBP_d}{N_d \cdot NP} \tag{2}$$

Where:

$SDA(d)$ ...Success rate of the dictionary attack on dictionary d.
NBP_d ......The number of passwords that would be broken by dictionary d.
N_d ....The size of dictionary d.
NP ....Total number of tested passwords used in the attack simulation.

Because we expect the attacker will not test words he has already tested, when sorting dictionaries we recursively remove the used words and reassess unused dictionaries. The overall process is described by the following algorithm.

Step 1: Gather passwords that were used in a given environment by a given kind of users.
Step 2: Gather all possible dictionaries that can contain passwords gathered in step 1. These dictionaries will be used for dictionary attack simulations.
Step 3: Create virtual dictionaries that consists of all one-character strings, two-character strings, and so on, and that can contain passwords gathered in step 1. The sizes of these dictionaries N_{VD} can be calculated by Eq. 1. These dictionaries will be used for brute force attack simulations.
Step 4: Calculate the success rate of the dictionary attack for every dictionary $SDA(d)$, using Eq. 2.
Step 5: If the success rate of the dictionary attack $SDA(d)$ for every dictionary is zero, stop this algorithm, otherwise continue.
Step 6: Select dictionary with maximum attack success rate. This dictionary will be used in the attack simulation model in the order this dictionary was selected.
Step 7: Delete all the words that the selected dictionary contains from the remaining dictionaries. A new set is created for the remaining reduced dictionaries.
Step 8: Repeat step 4 for the set of remaining reduced dictionaries.

3.3 Password Security Evaluation

The result of this algorithm is a sorted set of reduced dictionaries that the attacker can use in the event he wants to break a password in the most effective way. Now, it is easy to calculate the security of a password, which is defined as the expected value of number of attempts the impostor has to carry out to break a password, with help of Eq. 3.

$$S(p_i) = \frac{N_i + 1}{2} + \sum_{j=1}^{i-1} N_j \tag{3}$$

Where:

$S(p_i)$........Security of a password p that is a word from i-th reduced dictionary.
i ....The order of the reduced dictionary that contains a password p.
N_i............The size of the i-th reduced dictionary.

When calculating $S(p_i)$ a desire to express a variability of this variable can arise, due to the fact that dictionaries can have different sizes. We suggest, as the parameter of this variability, the range of security of a password that can be calculated by Eq. 4.

$$RS(p_i) = N_i - 1 \tag{4}$$

Where:

$RS(p_i)$ .....Range of security of password p that is a word from i-th reduced dictionary.
N_i ....The size of the i-th reduced dictionary.

On the base of Eq. 4, one can say that the best model is a model that consists of ordered sequences of one-word reduced dictionaries. Unfortunately this model loses the ability of generalization because it does not take into account the passwords that did not appear in previous surveys.

4. PASSWORDS SURVEY

In our study of real used passwords we gathered two sets of passwords users used on web pages. The first collection was used as training data for model parameters training and the second one was used as test data for model evaluation. The first set of 2,958 passwords was collected during 2005 [5]. All users who were selecting passwords were Czech speaking, and the requirements for password selection are as follows:

- The password has to contain a minimum of one character.
- Maximum length of the password is not restricted.
- Users had no time limit when selecting a password.
- A password could contain arbitrary characters typed using a keyboard.

The second set of passwords was collected three years after the first study (in 2008), and 1,895 passwords were collected. The requirements for password selection were the same as in the first study.

Firstly, Exploratory Data Analysis (EDA) was applied to the first password collection. The goal of this analysis was to create the basic assumptions about users' behavior, and for pertinent dictionaries selection. Diacritic characters were rarely used in passwords, only in 1.8% passwords. Further, only 10.6 % of passwords contained an uppercase character and 23.2 % of passwords contained a minimum of one numeral. After dividing the acquired passwords into four groups, in relation to the "randomness" of the password, it is possible to see that users prefer common words as their passwords, as you can see in Fig. 2. This assumption is proven when you test the correlation coefficients hypothesis between the frequencies of characters in passwords and the frequencies of characters in Czech words (Kendall rank correlation coefficient equals 0.78).

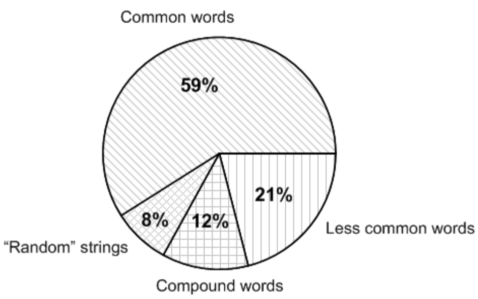


Figure 2. "Randomness" of Passwords (Source: own.)

Not only the structure but also the length of passwords from the first collection was investigated. Figure 3 depicts the frequency of the length of a password.

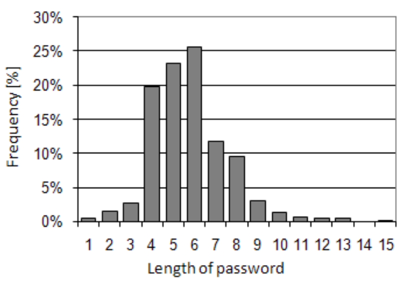


Figure 3. Length of Passwords (Source: own.)

When creating an attack model we selected, with help of the first data set, in accordance with the algorithm described before, 21 dictionaries from 35 original directories we acquired from different resources. The final order of these reduced dictionaries is as follows:

1. Czech First Names (490 words)
2. Common Czech Words (382 words)
3. Common Passwords (239 words)
4. Czech First Names (the first character uppercase) (490 words)
5. Years 1900 – 2029 (114 words)
6. Common Logins (2,131 words)
7. The Most Commonly Used English Words (391 words)
8. Czech and American Word Combinations (496 words)
9. Word Personages (437 words)
10. American Women Names (4,414 words)
11. American Men Names (3,020 words)
12. Slovak Dictionary (17,952 words)
13. Common Word Connection (796 words)
14. Electronic Firms (41,053 words)
15. Foreign First Names (8,801 words)
16. Czech Dictionary (157,228 words)
17. Bible Characters (10,654 words)
18. Unusual First Names (4,612 words)
19. English Dictionary (317,410 words)
20. States and Towns (68,729 words)
21. Big English Dictionary (581,000 words)

The next 15 complementary dictionaries were formed by virtual dictionaries that simulated a brute force attack that followed a simulated dictionary attack. Figure 4 shows the dependence of the False acceptance ratio on the number of tested passwords in the dictionary attack. Generally, the False acceptance ratio is a conditional probability that an impostor will be labeled by an authentication system as a valid user. This dependency in our model is expressed by Eq. 5. The results of the modeled brute force attack, that follow-up to dictionary attack, are not depicted for lack of space.

$$FAR(NA) = \frac{NBP_{NA}}{NP} \qquad (5)$$

Where:

FAR ....False acceptance ratio.
NA ....The number of attempts it took to break a password.
NBP_{NA} ....The number of broken passwords, after some number of attempts to break a password.
NP ....Total number of tested passwords used in the attack simulation.

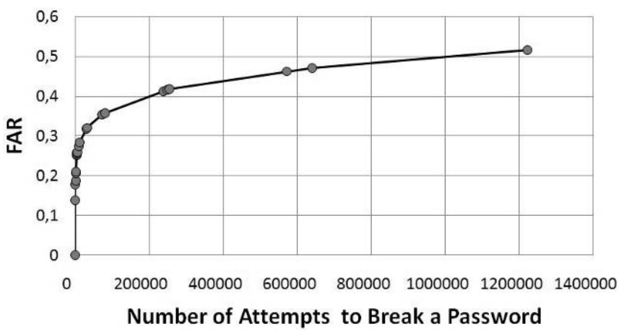


Figure 4. Simulation of a Dictionary Attack (Source: own.)

Table 1 the shows results of our study. The Pearson product-moment correlation coefficient between the expected frequency of passwords, gained from the first data set, and the actual frequency of passwords, gained from the second data set, equals 0.94. The high value of this coefficient indicates the stability of this model. Now we can make the conclusion that 10.85 % of all passwords have $S(p_1) = 246$ (with $RS(p_1) = 489$), and so on.

Table 1. Frequency of Passwords (Source: own.)

Type of an attack	No. of reduced dictionary	Security of a password	Expected frequency of passwords [%]	Actual frequency of passwords [%]	Type of an attack	No. of reduced dictionary	Security of a password	Expected frequency of passwords [%]	Actual frequency of passwords [%]
Dictionary	1	246	13.83	10.85	Dictionary	19	412402	4.60	1.32
	2	682	4.02	3.86		20	605471	0.74	0.12
	3	992	0.98	0.53		21	930336	4.56	4.25
	4	1355	1.89	1.25	Brute force	22	1220854	0.07	0.02
	5	1655	0.41	0.58		23	1221520	0.27	0.31
	6	2777	4.06	3.26		24	1245496	0.57	0.78
	7	4038	0.24	0.38		25	2108632	5.71	5.65
	8	4482	0.30	0.45		26	$3.32*10^{7}$	5.98	6.28
	9	4948	0.24	0.12		27	$1.15*10^{9}$	12.44	16.50
	10	7374	1.45	2.37		28	$4.14*10^{10}$	5.95	5.39
	11	11091	0.95	0.25		29	$1.49*10^{12}$	5.44	5.42
	12	21577	3.58	2.13		30	$5.37*10^{13}$	1.69	1.28
	13	30951	0.14	0.16		31	$1.93*10^{15}$	1.28	1.45
	14	51875	3.41	2.80		32	$6.96*10^{16}$	0.30	0.42
	15	76802	0.34	1.28		33	$2.50*10^{18}$	0.24	0.15
	16	159817	5.51	3.56		34	$9.02*10^{19}$	0.27	0.32
	17	243758	0.34	0.28		35	$3.25*10^{21}$	0.10	0.05
	18	251391	0.07	0.02		36	$1.17*10^{23}$	0.00	0.05

5. CONCLUSIONS

In this paper a model for dictionary attack and brute force attack is developed. On the base of this model the security of passwords can be evaluated from a user's perspective. The proposed model has an important variable that must be set, the ordered set of reduced dictionaries (no dictionary contains a word another dictionary has). For accuracy of the model, it is necessary to carefully select and sort these dictionaries. In our study we used an empirical approach, but an expert's estimates can be involved as well.

If an entity doesn't have the resources to set the ordered set of reduced dictionaries carefully, we recommend that the entity consider the calculated value for the security of password $S(p_i)$ as the maximum expected value because an impostor is more likely to create a more effective ordered set of reduced dictionaries.

We are confident that this variable differs for different types of conditions. Some users can prefer different types of passwords, and moreover they can prefer different passwords to different systems. In our study we investigated the same information system and likely the same kind of users, but the results are not exactly the same. Of course, the first source of this disparity is randomness (selected password can be though as being a random variable), but it is probable that there can be other sources, perhaps a rising awareness in security.

Our method evaluates the security of passwords only from a dictionary attack and a brute force attack point of view. Indeed other aspects have to be considering, for example whether users write down their passwords. The next aspects are technology factors that are not considered in this paper. Although only selected factors are considered in our method we are confident about its usefulness because it can serve as a means for future password studies.

6. ACKNOWLEDGMENTS

This paper was created with the support of the Grant Agency of the Czech Republic, grant No. 402/08/P202 with title Usability Testing and Evaluation of Public Administration Information Systems and grant agency of Ministry of Interior of the Czech Republic, grant No. VD20062010A06 focused on research of new principles of risk management and citizen security.

REFERENCES

[1] M. Burnett, D. Kleiman. ed. Perfect Passwords. Rockland, MA: Syngress Publishing. 2006. p. 181. ISBN 1-59749-041-5.

[2] Ch. P. Garrison, An Evaluation of Passwords, On line CPA Journal- May 2008, Accesable http://www.nysscpa.org/cpajournal/2008/

[3] N. Haller, C. Metz, P. Nesser, M. Straw, A One-Time Password System, RFC2289, February 1998.

[4] L. J. Hoffman, Modern Methods for Computer Security and Privacy. 1st ed. New Persey: Prentice-Hall, Engelwood Cliffs, 1977.

[5] M. Hub, Data security – authentication (Originally in Czech). Pardubice: University of Pardubice, 2005. ISBN 80-7194-825-X.

[6] A. Jones, Password Authentication with insecure communication, ACM communications, volume 24, number 11, November 1981.

[7] D. Klein, A Survey of, and Improvements to, Password Security. In Unix Security Workshop II, USENIX Association, 1990.

[8] K. Renaud, Evaluating Authentication Mechanism. Security and usability. O'Reilly Media, Inc. 2005. pp 103-128. ISBN 0-956-00827-9.

[9] B. Soh, A. Joy, A Novel Web Security Evaluation Model for a One-Time-Password System, In Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), Issue , 13-17 Oct. 2003 pp 413 – 416.

[10] W. Stallings, Network Security Essentials: Applications and Standards, Upper Saddle River, NJ: Prentice Hall, 2000.

[11] J. Yan, A. Blackwell, R. Anderson, A. Grant, The Memorability and Security of Passwords. Security and usability. O'Reilly Media, Inc. 2005. pp 129-142. ISBN 0-956-00827-9.

Miloslav Hub is a senior lecturer in Institute of System Engineering and informatics of Faculty of Economics and Administration University of Pardubice, Czech republic. He graduated at Jan Perner Transport Faculty in 2002 in Transport Management, Marketing and Logistics branch; in 2005 he got Ph.D. with specialty of system engineering. During his career he was visiting Marmara University, Istanbul (Turkey), Siauliai University, Siauliai (Lietuva) and Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Lisabon (Portugal). Now days he is involved in several important projects, for instance projects of the Grant Agency of the Czech Republic with titles Usability Testing and Evaluation of Public Administration Information Systems and Usability of software tools for support of decision-making during solving spatially oriented problems. He has published over 30 papers. His research interests are in authentication systems and usability of public administration information systems.

Jan Capek served as professor of Managerial Informatics in Institute of System Engineering and Informatics of Faculty of Economics and Administration (FEA), University of Pardubice (UPa), Czech Republic. He was born 1946, graduate VŠDS Žilina at Slovakia year 1970, PGS ČVUT Praha 1976, candidate of science (PhD) ČVUT Praha 1981, habilitate assoc. prof. VŠDS Žilina 1990, professor of managerial informatics VŠB TU Ostrava 2002, 1990 – 1993 head of department of reliability and diagnostics VŠDS Žilina, 1993 – 1995 head of department of information systems UPa, 1995-1997 vice-dean FEA - UPa, 1997- 2000 vice rector UPa, 2002- 2007 dean FEA - UPa 2008 till now vice-dean for research FEA- UPa).

FINANCIAL RISKS AND CIRCUMVENTION OF PITCH POINT ENTERPRISE IN SUPPLY CHAIN

——AN ANALYSIS ON DANGDANG E-BOOKSTORE SUPPLY CHAIN MODEL

Wan Yanru
School of Management, Wuhan University of Technology

ABSTRACT

This paper analyzed the relationship between pitch point enterprises and discussed the features and carriers of their financial risks. Through a analysis on Dangdang E-bookstore supply chain model, this paper has discussed the practical circumvention of financial risks and offered the theoretical base of Dangdang's success by using Long Tail Theory.

Keywords: Supply Chain, Pitch Point Enterprise, SCM.

1. INTRODUCTION

SCM is a new management model emerged with enterprise collectivization and internationalization in global manufacturing industry[1].Through feed forward information stream (from demand side to supply side, such as order contract, manufacturing bill or purchasing bill etc.) and feedback material stream (material stream from supply side to demand side as well as its accompanied information stream, such as bill of lading, warehouse warrant, completion report, etc.), supplier, manufacturer, distributor and retailer are combined together to form a supply chain system with all the processes including planning, coordinating, operating and optimizing. SCM refers to the planning and controlling process centered on the material stream, information stream and capital stream between supplier and demander. In an era of competition between different supply chains rather than different enterprises, enterprise should play an active role in supply chain management, the core of which is to strengthen financial risk prevention. This paper aims to talk about the profound meaning of supply chain management on the perspective of financial risk circumvention in supply chain pitch point enterprises.

2. LITERARY REVIEW

Zhu Yanjun（2004）has done a research on supply chain coordinating system of clustered network model. He offered that clustered network supply chain is a network form between market and enterprise, which has been rapidly developed theoretically and practically recently. Its emergence has a great impact on industrial and theoretical circles. He also has elaborated related concepts on clustered network supply chain, and analyzed the coordinating system in clustered network supply chain.

Li Jifang（2008）has discussed the relationship among strategic partnership enterprises under supply chain management, and pointed out that the supply chain is the demand-supply network made up of raw material supplier, manufacturer, distributor, retailer and end customer in the process of product manufacturing and circulation. The essence of supply chain management is managing the relationship among strategic partnership pitch point enterprises. In a supply chain, the core enterprise is the coordinator and organizer in the supply chain. The core enterprise chooses its strategic partner enterprises, while the strategic enterprises run their business around the core enterprise, serving as its accessory appliances. This connection will help to achieve the overall optimization in a supply chain, and finally to reach Pareto Improvement among all the strategic partner enterprises in the supply chain.

Liu Kaijun and Zhang Zigang(2004)have done a research on motivation system of sharing information in scattered supply chain, and pointed out that the supply chain is the clustered enterprise network featured with resource external usage in the supply chain. Supply chain management emphasizes on improving the overall efficiency in the supply chain systematically. Due to the information asymmetry among the member enterprises, it is hard for a scattered chain to achieve overall performance optimization without effective motivation system. This research demonstrated the necessity and mission of information sharing motivation system through analyzing the value of sharing information in a scattered supply chain as well as the existent problems. Moreover, they offered a motivation system designing model based on delegating-surrogating theories,and analyzed several effective motivation methods.

Qi Guiqing, Yang Xihuai and Li Sen(2006), based on picturing the features of the clustered network supply chain, have used Dynamic Repeated Game Theory to analyze the competitive and cooperative situation among the upstream and down streams pitch point enterprises in the single stranded network supply chain, as well as in the common-valued paralleled supply chain. Furthermore, they offered a cooperative leveling situation and came to the conclusion that in the single stranded supply chain pitch points there is a huge cooperative chance, while in the paralleled common-valued supply chain there is a huge competitive chance rather than a cooperative one.

Zhou Yongqiang and Yang Zhongzhi(2004)mainly researched on the risk control in the competitions on the network supply chain. They maintained that the change in the competition environment leads to the competition among supply chains taking the place of among enterprises. In order to achieve a satisfactory result among supply chain partner enterprises, an enterprise should realize the origins of the risks in supply chain operation, and take effective methods to circumvent the risks. Risks in a supply chain can be divided into relationship risks and performance risks; the former one referring to the risks resulted from cooperation and the latter one referring to the possibility of not accomplishing the goals though enterprises enjoy a friendly cooperation. Thus, the key to control risks effectively is to enhance information communication, to cut redundant processes, to simplify supply chain and to build a appropriate interest division system as well as a soft supply chain.

All the researches above mainly focused on coordination and cooperation among supply chain enterprises and risk control of competitions in the supply chain. There is no standout research on financial risk on supply chain enterprises up till now.

3. DEFINITION ON FINANCIAL RISKS ON SUPPLY CHAIN PITCH POINT NTERPRISES

3.1 Supply Chain Pitch Point Enterprises

Porter (1985) has mentioned his Value Chain Theory in his book, Competitive Advantage, which demonstrating the value increasing process in a supply chain. As far as Porter concerned, firstly, flow of people and materials has been changed into output logistics, and then goes to marketing and sales, finally goes to the customers. This process, accompanied with a series of supporting actions, has formed the overall activities of an enterprise. Actually, in every value activity, it is possible to add the value. These activities is linked together to form the value chain of an enterprise. When the final value exceeded the overall cost, the enterprise can earn the profits and realize value increasing. Secondly, any enterprise supply chain exists in a value system, which is made up of the values linked together among the supplier, manufacturer, distributor and customer. Some experts said that the supply chain, in its essence, is a value chain which is based on the logistics and is increasing its value constantly during its process of transferring the materials. This transferring process of different materials in a supply chain is a value increasing process of the market values as well as the additional values. APICS has defined a supply chain as “A supply chain includes the value chain of an enterprise manufacturing products and offering services to customers internally and externally” in the ninth edition of the dictionary. In this way, a supply chain is an extend value chain. In other words, a supply chain is essentially made up of the extension of supplier value chain, customer value chain as well as the enterprise internal value chain, or a value system made of the overall value chain. Porter’s theory has not only described the single value chain within the enterprise, but also elaborated the operating process of the industrial supply chain on the perspective of a value chain, which may give rise to realizing cycle economic model of industrial supply chain.

A supply chain refers to a network structure based on the upstream and down stream linkage of raw material supplier, manufacturer, distributor, retailer and end customer in products manufacturing and transferring. In other words, it is a network including getting raw materials, manufacturing, sending the products to the customers as well as all the departments and enterprises involved in the process. This network can be divided into internal supply chain and external supply chain; the former one referring to a demand-supply network including enterprise internal manufacturing and transferring, along with purchasing department, manufacturing department, storing department, sales department involved in the process; while the latter one referring to the demand-supply network including enterprise external manufacturing and transferring, along with raw material supplier, manufacturer, storage and transportation company, retailer and end customer involved. Internal and external supply chain combined together to form a supply chain from raw material to customer. In this way, internal supply chain is a minimized external supply chain. For example, for the manufacturer, purchasing department equals to supplier in an external supply chain. The only difference lies in the larger scale, more complicated coordination work and more involved enterprises of external supply chain.

A supply chain pitch point enterprise refers to the enterprises participating in products manufacturing, offering service, and help to realize the products values in a industrial value chain. Picture 1 shows the interactive relationship among pitch point enterprises in several paralleled supply chains. Every pitch point enterprise has several upstream enterprises and down stream enterprises. In a vertical supply chain, upstream and down stream pitch point enterprises could be either cooperative or competitive, which can be seen in enterprise A, B, C. The enterprises in the paralleled situation are not the pitch point enterprises in this paper. They form an industry, the relationship among which could be either competitive within the same industry, or cooperative in different industries, which can be seen in enterprise B1, B2 and B3. Pitch point enterprise financial risk only exists between upstream and down stream enterprises in vertical supply chain.

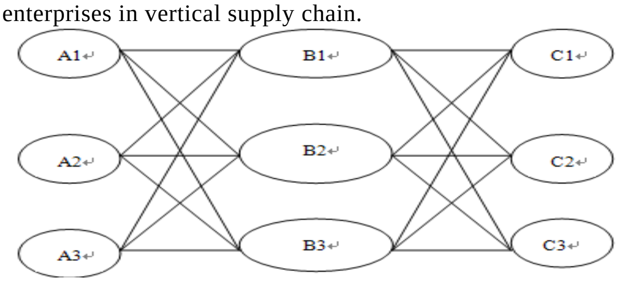


Figure 1. Relationship among Pitch Point Enterprises

3.2 Definition of Pitch Point Enterprise Financial Risk

Generally speaking, enterprise financial risk refers to unexpected and uncontrollable factors in an enterprise’s financial activities leading to the deviation between final financial results and expected business goals in a certain period within a certain scale, which may bring benefit loss to an enterprise. An enterprise’s financial activities line through the whole process of business operation. Risks could appear in collecting funds, long-term and short-term investing, or interest division.

However, for a pitch point enterprise, supplier and demander are linked together based on the transfer of materials or money. If upstream and down stream enterprises fail to effectively manage the logistics and fund flow simultaneously, it may cause loss to either or both side of the pitch point enterprises. The reflection of this possibility in fiancé is called pitch point enterprise financial risk.

The fundamental differences between pitch point enterprise financial risk and common enterprise financial risk lies in the incentive of the risks. The incentive of common enterprise financial risk could be either internal or external. The internal factor is resulted from an enterprise’s internal decision, such as China Aviation Oil Singapore Office failed in cash debt investment. The external financial risk is influenced by external environment or issues, such as due to American sub prime lending crisis, many of our export enterprises along the coast line are one the verge of bankrupt.

The incentive of financial risks in a pitch point enterprise (risk recipient) is caused by another pitch point enterprise linked with it (risk maker). In other words, financial risk of the recipient is made by risk maker’s activities, which bring financial losses to risk recipient. There is a direct cause effect sequence between risk maker’s activities and risk recipient’s losses. For example, upstream enterprise in a supply chain violates the contract and stops to supply materials to the down stream enterprise, or, lift the price unilaterally, and the liquidated damages could not make up for the losses of down stream enterprises. Or, down stream enterprise overruns the time limit and refuses to pay, which increases the fund cost and account cost on the upstream enterprise.

In this way, the carrier of pitch point enterprise financial risk is logistics and capital flow, the former one from the upstream enterprise to the down stream enterprise will bring about purchasing risk, while the latter one from down stream enterprise to upstream enterprise will lead to account risk.

4. DANGDANG MODEL AND IRCUMVENTION OF PITCH POINT ENTERPRISE FINANCIAL RISK

Statistics have shown that Dangdang now has 37.56 million registered members,15.6 million loyal customer (purchased more than twice), which account for 40% of all the online retailing users. Every month, Dangdang will load more than 4 million items of products, with 20 million daily page-view number and 1.2 million independent IP number every day. After the "Cross-provincial Extension Plan" in 2007, Dangdang has started its "Transparent Extension Plan" this year, aiming to expand its website from GradeⅠ,Ⅱ to Grade III Cities.In its development,Dandang's speed is six times faster than the average developing speed in B2C market Statistics in Chinese B2C Market Report in the 4th Season, 2007 has shown that Dangdang has ranked first in Chinese B2C market, covering 17% of market share, followed by Joyo with 14% of market share. Li Guoqing said, "Dangdang has got the key to profit. But to maintain the expansion in scale, profit is not the number one issue for Dangdang up till now."He also said, by 2009, Dangdang will have sold 2 to 3 billion yuan of books. Figure2 is the supply chain model of Dangdang.

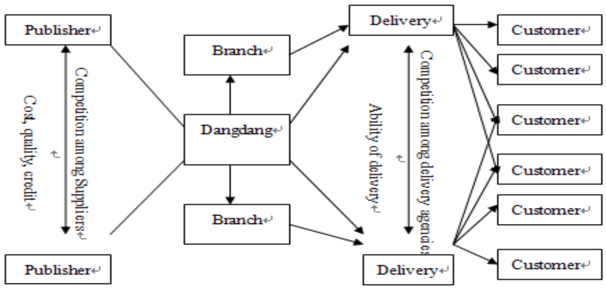


Figure 2. Dangdang Supply Chain

As a online bookstore, Dangdang is only a pitch point enterprise in the traditional book supply chain, but it has made such great performance. How does it develop so rapidly through supply chain management and financial risk circumvention?

(1)Dangdang Logistics Risk Control

The virtual storage online has helped Dangdang to lower its purchasing cost and purchasing risk. Through collecting the electronic information of published books, Dangdang has built a e-menu of books on sale. The books are stored in the warehouses in publishers or branch bookstores. Once Dangdang received an order, the books would be delivered to customers directly from publishers or branch bookstores. Virtual storage helped Dangdang break the limitation of physical space of traditional bookstores and offer millions of book lists to readers.

Meanwhile, due to the huge consuming figure of Dangdang, it has gained a great advantages in bargaining with upstream suppliers. In this way, Dangdang could attract many suppliers and get a bigger discount. Dangdang offers this discount to customers to stimulate book selling. Thus, a benign cycle is formed.

(2)Dangdang Capital Flow Risk Control.

Though Dangdang is a B2C network, there is no direct cash connection between Dangdang and the customers. Capital is recovered through delivery. It is well acknowledged that in Dangdang, customers pay on delivery. When the book is delivered to the customers, deliverer will get the money of the books. But what is the next step between deliverer and Dangdang? Will there be unrecorded revenue earned? Actually, Dangdang, including its branches, never sell on credit to deliverers or customers. Instead, Dangdang requires deliverer to buy the books from Dangdang on the discounted prices and then send the books to the customers. This could avoid the cost of money recovering, hence control money recovering risks.

5. AN ANALYSIS ON DANGDANG MODEL BY USING LONG TAIL THEORY

Long Tail Theory is an economic theory about breaking the bottle necks and to satisfy individual need. Large-scaled manufacturing to meet the need of customers is a exception under industrial conditions. Customers' real need will reveal without the physical restrictions. Long Tail is a special form of demander's economic scale. It focused on the comparative scale of the subdivided market in every Long Tail section, or the so-called scale that is made up of various demand and reflected on profits, rather than the demander's scaled economy on the short end.

In Long Tail coordinate, one dimension is quantity, the other is item; the head is the minor products with a big sales figure, while the tail is the major products with a small sales figure. Reading is a highly individual consuming. Different people have different reading habits and preference. Millions of readers will create astonishing buying power although a single person's buying power is comparatively small.

At the same time, Dangdang has provided a great many products choices by using the research function and other information advantages. For digital music and video products, as long as storing these products in the hardware, it is possible to transfer them through board-band network. Thus, there is almost no cost of manufacturing, storage and sales. For published books, electronic menu could also lower the sales cost, which enable Dangdang to make a great leap on the potential long tail. That is to say, website combines countless books together and makes them easy to find, convenient to get.

Besides, Dangdang closely connected demand and supply to reach a maximized matching. Dangdang's recommending system and searching function is like a filter that helps customers to find the information of the products they need in a short period of time. This saves the customers much time and enhances the consuming efficiency.

6. CONCLUSIONS

Pitch point enterprise financial risks in a supply chain appears if upstream and down stream enterprises fail to effectively manage the logistics and fund flow, which may cause loss to either or both side of the pitch point enterprises. The success of Dangdang model has shown that the key to circumvent financial risk is to use modern technology to alter the traditional supply chain. Therefore, the enterprise can take a favorable position in the supply chain game, bring more benefits to its partners and achieve win-win or multilateral wins among pitch point enterprises in the supply chain.

REFERENCES

[1] Li Yanjun. “A Research on Clustered Network Supply Chain”. Information Magazine, 2004(6)

[2] Zhang Yongcheng. “Supply Chain Protection Strategies Based on Guangxi Sugar Cycle Economic Industrial Chain”. Academic Forum, 2008(5)

[3] Li Jifang. “On Relationship among Strategic Partner Enterprises in Supply Chain Management”. Shandong Economy, 2008 (3)

[4] Qi Guiqing, Yang Xihuai, Li Sen. “ An Analysis on Competitive-cooperative Relationship in Clustered Network Supply Chain, Based on Repeated Game Theory”. Northeast University Journal, 2006(2)

[5] 5.Zhou Yongqiang, Yang Zhongzhi. “Risk Control in Network Supply Chain Competition”. Northwest Agro forestry Science and Technology University Journal, 2004(6)

[6] Yang Mincai, Wang Huailin. “ An Research on Joint Storage Control in Supply Chain Management”. Logistic Technology. 2003(3)

[7] Liu Kaijun, Zhang Zigang. “ An Research on Motivation System in Scattered Supply Chain Information Sharing”. Technical Management Research. 2004(2)

[8] Jerom Jays, trans Kang Yixiang. Wal-mart Operation. Harbin Press. 2004

[9] Sam Walton, John Hugh, trans Shen Zhiyan. Biography on the Founder of Wal-mart [M]. Shanghai Translation Publishing House. 2002

FAST SIGNATURE SCHEME FOR NETWORK CODING *

Mingxi Yang[1], Wenjie Yan[1], Huajing Fang[2]
[1]School of Computer Science and Technology, Wuhan University of Technology
[2]Dept. of Control Science and Engineering, Huazhong University of Science andTechnology
Wuhan, Hubei, 430070, China
Email: yangmx@whut.edu.cn, jie.wing@163.com, hjfang@mail.hust.edu.cn.

ABSTRACT

It has been proven that network coding can be used to improve the throughput for multicast transmission in networks, but such systems are very vulnerable to pollution attacks. In recent years, many schemes have been designed to prevent this attack, but most of these schemes are based on expensive discrete logarithms and Weil pairing operations on elliptic curves, these schemes are inefficient in verifying the integrity of messages, they are not suitable for those scenarios with low computing capability such as mobile Ad hoc networks and wireless sensor networks. In this paper, we proposed a novel signature scheme for network coding based on a homomorphic public cryptography. In our scheme, we use fast computation which greatly improved the efficiency of authentication when counteracting pollution attack for network coding. The time complexity of verification in our signature scheme is much less than those in the existing algorithms.

Keywords: Network Coding, Signature, Homomorphic.

1. INTRODUCTION

As the applications of network coding in traditional computer networks, wireless sensor networks [1] and peer-to-peer systems [2], more and more people begin to pay attention to this new technique. Network coding was first proposed by Ahlswede et al. [3]. They proved that the throughput of multicast in networks could be increased by allowing the node to encode its input messages to generate a new output one. Later, Li et al. [4] further proved that linear network coding is sufficient to achieve this purpose. Based on it, Ho et al. [5] [6] proposed a random linear network coding, which make the coding more easy and effective.

However, with the development of network coding comes many new challenges for the security. One of the security challenges is pollution attack, in which the adverse nodes can intentionally pollute the transmitted messages or inject the forged messages into the networks. Signature approaches based on hash functions such as SHA or MD5 used in traditional networks are no longer suitable for network coding, because the source's signatures were destroyed by the encoding process on each node.

1.1 Related Work

Homomorphic hashing function was first proposed by Krohn et al. [7], in their scheme, the source computes the hash values $h_1, h_2, \dots, h_n$ for the messages $m_1, m_2, \dots, m_n$ and distributes these hash values to all the nodes in the network. When a combination M of messages is received by a node, the hash value h_M of M can be compute from the hash function, and from the hash values $h_1, h_2, \dots, h_n$, then the message M is verified. A main drawback of this signature scheme is that it requires the sender to know the entire file in advance, before the authentication information can be computed. This limits the applicability of the scheme for transmission of streaming data; In Zhen Yu et al. [8], the forwarders compose the signatures for their output messages from those of input messages using the similar way that the output messages are composed from the input messages. Since each node appends the signatures to its output messages, its downstream nodes can verify the messages; Charles et al.[9] proposed a new homomorphic hashing scheme which is built on top of expensive Weil pairing operations [10], [11] over elliptic curves. All the schemes described above require expensive computation in verification, which greatly slow down the efficiency of verification. Jing Dong et al.[12] proposed a scheme which uses time-based checksum to allow the intermediate nodes to authenticate the received messages. However, this scheme requires time synchronization between senders and receivers, which is unpractical.

1.2 Our Contribution

The verification speed is a very important metric for evaluating performance of schemes. In this paper, we proposed a new signature scheme for network coding which could generate hash value of received packets quickly by using linear computation, it greatly improve the efficiency of verification, the verification speed in our scheme is much greater than those in [8] and [9] (it is shown in section 5). The homomorphic cryptography function in our scheme is based on Bresson's homonorphic encryption Algorithm which is first proposed by Emmanuel Bresson et al.[13] for encryption.

In the next section we will introduce the model of network coding briefly, our signature is proposed in section 3, and the proof of security is explained in section 4. Section 5 is the performance analysis of our scheme. Finally, this paper is concluded in section 6.

2. BACKGROUND

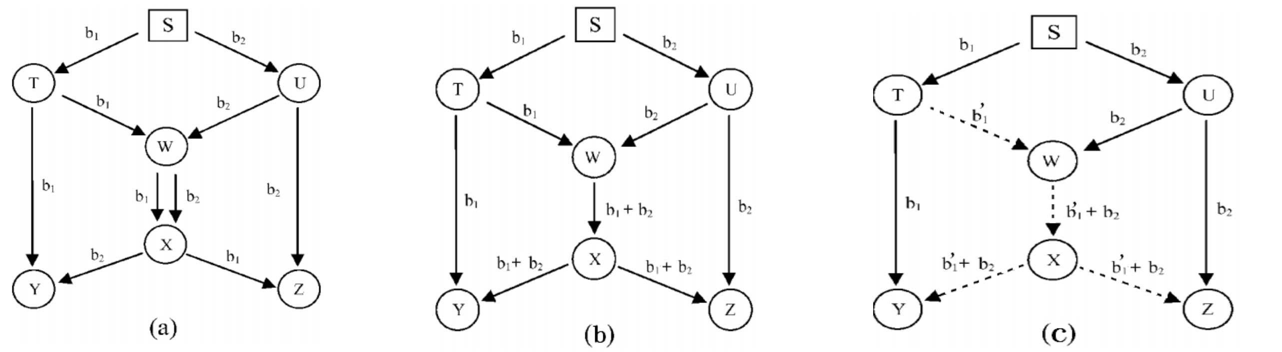
2.1 Linear Network Coding

In this paper, we model the network with a directed graph $G_d = (E, V)$, where E is the set of links and V is the set of vertices in the network. There is a source node $S \in V$ wish to send a set of original messages $MES=\{m_1, m_2, \dots, m_n\}$ to the set of destination nodes, $T \subset V$, each original message is augmented as: $M_i = (m_i, \overbrace{\underbrace{0,\dots,0}_{i-1},1,0,\dots,0}^{n})$, where $m_i \in Z_q$ and $(0,\dots,0,1,0\dots,0) \in Z_p^n$, thus the message M_i is separated into two parts, the first part is data part (m_i), and the second part is coding vector ($\overbrace{\underbrace{0,\dots,0}_{i-1},1,0,\dots,0}^{n}$), the data part is an element in Z_q and the coding vector part is a vector in Z_p^n.

Each node in network processes the received messages as follows. Upon receiving messages (vectors) $y_1, y_2, \dots, y_l$ on its l

* This research is partially supported by the National Natural Science Foundation of China (under Grant No.60874053).

income edges, a node computes and transmits message



(a) Traditional Multicasting (b) Multicasting over network coding (c) Pollution attack

Figure 1. Multicasting and Attacking.

$y=\sum_{i=1}^{l} c_i y_i$ on its outgoing edges, where each c_i acts as the coding coefficient.

When any destination node receives n linearly-independent vectors $w_1, w_2, \ldots, w_n$, it can recover the original messages as follows:

$$\begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{pmatrix} = \begin{pmatrix} w_1^R \\ w_2^R \\ \vdots \\ w_n^R \end{pmatrix}^{-1} \begin{pmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{n1} \end{pmatrix}$$

Where w_i^R denote the coding vector of w_i. Thus for any received message $w_i=(w_{i1}, w_{i2}, \ldots, w_{i,n+1})$ and $w_i^R=(w_{i2}, \ldots, w_{i,n+1})$.

Such as shown in Fig.1, assume that we multicast two data bits b_1 and b_2 from the source S to both the nodes Y and Z. In the Fig.1 (a), every channel carries either the bit b_1 or the bit b_2 as indicated, and every forwarder simply replicates and sends out the bit(s) received from upstream. Therefore, the channel from W to X is used twice. Fig.1 (b) depicts a different way which used in network coding to multicast the two bits b_1 and b_2 on the same network as in Fig.1 (a). This time, the node W derives the bit b_1+b_2 from the received bits b_1 and b_2, the channel from W to X transmits b_1+b_2, which is then replicated at X for passing on to Y and Z. Then, the node Y receives b_1 and b_1+b_2, from which the b_2 can be decoded, and the same to Z to decode b_1 from b_2 and b_1+b_2, in this way, all the channels in the network are used exactly once.

2.2 Pollution Attacks

A main security challenge is pollution attack in network coding, in which the adverse nodes can intentionally pollute the transmitted messages or create the forged messages and inject these messages into the encoded messages. These attacks prevent the destination nodes from recovering the source message from the encoded messages correctly, and what's more, the contaminated messages can quickly propagate into the networks and infect a large proportion of nodes, because these contaminated messages will be multicasted by the downstream nodes. As depicted in Fig.1 (c), the dotted lines denote the paths along which the contaminated messages propagate, b_1*'* is a corrupted message from node T, when the corrupted message b_1' combine with b_2, the new combination b_1'$+b_2$ is also corrupted, then the contamination passed down to the path WX, XY and XZ. When Y receive message b_1 and b_1'$+b_2$, it can't decode b_1 and b_2 correctly. Therefore, it is necessary for each node in the network to verify the integrity of every received message.

2.3 Bresson's Homonorphic Encryption Algorithm

In this paper, our homomorphic hash function is constructed based on Bresson's homonorphic encryption algorithm, his algorithm works as follows:

Let Q be a composite modulus product of two large primes, let CG be the cyclic group of quadratic residues modulo Q^2.

Key Generation: Choose a random element $\alpha \in Z_{Q^2}^*$, a random value $a \in [1, ord(CG)]$ and set $g=\alpha^2 \bmod Q^2$ and $h=g^a \bmod Q^2$. The public key is (Q, g, h) while the corresponding secret key is a.

Encrypt: Given a message $m \in Z_Q$, a random pad r is chosen uniformly and at random in Z_{Q^2} the cipher text (A, B) is computed as:

$A=g^r \bmod Q^2$ $\quad B=h^r(1+mQ) \bmod Q^2$.

Decrypt: Knowing a, one can compute m as follows

$$m=\frac{B/(A^a)-1 \bmod Q^2}{Q}.$$

For any two messages m_1, $m_2 \in Z_Q$, their cipher texts are $E(m_1)=(A_1, B_1)$ and $E(m_2)=(A_2, B_2)$, where:

$$A_1=g^{r_1} \bmod Q^2,\quad B_1=h^{r_1}(1+mQ) \bmod Q^2$$

$$A_2=g^{r_2} \bmod Q^2,\quad B_2=h^{r_2}(1+mQ) \bmod Q^2$$

Define: $E(m_1)\times E(m_2)=(A_1A_2, B_1B_2)$,

$$\because A_1A_2=g^{r_1+r_2} \bmod Q^2$$

$$B_1B_2=h^{r_1+r_2}[1+(m_1+m_2)Q] \bmod Q^2$$

$$\therefore E(m_1)\times E(m_2)=E(m_1+m_2).$$

So Bresson's homonorphic encryption Algorithm is an additive homomorphic function [13] [14]).

3. PROPOSED SIGNATURE SCHEME

In our signature scheme, the source node generates the signature for each message and appends the signature to the corresponding message. And the forwarder nodes use the public key and the appended signatures to verify the integrity of the received messages, and if needed they could use these signatures and the homomorphic function to compute new signatures for the corresponding encoded messages without knowing the source secrete key.

3.1 Homomorphic Hash and Signature Function

Bresson's algorithm [13] is originally used to encrypt message. Due to its additive homomorphic property, we recognized it can be used to construct homomorphic signature for network coding finely. Considering there is no need to encrypt or decrypt the messages in the signature scheme, and for the computation efficiency, we omit some operations which were designed for decryption, such as $A=g^r \bmod Q^2$, h^r and the decryption. Our homomorphic hash and signature functions are constructed based on Bresson's algorithm which works as follows:

Select a large prime q. $M=(m, c_1,\ldots,c_n)$, where $m \in Z_q$ is the combination of original messages, $(c_1,\ldots,c_n)$ is the global coding vector of M. We define the hash value of M as:

$$h(M)=h(m)=(1+mq) \bmod q^2$$

The signature of M is computed as:

$$\sigma(M)=h(M)^d$$

where d is the secret key of source node.

Both functions have the following two properties :

Property 1: $h(M)$ and $\sigma(M)$ are both additive homomorphic function, thus

$$h(M_1)\times h(M_2) = h(M_1+M_2)$$
$$\sigma(M_1+M_2)=\sigma(M_1)\times\sigma(M_2)$$

Proof:

(1) We first prove $h(M_1)\times h(M_2) = h(M_1+M_2)$.

$\because$ $h(M)= h(m)=(1+mq) \bmod q^2$

for any two messages M_1, M_2, we have

$$\begin{aligned} h(M_1)\times h(M_2) &= [(1+m_1q)][(1+m_2q)] \bmod q^2 \\ &= [1+(m_1+m_2)q+m_1m_2q^2] \bmod q^2 \\ &= [1+(m_1+m_2)q] \bmod q^2 \\ &= h(M_1+M_2) \end{aligned} \qquad (3\text{-}1)$$

Hence, this function is an additive homomorphic function.

(2) Then we prove $\sigma(M_1+M_2)=\sigma(M_1)\times\sigma(M_2)$.

$\because$ $\sigma(M)=h(M)^d \bmod N$

$\therefore$ for any two messages M_1, $M_2 \in Z_q$, we have

$$\sigma(M_1+M_2)=[h(M_1+M_2)]^d \bmod N$$

from formula (3-1) we have just proved, we know that: $h(M_1+M_2)= h(M_1)\times h(M_2)$, then we get

$$\begin{aligned} \sigma(M_1+M_2) &= [h(M_1)\times h(M_2)]^d \bmod N \\ &= h(M_1)^d\times h(M_2)^d \bmod N \\ &= \sigma(M_1)\times\sigma(M_2) \end{aligned} \qquad (3\text{-}2)$$

Thus $\sigma(M)$ is still an additive homomorphic function. ■

Property 2: Given a message M and an integer $c \in Z_p$, where $M=(m, c_1,\ldots,c_n)$, $m \in Z_q$ and $(c_1,\ldots,c_n) \in Z_p^n$, we have

$$\sigma(cM)=[\sigma(M)]^c \text{ and } h(cM)=[h(M)]^c$$

Property 2: Given a message M and an integer $c \in Z_p$, where $M=(m, c_1,\ldots,c_n)$, $m \in Z_q$ and $(c_1,\ldots,c_n) \in Z_p^n$, we have

$$\sigma(cM)=[\sigma(M)]^c \text{ and } h(cM)=[h(M)]^c$$

Proof: We use mathematical induction to prove it, since the two equations have similar proof, we only prove

$$\sigma(cM)=[\sigma(M)]^c. \qquad (3\text{-}3)$$

$$\begin{aligned} \sigma(cM) &= \sigma((c-1)M+M) \\ &= \sigma((c-1)M)\times\sigma(M) \quad \text{(from equation 3-2)} \\ &= [\sigma((c-2)M)\times\sigma(M)]\times\sigma(M) \\ &\cdots \\ &= \sigma(2M)\times\underbrace{\sigma(M)\times\cdots\times\sigma(M)}_{c-2} \\ &= \sigma(M+M)\times\sigma(M)^{c-2} \\ &= [\sigma(M)]^c. \end{aligned}$$

To sum up the above arguments, we conclude that for any $c \in Z_p$, the equation $\sigma(cM)=[\sigma(M)]^c$ is always valid. ■

3.2 Our Signature Scheme

Our scheme is defined as follows:

- ***Setup:***
  Select u, v, q from large primes, $length(u)\approx length(v)$, $length(uv)\approx length(q^2)$ and $q^2<uv$ (where $length(x)$ denotes the bit length of x), let $N=uv$, keep u and v secretly.
  And random select n different elements $r_1,\ldots,r_n$ from G, G is a multiplicative group with prime order p.
  We also select integers d, e_1, $e_2<\varphi(N)$ such that $d\times e_1+e_2\equiv 1 \bmod \varphi(N)$, where $\varphi(N)=(u-1)\times(v-1)$. The secret key is d and the public key is $pk=(N, e_1, e_2, r_1,\ldots,r_n)$.
- ***Sign:*** For $i=1,2,\ldots,n$, the source node outputs the signature $\sigma(M_i)$ on source message M_i as:
  $$\sigma(M_i)=\frac{[(1+m_iq) \bmod q^2]^d}{r_i} \bmod N.$$
  Where $M_i=(m_i,\overbrace{\underbrace{0,\ldots,0}_{i-1},1,0,\ldots,0}^{n})$, $m_i \in Z_q$ and $\overbrace{\underbrace{0,\ldots,0}_{i-1},1,0,\ldots,0}^{n} \in Z_p^n$.
- ***Combine:*** Given a set of coefficients $c_1, c_2, \ldots, c_n$, and a sequence of messages $W_1,\ldots,W_l$ along with their signatures $\sigma(W_1),\ldots, \sigma(W_l)$, where $W_i=(w_i,c_{i1},\ldots,c_{in})$, outputs the combination message W_0 as: $W_0=(w_0, c_{01},\ldots,c_{0n})$,
  where $w_0=\sum_{i=1}^{l} c_iw_i \bmod q$ and $c_{0j}=\sum_{i=1}^{l} c_ic_{ij} \bmod p$ for $j=1,\ldots,n$. The signature on W_0 is computed and outputted as: $\sigma(W_0)=\prod_{i=1}^{l}\sigma(W_i)^{c_i} \bmod N$.

From this phase, we can conclude that the intermediate nodes can generate valid signature for any messages in subspace $V \overset{def}{=} span\{M_1, M_2,\ldots,M_n\}$ without knowing source private keys, where $M_1,\ldots,M_n$ are the source messages.

- ***Verify:*** Given encoded message $W_0=(w_0, c_{01},\ldots,c_{0n})$ and signature $\sigma(W_0)$, $\sigma(W_0)$ is a valid signature on W_0 iff
  $$[\sigma(W_0)\cdot\prod_{i=1}^{n} r_i^{c_{0i}}]^{e_1}\cdot h(W_0)^{e_2} \bmod N=1.$$

Correctness:

Since $W_0=(w_0, c_{01},\ldots, c_{0n})$, we know that $(c_{01},\ldots, c_{0n})$ is the global coding vector of W_0, then w_0 is a linear combination of original messages with this coding vector, thus

$$W_0=\sum_{i=1}^{n} c_{0i}M_i$$

And

$$w_0=\sum_{i=1}^{n} c_{0i}m_i \bmod q,$$

From the *combine* phase, we know that signature $\sigma(W_0)$ on W_0 is also an combination of original signatures $\sigma(M_i)$ ($i=1,\ldots,n$) with this coding vector, thus

$$\begin{aligned} \sigma(W_0) &= \sigma(\textstyle\sum_{i=1}^{n} c_{0i}M_i) \\ &= \textstyle\prod_{i=1}^{n}\sigma(c_{0i}M_i) \\ &= \textstyle\prod_{i=1}^{n}\sigma(M_i)^{c_{0i}} \bmod N \end{aligned}$$

. (By equation 3-3), where M_i is the source message.

According to $h(W_0)=\prod_{i=1}^{n} h(M_i)^{c_{0i}}$, we have

$$\sigma(W_0)=\prod_{i=1}^{n}\sigma(M_i)^{c_{0i}}=\prod_{i=1}^{n}\left(\frac{h(M_i)^d}{r_i}\right)^{c_{0i}}$$
$$=\frac{\prod_{i=1}^{n}[h(M_i)^{c_{0i}}]^d}{\prod_{i=1}^{n} r_i^{c_{0i}}}=\frac{h(W_0)^d}{\prod_{i=1}^{n} r_i^{c_{0i}}} \qquad (3\text{-}4)$$

Then

$$[\sigma(W_0)\cdot\prod_{i=1}^{n} r_i^{c_{0i}}]^{e_1}\cdot h(W_0)^{e_2} \bmod N$$
$$=[\frac{h(W_0)^d}{\prod_{i=1}^{n} r_i^{c_{0i}}}\cdot\prod_{i=1}^{n} r_i^{c_{0i}}]^{e_1}\cdot h(W_0)^{e_2} \bmod N$$
$$=[h(W_0)^d]^{e_1}\cdot h(W_0)^{e_2} \bmod N$$
$$=h(W_0)^{d\cdot e_1+e_2} \bmod N$$
$$=h(W_0)^{\varphi(N)} \bmod N$$
$$=1$$

4. SECURITY ANALYSIS

In this section, we will prove that our scheme is secure for pollution attacks. We describe our proof as follows.

a. *Forge a message for a given signature.*

In this phase, we focus on the problem that whether an adversary could compute a forged message $Y' \notin V$ by a given signature.

Let $Y \in V$ and $\sigma(Y)$ is the signature on Y, where $Y=(y_1, y_2,\ldots,y_{n+1})$, if an adversary wants to derive a forged message Y' from packet $(Y\|\sigma(Y))$, he can proceed as follows:

(1) find a forged message $Y'=(y_1', y_2, \ldots, y_{n+1})$ and $y_1' \neq y_1$ (namely Y' and Y have the same coding vector) such that $\sigma(Y')=\sigma(Y)$.

Theorem 1: For any two elements x and y in Z_q, if $x\neq y$, then $h(x)\neq h(y)$.

Proof: we assume that $h(x)=h(y)$ for $x\neq y$, without loss of generality, let x>y.

Since $h(x)=(1+xq) \bmod q^2$ and $h(y)=(1+yq) \bmod q^2$, and $h(x)-h(y)=0$, we have

$$[(1+xq) - (1+yq)] \bmod q^2=0,$$

Thus $(x-y)q \bmod q^2=0$, then $(x-y)q=r\times q^2$, where $r \in Z$, thus $x-y=rq$, but $x,y \in Z_q$, it means that $x-y<q$, then $r=0$, thus $x=y$, this is contradictory to our assuming. ■

Theorem 1 told us that there is no such message $Y'=(y_1', y_2,\ldots,y_{n+1})$ such that $h(Y')=h(Y)$. from equation (3-4) we know that $\sigma(Y)=\frac{h(Y)^d}{\prod_{i=1}^{n} r_i^{y_{i+1}}} \bmod N$, as $h(Y') \neq h(Y)$, then we have $\sigma(Y')\neq \sigma(Y)$, this prove that method (1) is infeasible.

(2) Find a forged message $Y'=(y_1, y_2',\ldots,y_{n+1}')$ (namely Y' and Y have same data part y_1 but different coding vectors) such that $\sigma(Y')=\sigma(Y)$.

As Y' and Y have the same data part y_1, then $h(Y')=h(Y)$. From equation (3-4) we can find that if $\sigma(Y')=\sigma(Y)$, thus $\frac{h(Y)^d}{\prod_{i=1}^{n} r_i^{y_{i+1}}} \bmod N=\frac{h(Y')^d}{\prod_{i=1}^{n} r_i^{y_{i+1}'}} \bmod N$, as $h(Y')=h(Y)$, then $\prod_{i=1}^{n} r_i^{y_{i+1}}=\prod_{i=1}^{n} r_i^{y_{i+1}'}$, it means that an adversary has to find a coding vector $(y_2',\ldots,y_{n+1}')\neq(y_2,\ldots,y_{n+1})$ such that $\prod_{i=1}^{n} r_i^{y_{i+1}}=\prod_{i=1}^{n} r_i^{y_{i+1}'}$, this is equivalent to solve the *Diffie-Hellman* problem.

b. *Generating a forged signature for a given message vector.*

In this phase, we assume that the given message is not an element in subspace V.

Theorem 2: given a message $Y \notin V$, $Y=(y_1, y_2,\ldots,y_{n+1})$, it is infeasible for an adversary to generate a valid signature $\sigma(Y)$ on Y without source's secret key.

Proof: from the phase *sign*, we know that $\sigma(M_i)=\frac{[(1+m_iq) \bmod q^2]^d}{r_i} \bmod N$, where M_i (i=1,…,n) is a source message. Form this equation, we can find that the computation of the signature on message Y could be processed as $\sigma(Y)=\frac{[(1+y_1q) \bmod q^2]^d}{\prod_{i=1}^{n} r_i^{y_{i+1}}} \bmod N$, but this computation requires the knowledge of secret key d.

5. VERIFICATION EFFICIENCY

In network coding, verification at the forwarders should be done as fast as possible, it's the bottleneck of the whole network and prevents the source from sending messages at the optimal rate. Hence, verification speed is the most important metric for evaluating performance of schemes.

Let φ be a prime number and ψ a power of different prime with $\varphi<<\psi$, E is an elliptic curve over Z_ψ. In scheme [8] and [9], every original message is a vector with dimension k, the source then append a n-dimension coding vector on it, such as $X=(x_1, x_2, \ldots, x_k, c_1,\ldots, c_n)$, where $x_i, c_i \in Z_\varphi$.

Charles et al.[9] have shown the verification of a signature in their scheme requires $O(klog^{2+\epsilon}\psi)$ bit operations, where $\epsilon>0$. In [8], the verification of a signature requires about $k+n+1$ modular exponentiations, every modular exponentiation requires $O(log(1+\epsilon)(log^2\varphi))$ bit operations, where $log(1+\epsilon)$ is the length of coefficient, then the verification requires $O[(1+n+k)log(1+\epsilon)(log^2\varphi)]$ bit operations.

In our scheme, the source message is given as $M_i=(m_i, c_1,\ldots, c_n)$ (i=1,..,n), thus the original message m_i is 1-dimension, but they have the same size with those original messages in [8] and [9] (namely $length(m_i)=length(x_1, x_2, \ldots, x_k)$). We do not separate the original message into k dimensions, the verification of a signature requires about n+2 modular exponentiations: so the time complexity is $O[(2+n)log(1+\epsilon)(log^2\varphi)]$. TABLE 1 shows the time complexity of verification.

Table 1. Verification of message (bit operation)

our scheme	$O[(2+n)log(1+\epsilon)(log^2\varphi)]$
[8]	$O[(1+k+n)log(1+\epsilon)(log^2\varphi)]$
[9]	$O(klog^{2+\epsilon}\psi)$

Since $\varphi<<\psi$ and in practical network, $k>>n$ and the length of the coefficient $log(1+\epsilon)$ should be large enough, then as depicted in TABLE 1, it's easy to conclude that the verification efficiency in our scheme is much less than those in others:

$$\begin{aligned}
[9]&=O(klog^{2+\epsilon}\psi)\\
&=O(klog^{\epsilon}\psi log^2\psi)\\
&>O(klog^{\epsilon}\psi log^2\varphi)\\
&>O[(k+1)log(1+\epsilon)(log^2\varphi)]=[8]\\
&>O[(2+n)log(1+\epsilon)(log^2\varphi)]\\
&=\text{ours}
\end{aligned}$$

Thus the comparison is [9] >[8]>ours.

The authors in [8] have compared the verification between [8] and [9] in experiments, the result is shown in TABLE 2.

Table 2. Verification efficiency in [8] and [9]
(*k*: 256-bit , *n*: 128-bit and codeword: 1024-bit)

	[8]	[9]
Message verification (per message)	1.44s	16.54s

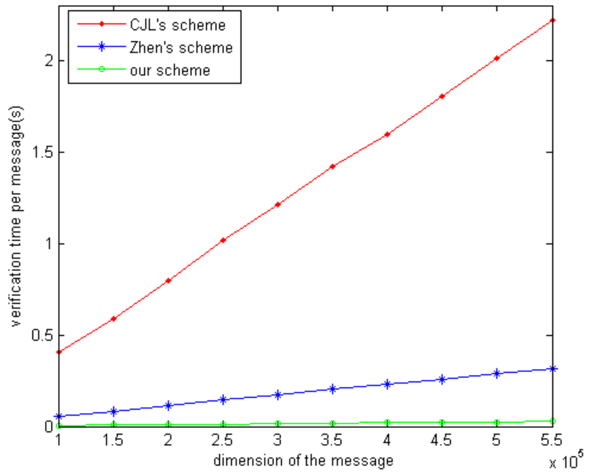


Figure 2. Comparison of verification efficiency (The length of each codeword is 1024 bits).

Fig. 2 shows the comparison of verification among [8], [9] and ours. It's clearly that our scheme perform much greater than those in [8] and [9].

6. CONCLUSIONS

This paper proposed a novel signature scheme to defend against pollution attacks. In this scheme, the corrupted messages could be filtered out by intermediate nodes. As network coding requires each node in network to verify the integrity of received messages for security, verification speed is the most important metric for evaluating performance of schemes. Our signature scheme proposed in this paper is more efficient for forwarders to verify the received message in network coding, and it have been proven that the verification efficiency of this scheme is much greater than those in [8] and [9]. This scheme is suitable for those network coding systems whose devices have low computing capability.

REFERRENCES

[1] D.Petrovic, K.Ramchandran, and J.Rabaey, "*Overcoming Unturned Radios in Wireless Networks with Network Coding*", in IEEE Transactions on Information Theory, Vol. 52, No. 6, pp. 2649-2657, 2006.

[2] C.Gkantsidis and P.Rodriguez, "*Network Coding for Large Scale File Distribution*", in Proc. IEEE INFOCOM, 2005.

[3] R. Ahlswede, N. Cai, S.Li, and R. W. Yeung, "*Network information flow*," IEEE Trans. Inf. Theory, vol. 46(4), pp. 1204-1216, 2000.

[4] S. Li, R. Yeung, and N. Cai, "*Linear Network Coding*", in IEEE Transactions on Information Theory, Vol 49, No. 2, pp. 371381, 2003.

[5] T. Ho, R. Koetter, M. M´edard, D. R. Karger, and M. Effros, "*The benefits of coding over routing in a randomized setting,*" in International Symposium on Information Theory (ISIT), 2003.

[6] T. Ho, M. M´edard, J. Shi, M. Effros and D. R. Karger, "*On randomized network coding,*" In proc. 41st Annual Allerton Conference on Communication Control and Computing, Oct. 2003.

[7] M.N.Krohn, M.J.Freedman, and D.Mazi´eres, "*On-the-fly verification of rateless era-sure codes for efficient content distribution,*" IEEE Symp. Security and Privacy, Oak-land, CA, pp. 226-240, May 2004.

[8] Zhen Yu, YaWen Wei, Bhuvaneswari Ramkumar, and Yong Guan, "*An Efficient Signature-based Scheme for Securing Network Coding against Pollution Attacks*" INFOCOM 2008. The 27th Conference on Computer Communications. IEEE, April 2008.

[9] D. Charles, K. Jian, and K. Lauter, "Signature for *Network Coding*", Technique Report MSR-TR-2005-159, Microsoft, 2005.

[10] A. Menezes, T. Okamoto, and S. Vanstone, "*Reducing Elliptic Curve Logorithms to Logorithms in a Finite Field*", in IEEE Transactions on Information Theory, Vol 39, No. 5, pp. 1639-1646, 1993.

[11] V. Miller, "*Short Programs for Functions over Curve*", unpublished manuscript, crypto.stanford.edu/miller/, 1986.

[12] Jing Dong, Reza Curtmola, Cristina Nita-Rotaru, Practical Defenses Against Pollution Attacks in Intra-Flow Network Coding for Wireless Mesh Networks, Proc. of The Second ACM Conference on Wireless Network Security(WiSec 2009), Zurich, Switzerland, March 2009.

[13] Bresson E, Catalano D, Pointcheval D. "*A simple public key cryptosystem with a double trapdoor decryption mechanism and its applications*," In: Laih CS, ed. Aciacrypt 2003. LNCS 2894, Berlin: Springer-Verlag, 2003. 37−54.

[14] SUN Zhong-Wei, FENG Deng-Guo, WU Chuan-Kun, "An Anonymous Fingerprinting Scheme Based on Additively Homomorphic Public Key Cryptosystem". In Journal of Software: 2005,vol.16, No.10,pp1816-1821.

Mingxi Yang received her Ph.D. degree in Computer Applied Technology from Wuhan University of Technology, China in 2007, and B.S. degree from Huazhong University of Science and Technology, China in 1982. She works as a associate professor and Director of the Institute of Software in School of Computer Science and Technology at Wuhan University of Technology, in China now. Her research interests are in computer network and network security.

Wenjie Yan received his B.S. degree in Science of Information and Computation from Wuhan Institute of Technology, China in 2007. He is now working toward his M.S. degree in School of Computer Science and Technology at Wuhan University of Technology, China. His research interest is network security.

ENHANCING PRIVACY PRESERVATION FRAMEWORK IN DISTRIBUTED TRUST MODEL

Yusheng Zhong[1], Bo Yang[2], Mingwu Zhang[3], Wenzheng Zhang[4]
[123]College of Informatics South China Agriculture UniversityGuangzhou, Guangdong, 510642 China
[4]National Laboratory for Modern CommunicationsChengdu, Sichuan, 610041 China
Email: zhongyusheng3000@163.com, csmwzhang@gmail.com, byang@scau.edu.cn, wzzhang@163.com

ABSTRACT

Distributed trust management implements authorization, delegation, and access control with a decentralized and flexible manner. Credentials submitting and policies interacting will disclose owner's sensitive attributes and privacy. In this paper, several solutions about establishing trust with privacy preserving are introduced, including secure and anonymous credentials, sensitive policies, balance of trust and privacy, risk analyzing and evaluating. The privacy enhancing frameworks for trust establishing are described to enforce the trust and privacy balance. Several potential solutions are also proposed and discussed.

Keywords: Trust Management, Credential, Sensitive Policy, Privacy Preservation

1. INTRODUCTION

Decentralized trust management, introduced by Blaze[1], is a flexible approach to implement distributed access control where the access decisions are based on the policy statements. Each credential contains a set of attributes about the owner. Privacy and security are often intertwined when establishing trust relation using credential-based model in decentralized environment. Privacy not only focus on user data but also include access policy, credential, and even the ownership of the credential. Trust is a monotonic model that $s \preceq t$ means that *t* denotes at least as high a trust-level as *s*. For scalability and efficiency considerations, trust evaluation is constrained to information provided by directly connected nodes, i.e. it is based on local interactions.

However, there is a tension between the privacy preserving and the releasing controlled information when a user submits credentials for establishing and verifying trust metric. While the focus of current research focuses on how to design a secure protocol to protect user's privacy, identity, and data with hidden identity or pseudonym fashion. The security and privacy issues in current study have not been explored in any depth.

Privacy enhancing security in trust management system will aspect for:

- Protecting the requester and the provider anonymity, pseudonymity, unlinkability, unobservability.
- Protecting confidentiality and integrity of attribute data in submitted credential.
- Hiding and protecting service provider's sensitive policy.
- Privacy is controllable and risk is aware and can be evaluated during trust negotiation and trust establishment.

2. TRUST MODEL WITH RISK EVALUATING

Definition 1(**Trust structure with risk metric**) A trust module with privacy risk metric is defined as a trust value domain *D* with two relation operation denote as:

$$T = (D, \preceq, \unrhd) \tag{1}$$

where *D* is a lattice which has a bottom element $\perp$ and top element $\top$ satisfying a trust domain of authority and delegation, and two partial relation, the trust ordering $\preceq$, and the risk ordering $\unrhd$. We use partial relation operator $\preceq$ to describe the trust operator when trust evidence submits, and $\unrhd$ to describe the risk value for evidence and credential be disclosed In order to safeguard the sensitive credential from malicious access, we use privacy entropy to describe credentials' attribute sensitivity. Before principal submit credentials, he or she expects the least privacy loss for attributes disclosure in the set of credentials. A trust model with risk evaluating and privacy considering refers to Fig.1. In the figure, it introduces a risk analysis including risk evaluating algorithm, privacy rule database, evidence rule, risk/privacy computing module, risk monitor module, and trust cycle management module.

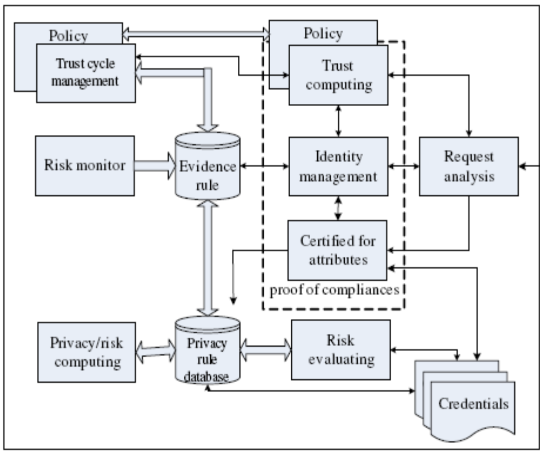


Figure 1. Trust with Risk Evaluating

-Proof of compliance modules, including trust computing module, identity management module and certified for attributes module, is designed for compliance check in the credentials chain that satisfying the users' request.

-Risk evaluating module, associated by credentials disclosing and privacy/risk rule database, is to evaluate the privacy risk when credential be submitted during the trust negotiation.

-Privacy/risk computing module, is called by risk evaluating module that use two dimension opinion space including a trust opinion space $OS_t \in [0,1]$ and a privacy opinion space $OS_p \in [0,1]$ with a initial value $(OS_t, OS_p) = (0,0)$ which means a new user has not any trust metric and no privacy disclosing, and $(OS_t, OS_p) = (1,0)$ is a best model which means that the system has established the trust relation without any privacy leaking.

-Risk monitor module manages a threshold and interactive with evidence rule database and PoC module.
-Evidence rule and privacy rules database store the policy rule, credential risk rule, privacy opinion description, sensitive resource rule etc., which can be a policy expression, privacy formula or resource description.

3. SECURE AND PRIVACY CONSIDERATION

3.1 Secure and Anonymous Credentials

Anonymous credential(AC) scheme, let A prove to B that C has given him a credential in a hidden manner, is a useful way to protect privacy of the credential owner. Non-linkable and forward secure are important for an AC system. Up to now, the key model of the AC scheme is to use zero-knowledge proof(ZKP), recent following by non-interactive ZKP, which is a time-consuming operating. Moreover, in existing AC systems, the revocation authority can link the transaction of any user in a subset T of users in $O(\log|T|)$ fake failed sessions. Furthermore, anonymous credential can only protect the identity privacy, but it cannot protect the attributes value in attribute-based credentials.

Hidden credential[5] and OACert[6] are useful schemes for attribute value preserving. Hidden credentials allow messages to be encrypted against complex policies, protecting policies from leaking to unqualified recipients and allowing recipients to use combinations of credentials without even knowing their existence. In Oblivious Attribute Certificates (OACerts), certificate holder can select which attributes to use and how to use them. In particular, a user can use attribute values stored in an OACert obliviously, i.e., the user obtains a service if and only if the attribute values satisfy the needed policy of the service provider, yet the service provider learns nothing about these attribute values.

To improve the security of credential for privacy consideration, we consider the following schemes in the system:
Attributes control. In general, credential owner has no control over what happens with his attribute data after his transactions end. We adopt a verifiable & undeniable encryption scheme to guarantee the provider knowing the user might decrypt and operate the information.
Forward secure credentials. Forward secure credential means that credential attribute disclosure will not expose the security and privacy that has used it before. Unlinkability and untransferability of credential submitting is crucial properties for forward secure credential systems.

3.2 Sensitive Policy Protect

Policies guarding resources are often involve multiple credentials. Trust access control policy π formalized as follows:

$$\begin{aligned} \pi &::= p|\lambda x : P.\tau \\ p &::= a \in P|x : P \\ \tau &::= \exp(p)|b \mapsto \tau;|\tau|\pi(p)|op(\tau_1, \tau_2, \ldots, \tau_n), \end{aligned} \tag{2}$$

where op is an arithemetic or logic operator.
b:=p=$|b$ bop b; where bop is a Boolean operator.

Existing policies may not provide adequate guarantees to deal with new exposures and vulnerabilities. We can introduce a policy-based cryptography to perform policy enforcement. Hidden credential be used to encrypt the multiple attribute value according to the corresponding policy, and only the credentials that satisfies the policy can decrypt the information.

The sensitive policy protection schemes above have two disadvantage:
(1)Policy must be fix and static, which means that policy can be expressed in advance;
(2)Policy must be monotonic so that more credentials disclosing will obtain access abilities. Most trust management systems assume monotonicity that additional credentials can only result in the increasing of privilege.

It considers further studies as follows:
Dynamic sensitive policy secure. Prasad and Zhou introduced a dynamic policy that we can adjust the policy expression during trust negotiation dynamically in [9]. As the dynamic of the policy, we cannot use hidden credential and OACert schemes to encryption policy. It is an open issue to study it.
Non-mononotonic policy Compared with nonmonotonic policies, monotonic ones are less flexible and cannot express policies such as"Chinese wall policies" and"separation of duties(SoD)". Attributed based cryptography allows a user's private key to be expressed in terms of multiple attribute expressions. We are constructing non-monotonic policy security model with attributed-based cryptography in a special set model.

3.3 Risk Evaluation in Trust Management

In several distributed environments, trust establishing will require some levels of risk to be tolerated. We proposed an ordered semiring-based trust model in [10], which can evaluate the risk during the trust establishing. In [2], authors introduced a solution to establish trust based on history interactions among parties, which held the untraceability and anonymity for interactions submitter, which designed a anonymous protocol using history-based signature that is an extension of group signature scheme.

A widespread risk and privacy analysis is reinforced as follows:

Risk evaluation vs privacy dosclosure. Privacy metric with trust be formulated to measure the quantify privacy loss. Trust is a measure of uncertainty with its value represented by entropy, and presented two trust models: entropy-based model and probability-based model in [8]. It is a formal way to describe risk metric using logic-based operator.

Balance of privacy and trust. It goes without saying that there is a great need for privacy in decentralized environment. However, trust and privacy tradeoffs are a normal part of attributes, identifiers, and possess of data etc. For example, what is the balance point between privacy preserving and trust establishing? A semiring privacy model during trust establishing in [10]. It deem that information entropy and probability model can be adapted to describe the privacy disclosure metric, and use information theory viewpoint to describe privacy risk.

4. CONCLUSIONS

Establishing trust among entities with minimized privacy and risk disclosure is crucial in decentralized trust environments. In this paper, we proposed a trust framework with risk evaluating and privacy preserving, including secure credentials, sensitive policy, trust and privacy, and risk evaluation in distributed trust system. Security requirements, such as anonymous credential, identity hidden, attribute protect, and balance of privacy, security and trust, are discussed.

5. ACKNOWLEDGEMENT

The authors thank the anonymous reviewers for their helpful suggestion. This work is supported by the National Natural Science Foundation of China under Grant 60773175 and 60673077, the Foundation of National Laboratory for Modern Communications under Grant 9140C1108020906, and the Innovation Fund for Technology Based Firms of China/Guangzhou under Grant 08C26214411225 and 2007V41C0311.

REFERENCES

[1] M. Blaze, J. Feigenbaum, and J. Lacy. Decentralized trust management. In Proc. of the 1996 IEEE Symposium on Security and Privacy, pages 164-173. IEEE Computer Society Press, May 1996.

[2] Blaze, M., Feigenbaum, J., Ioannidis, J., Keromytis, A.: Rfc2704: The keynote trust-management system version 2 (1999).

[3] Laurent Bussard, and Refik Molva. Establishing Trust with Privacy. Security Protocols 2004, LNCS 3957, pp. 199-209,Springer-Verlag, Berlin, 2006.

[4] Andreas Pashalidis and Chris J. Mitchell. A security model for anonymous credential systems. Advances in Cryptology, EUROCRYPT 2004, LNCS vol. 2045, pp. 93-118. Springer-Verlag, Berlin, 2004.

[5] Mira Belenkiy, Melissa Chase, Markulf Kohlweiss, and Anna Lysyanskaya. P-signatures and noninteractive anonymousc credentials.LNCS vol.2058, pp. 125-134. Springer-Verlag, Berlin, 2007.

[6] Jason E. Holt, Robert W. Bradshaw, Kent E. Seamons, and Hilarie Orman. Hidden credentials. In Proc. of the 2nd ACMWorkshop on Privacy in the Electronic Society, 2003.

[7] J. Li, and N. Li. OACerts: Oblivious Attribute Certificates. ACNS 2005, LNCS 3531, pp. 301-317, 2005.

[8] Y. Zhong and B. Bhargava, Using entropy to trade privacy for Trust. in Proceedings of Security and Knowledge Management(SKM), 2004, Amherst, NY, Sept. 2004.

[9] YL. Sun, W. Yu, Z. Han. Information theoretic framework of trust modeling and evaluation for ad hoc networks. IEEE Journal on Selected Areas in Communications vol.24, no.2: 305-317. 2006.

[10] A. Prasad Sistla, and Min Zhou. Analysis of dynamic policies. Information and computation. vol.206, pp.185-212,2008.

[11] M. Zhang, B. Yang, S. Zhu, and W. Zhang. Ordered semiring-based trust establish model with risk evaluating. International journal of network security, vol.8, no.1:96-101, 2009.

[12] Changyu Dong, Giovanni Russello and Naranker Dulay. Privacy-Preserving Credential Verification for Non-monotonic Trust Management Systems. Computer Network Security Fourth International Conference, MMM-ACNS 2007 St. Petersburg, Russia, September 13–15, 2007.

[13] Jim, T. A trust management system with certified evaluation. In: P' 01: Proceedings of the 2001 IEEE Symposium on Security and Privacy, Washington, DC, USA, pp. 106–115. IEEE Computer Society, Press, Los Alamitos , 2001.

[14] Marcin Czenko, Sandro Etalle, Dongyi Li and William H. Winsborough, An Introduction to the Role Based Trust Management Framework RT, Lecture Notes in Computer Science, Volume 4677/2007.

[15] Junji Nakazato , Lihua Wang and Akihiro Yamamura, Privacy Enhancing Credentials, Lecture Notes in Computer Science, Volume 4846/2008.

THE IMPROVED SSL PROTOCOL BASED ON AH

Wang Haohao
Parallel Processing & Computing Lab, Parallel Processing Institute
Wuhan University of Technology
Wuhan, Hubei, China
Email:whhbest@126.com

ABSTRACT

With the rapid development of Internet, people depend more and more on computers, and the security of sensitive information transporting by Internet is paid much more attentions. So series of safe network transport protocols are given, such as application layer authentication protocol Kerberos, secure electronic transactions protocol SET, secure socket layer protocol SSL, and security protocol on network layer IPsec. But these protocols have their own advantages and disadvantages. If some of these protocols are combined together, with complementary advantages, the security will be greatly enhanced. Based on this idea, this paper designs the conjunction of AH (Authentication Header) and SSL after analyzing these two protocols and at the same time gives simple analysis on the performance and transport efficiency on conjunction.

Keywords: SSL, AH, Efficiency, Conjunction, Security

1. BRIEF INTRODUCTION TO SSL

SSL protocol known as Secure Sockets Layer protocol, which is between the transport layer and application layer, and is mainly used between the client and server to establish the credibility of a safe data transmission channel in order to ensure the security of Internet.

The layer SSL protocol locates in the TCP / IP protocol stack is shown below in Figure 1.

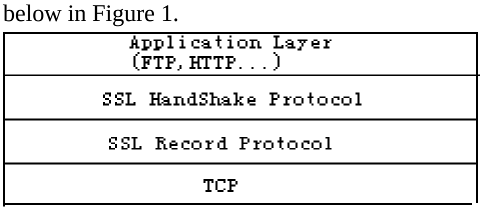


Figure 1. The Layer of SSL

1.1 Security Services

SSL provides such as Connection protection, One-way authentication for servers as well as message integrity and optional client mechanism.

1.2 Methods of Encryption

SSL uses some algorithms for encryption, such as RSA, DES, 3DES, MD5 HASH and so on.

1.3 The Working Process Is as Follows

1)The Initialization of connection process:
Client sends ClientHello messages to the server;
Server sends SeverHello messages to the client;
Server sends the digital certificate containing the RSA public key to the client;
Sever sends SeverHelloDone messages;
Client validates the certificate then if decrypt successfully it will get access to the server public key or else terminate the conversation;
Client generates a random DES key and uses the server public key for encryption and then send it to the server;
Server decrypt with its own private key to get the private key access to the client;

So far, Guest has completed the authentication to server. After this the client and server use the previous session key of DES to communicate safely.

2)The process of customers changing the session key:
Customer sends Change_cipher_Spec message to the server;
Customer sends Finished message to the server;
Server replaces the current key with the key customer sends

3)The end of the process
Client sends a Close_notify warning message and requests to the server for closing the connection.

1.4 The Advantages and Disadvantages of SSL

The advantages of SSL:
1)A simple setting up, low cost and the users need no special installation and making a job as long as the browser will do.
2)To protect the data above the transport layer through encryption.
3)To prevent monitor and attack from the middle-man; the flow data analysis attacks; interception the splicing attacks and re-message attacks.
4)To achieve point-to-point protection.

The shortcomings of SSL:
1)Only can provide a single authentication to server but can not provide a digital signature.
2)Can not guarantee the security outside the transfer process; An attacker can make a data-analysis attack according to analysis the data not protected in the IP message.
3)SSL need to be set up to establish a connection in each communication and in the process of SSL handshake is very time-consuming as well as much cost of server resources and network resources.
4)The session-recovery mechanism has the advantages of performance in frequent conversations but can easily become the vulnerable attacked point of the whole protocol.

2. AH PROTOCOL

IPsec is a security protocol of the IP layer. It mainly provides security services for IPv4 and IPv6 through encryption. It aims to provide safe handling to IP packets and security services for IP protocol and the protocol of upper layer. It includes the control of access, the integrity without connection, and the authentication of data source, anti-replay, and confidentiality of data and limited flow of business. Also IPsec provides the compress of IP datagram. These functions and services are mainly dependent on the cooperation of the AH protocol, ESP and IKE protocol in the IPsec security architecture.

In the IPsec security architecture, AH protocol is the abbreviations of Authentication Header protocol. It mainly provides authentication to information source and data integrity testing for IP data packet. Besides it has the function of anti-multicast. Importantly, the protection mechanism provided by AH is paragraph by paragraph.

2.1 Let Us See Which Layer The Ipsec Security Architecture Is Located In.

HTTP	FTP	SNMP	......
TCP/UDP			
IPsec/IP			

Figure 2. The Layer of IPsec

2.2 The Authentication Mechanism of AH

AH gets the achievement of identity through the process of data transmission on the IP packet with a high level of safety calculated by one-way function. These one-way functions are the core of the AH authentication mechanism, so there must be a high security to guarantee information transmission safety.

2.3 The Authentication Algorithms of AH

AH is defined by RFC2402.It is used to authenticate the data of IP layer with the MAC algorithm in Cryptography. That is Message authentication algorithm which can output a fixed length message named Message Digest accepting a combination of message with any length and a key as input. MAC generally evolved by the HASH algorithm, it first combinates the input message and key and then be an input message of HASH algorithm. We call such MAC Hash Message Authentication Code HMAC.It can be tied to any iterative hash function, but in default it is MD5.

2.4 The Main Principle of MD5 Authentication Algorithm

First, do computing hash to the first plaintext. The computing hash has an initial number of 4 rounds and 16 times computing a round.

Then initial the number of the next round with the result of current round which is a data with 128bit.
Finally, use the output of the final round's 128bit summary information for data authentication.

2.5 The Strengths And Weaknesses of AH Authentication Protocol

The strengths are AH is a two-way authentication. It uses SA-a one-way pipeline to guarantee the reliability of authentication. Only when both sides agreed to verify the SA the authentication would be successful.

The weaknesses of AH are as follows: It does not provide the protection of the business flow analysis. If the attacker observes and analysis the flow patterns of the communications business, that may result in the disclosure of information.

3. THE DESIGN OF THE COMBINATION WITH AH AND SSL

The biggest advantage of SSL is to achieve the security of point-to-point communication, but it only provides one-way authentication to server; it is easy to be attacked through an analysis of IP packets to carry out attacks. Besides it needs hand-shake for key-exchange each communication. Since the authentication algorithm used in SSL is RSA, RSA ensure the safety with the difficulties of the decomposition of large integer.The efficiency of encryption and decryption using RSA is more slowly comparing with the symmetric encryption algorithm. So if there are too many hand-shakes during the communication the speed of communication would be lower. IPsec and SSL can be very good to make up for the security risks posed by, as well as time-consuming and resource consumption. Reasoning from the theory, the two an be combined to perfect each other to further enhance the security of network communications, while using the least resources and speeding the communication and improving the efficiency.

After combination with AH and SSL, the layers of them in the TCP / IP protocol stack are not changed, which is shown below in Figure 3.

HTTP	FTP	SNMP	......
SSLHandShake	SSLChangeCIpherSpec		SSLAlert
SSLRecordProtocol			
TCP			
IP/IPsec			

Figure 3. The layer of combination

3.1 An Improved Program

We put forward the idea that the upper layer provides encryption and authentication and the lower layer is in charge for authentication and signature. That is to say we can use SSL protocol to encryption for data protection which is between the TCP layer and the application layer and in the IP layer IPsec charge for the signature and authentication between the two sides of the communication.

Here we use the transmission mode of AH which is one of the protocols in IPsec security architecture to authenticate the identities of both sides. So we need only one shake-hand with an end-to-end authentication during the whole communication, then we can make up for the shortcomings of SSL which need hand-shake each communication. It can help to improve efficiency especially in that case switching conversation quite frequently.

Improve the SSL handshake protocol and simplify SSL handshake appropriately.

Use the packaging method with UDP packet to solve the problem that AH can not penetrate NAT.

1)In the IPsec security architecture[9], ESP can be used to authenticate and encrypt. If ESP and AH are used alone, no doubt that ESP has higher security than AH. In here we don't use ESP but AH, because ESP protocol is quite more complex than the AH protocol and we just need authentication in IP layer. Let we see the IPsec security architecture changed in Figure 4.

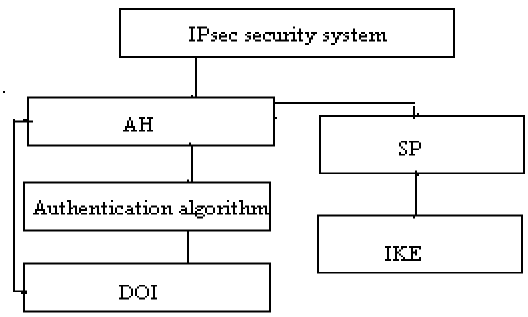


Figure 4. New IPsec Security Architecture

The figure above show the new IPsec security architecture when AH is used alone. We can know that AH and other parts of the system are not changed.

2)The change of SSL handshake process

At the first time, the connection between client and server needs "three-step handshake" as before. But once the connection is established, during the subsequent communications, there is no need to complete another three-step handshake to authenticate to sever. Let we see simple handshake process.

First: Client send Ask message ClientHello to Sever.
Second: Server response to Client with SeverHello message.
Third: If Client wants to change the key of communication, then send ChangeCipherSpec message, after that send Finished message to Sever.
Forth: Sever response to Client with the message ChangeCipherSpec and Finished message.

3)The process of data transmission

The figure below shows how the data of application layer transmit.

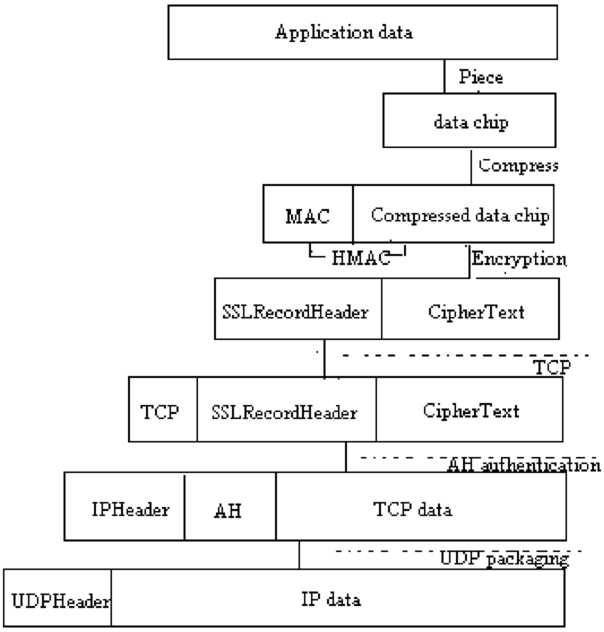


Figure 5. Transmission Process of Data

4)The workflow of SSLprotocol

First step: The complete initialization of SSL. Complete "three-step handshake" process, the establishment of security capabilities, authentication to sever and private key exchange, client identification and key exchange.

Second step: Communications. Divide the application message data into manageable pieces with SSL record protocol, compress the data, encrypt with the private key generated by client, and add header made up of the content type, the main version, minor version and the compressed length, then send the new packet to TCP layer. The process of accepting data is just the opposite.

Third step: Simple initialization. Just need a simple handshake process-a "two-step handshake" but not the complete one, give the responsibilities of authentication to IP layer that can do two-way authentication.

5) The workflow of AH protocol

The workflow of AH protocol is shown below.

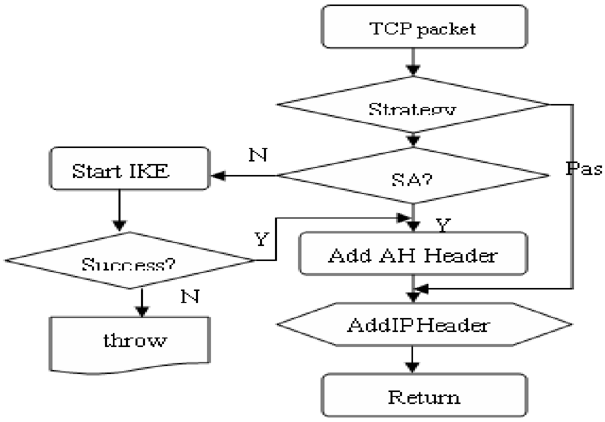


Figure 6. Workflow of AH protocol

6) Package with UDP

Package with UDP is proposed by IETF to help IPsec get through NAT. The idea is that package the IPsec packets with UDP packets before sending it to the host, when arrived, remove the IP header and the UDP package outside. Here we use UDP package to solve the problem getting through the NAT.

3.2 Performance and Security Analysis

3.2.1 Performance Analysis

In the process of unchangeable SSL handshake, which encryption component sever will choose is due to which encryption component the client selects and the performance of SSL are entirely dependent on the selected key exchange algorithm and digital signature algorithm. In the SSL signature algorithm is RSA. RSA is based on the difficulty of large integer decomposition, it has two operations-one with private key and one public key. Obviously the operation with private key needs more time than the operation with public key. But the SSL protocols have to shake hands with each communication, therefore there is a large consumption of resources. Here we let AH protocol and SSL protocol charge for authentication and encryption together with improving the SSL handshake process. After the Improvement, when client and server connect for the first time, SSL needs the three-step handshake and AH needs authentication to both sides of the communication, so the communication speed will drop. But during the subsequent communication, after the initialization is finished, SSL just needs a two-step handshake; AH charge for the authentication alone. Since AH needs one authentication during the whole communications. So this idea can help to reduce a lot of overhead, including server and client.

3.2.2 Security And Feasibility Analysis

1)Analysis from the perspective of network

TCP / IP model is divided into four layers from top to bottom, followed by the application layer, transport layer, network layer and physical layer and they have different functions: the application layer is responsible for raw data input and output, the transport layer (TCP protocol here) is responsible for data transmission, the network layer is for routing, the physical layer is for signal transmission. Here we only analyze the transport layer and the network layer. The reliable connection service of TCP ensures the information sent and received is the same. To some extent, it can prevent copying, inserting, changing the sequence and playback, as well as the destruction of data. So in this layer, encrypting can ensure the confidentiality of data based on the data integrity. For example, the SSL protocol is based on TCP protocol. Generally speaking, SSL has done two things, one is the exchange of keys, and the other is encrypting the data in communications. Therefore, encrypting in TCP layer dose no conflicts with other services or protocols.

The network layer is primarily for routing, that is, to determine a route to destination. If the destination of communication has been modified, certainly the route selected is wrong. Thus, while IP routing, the authentication to destination is more effective. Here we use AH for authentication which is based on the IP address of the public network, so the AH authentication is better than SSL authentication. If the server has been attacked, the attacker disguised as a server to communicate with the clients. During the initialization, we can consider that the initialization of lower layer completes first and then initialize the services top. Therefore, authenticating in network layer and encrypting in TCP layer are consistent with the principle of authenticating first and then encrypting. Also the AH authentication can well solve the problem of data source deceiving.

2)Analysis from the protocol

In this paper, we let AH and SSL work together for the safe communication. Since the AH and SSL are not in the same layer and the IPsec security architecture is transparent to the upper layers, the AH will have no conflicts with SSL.Our goals are reducing the communication delay, accelerating the effectiveness of communications, at the same time improving the effectiveness of authentication by improving the SSL handshake process. The session mechanism also improves the performance by changing the process of shake-hand. Compared with it, the data packet is a little bigger because of the AH header. But the AH header is very small for the current network bandwidth. However, the session mechanism stores the information of clients on the sever, which no doubt add the burden to the server. If there are a lot of clients, the speed of response must be lower and the performance of communications between client and server will also reduce.

3)Analysis from cryptography

During the initialization of communication, the transport layer uses a certificate for authentication and RSA for key-exchange while the IP layer uses AH for authentication. The two authentications are more reliable than SSL.AH uses diffie-Hellman,IKE and MD5 for authentication and these classic encryption algorithms have been proved that to a certain extent it is safe and reliable.

4)Analysis from attacks

Since the SSL protocol is based on a reliable connection-TCP,it can prevent some attacks, such as Man-in-the-middle attacks, traffic analysis attacks and replay attacks. But it also has shortcomings such as business streaming attacks, key exchange algorithm deception, Change-Cipher-Spec discarded attack. These problems of SSL can be solved well after the using of AH.

Preventing business streaming attacks: we can use the tunnel model of AH to protecting the whole IP packet by AH authentication and preventing the attackers analyzing the fields and attributes which are not protected. Preventing key exchange algorithm deception: during the initialization of AH,it needs to establish SA with the authentication to both sides, so the SSL authentication based on AH has no doubt preventing the key exchange algorithm deception. Preventing Change-Cipher-Spec discarded attack: while changing the key or finishing the communication, the client and sever will send Change-Cipher-Spec message to each other. But it is not protected by authentication when SSL is used alone. Here this message is protected by AH when it is sent to the IP layer from TCP layer.

4. CONCLUSIONS

People pay more and more attention to the security of network and the technology of network security has been greatly developed, but the technology on network intrusion has been greatly developed too showing the characteristics of uncertainty, complexity and diversity. As each protocol on network security has some shortcomings, it is difficult to resolve the problems caused by the invasion. If we let different protocols work together, the security of information transmission on network would be enhanced greatly.

REFERENCES

[1] Zhang Xuefeng, Analysis and comparison of several common protocols on network security, Network security and technical applications, 2009.

[2] Zhang Aike, Tan Yaojian, The research and solution on compatibility of IPSec and NAT, Fujian computer, No.4, 2009.

[3] Deng Xiaojun, Huo Xunyu, The Analysis and Research on SSL / TLS Handshake Protocol, Modern computer, No.4, 2008.

[4] Yang Fuhua, The problem on security of the IP network security protocol, Computer learning, No.1, 2009.

[5] Yang Bo, Modern Cryptography, Tsinghua University Press, April 2007.

[6] Schneier, B. Applied Cryptography Protocol, Machinery Industry Press, March 2000.

[7] Xie Xiren, Computer Networks (5th Edition), Publishing House of Electronics Industry, Jan 2008.

[8] Wang Guiren, Bi Hongjun, A application of improving the SSL with Kerberos, Network security and technical applications,No.2,2008.

[9] Ge Dongxia, Jiang Hong, IPsec Security Policy System, Computer Security, No.10, 2007.

Wang Haohao is a master candidate of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology, focuses his study on network security.

A FRACTAL ENCRYPTION ALGORITHM

Hai Peng ,Qingping Guo
1,2 Computer Processing & Distributing Laboratory, Wuhan University of Technology
Wuhan , Hubei 430063, China
Email: percyph@163.com; qpguo@mail.whut.edu.cn

ABSTRACT

In this paper a new method for symmetric encryption algorithm is suggested. The method is based on fractal graphics, and it can be used for future research in the cryptography. Using the multi-branches tree, this algorithm can be easily implemented on the computer.

Keywords: Fractal Theory, Encryption Algorithm, Multi-branches Tree

1. INTRODUCTION

Cryptographic algorithm is the core techniques in cryptography. Efficient, stable and reliable cryptographic algorithm is the key to information security. At present, there are the popular RSA, DES, AES, IDEA, ECC, etc. Some of these algorithms is still reliable, while others have been deciphered. This paper will introduce a new cryptographic algorithm which based on characteristics of the fractal graphics. Fractal graphics' nonlinear feature make the algorithm difficult to be deciphered. And at the same time, the self-similarity also make the algorithm easy to implement.

2. INTRODUCTION TO FRACTAL GEOMETRY

The fractal geometry was set up by Mandelbrot at 1973.Soon it aroused the concern of many disciplines, include biology, geography, materials science, etc. This is because it is very valuable not only in theory, but also in practice. Fractal geometry is now already make a lot of great application in the structure and control of materials, fracture mechanics, natural graphics simulation, the growth of enzymes and many other fields.

Fractal geometry is a new geometry that different from the classical Euclidean geometry. It has a great of advantages in descripting the real nature things. Because many objects like snow, in the nature has the self-similar structure. Quite a number of complex physical phenomena behind the hierarchical structure is reflected in the fractal geometry.

2.1 Famous Fractal Graphics

2.1.1 Cantor Set

Cantor set is a one-dimension space self-similar structure, which introduced by German mathematician Georg Cantor in 1883.

The Cantor set is created by repeatedly deleting the open middle thirds of a set of line segments. One starts by deleting the open middle third (1/3, 2/3) from the interval [0, 1], leaving two line segments: [0, 1/3] ∪ [2/3, 1]. Next, the open middle third of each of these remaining segments is deleted, leaving four line segments: [0, 1/9] ∪ [2/9, 1/3] ∪ [2/3, 7/9] ∪ [8/9, 1]. This process is continued ad infinitum. At last, we get a graphics like figure 2-1 following:

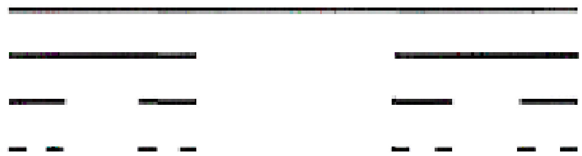


Figure 2-1. Cantor Set

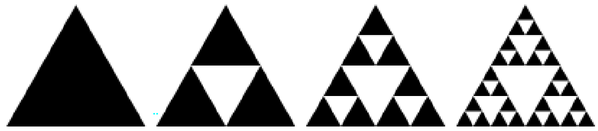
Apparently, Cantor set is a infinite self-similar structure.

2.1.2 Sierpinski Triangle

The Sierpinski triangle, also called the Sierpinski gasket or the Sierpinski Sieve, is a fractal named after the Polish mathematician Wacław Sierpiński who described it in 1915.

Originally constructed as a curve, this is one of the basic examples of self-similar sets, i.e. it is a mathematically generated pattern that can be reproducible at any magnification or reduction.

Start with any triangle in a plane. The canonical Sierpinski triangle uses an equilateral triangle with a base parallel to the horizontal axis,as the figure 2-2-a. Then divide the triangle into 4 identical triangles and remove the middle square as figure 2-2-b. Then divide each of the remaining squares into 4 triangles like before, and remove each middle triangles as figure 2-3-c. Repeat the process. Finally, we get the Sierpinski triangle.

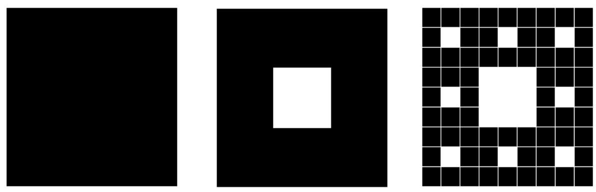


a b c d

Figure 2-2. Sierpinski Triangle

2.1.3 Sierpinski Carpet

The Sierpinski carpet is a plane fractal first described by Wacław Sierpiński in 1916. The carpet is a generalization of the Cantor set to two dimensions It is obtained as follows: Start with a square S . Divide S into 9 identical squares via vertical and horizontal line segments. Remove the middle square. Then divide each of the remaining squares into 9 squares like before, and remove each middle square. Repeat the process. Finally, we get the Sierpinski carpet.



a b c

Figure 2-3. Sierpinski carpet

2.1.4 fractal dimension

Fractal dimension has many kind of definitions. Here is the definition of fractal dimension in this paper.

Take the Sierpinski triangle as a example. When it is a triangle, it has 3 edges. Then it is divided into three triangles, and then every sub-triangles turn to tree triangles…Suppose that at the step i, there are $s = ((m-1)^2 - 1)^i$ triangles. Here we define n=i+1 as the fractal dimension.Generally,

$$n = \frac{\ln s}{\ln((m-1)^2 - 1)} + 1.$$

According to this definition, the fractal dimension of model in figure 2-2-c is 3,in figure 2-2-d is 4.

3. FRACTAL ENCRYPTION ALGORITHM

Fractal graphics have good non-linear features, so it can be Applicated in cryptography. The following steps will demonstrate the algorithm with a Sierpinski triangle. The detail encryption process as followed.

Plaintext: 100010111
Key: 0000 0011 1110
Ciphertext: 100001111

Suppose the plaintext is 100010111, the key is 0000 0011 1110. That the encryption process is following:
First select a Sierpinski triangle (n=3).Then put the each bit of plaintext to the black triangles with the order of the triangle generation(clocksize).

The key include three segment.The beginning segment has four bits which identify the model of fractal graphics. For example, 0000 stand for Cantor set, 0001 stand for Sierpinski ttiangle. And the second segment which also has four bits is identify the dimension of the fractal graphics, here is 0011, it is 3 in decimal. The rest bits of the key are the rotate bits. When put these bits into the white triangles, the triangle turn into figure 3-1. And then rotate the sub-triangles, as the following step.

If the value in the center triangle is 0, the sub-triangles will be rotated clockwise direction once, or if it is 1, the sub-triangles will be rotated twice. The following figure is the rotation process. See figure 3-2 and figure 3-3. Decryption process is reverse. If the value is 0, the sub-triangles will be rotated twice, otherwise, they will be rotated once.

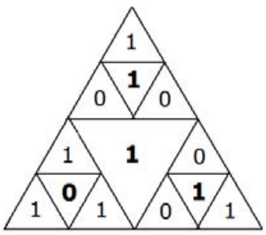


Figure 3-1. Input Data

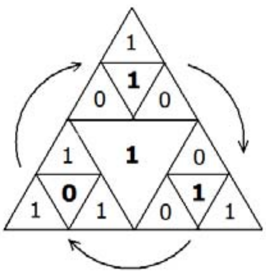


Figure 3-2. Rotate Plaintext

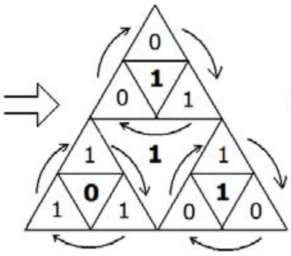


Figure 3-3. Rotate Plaintext

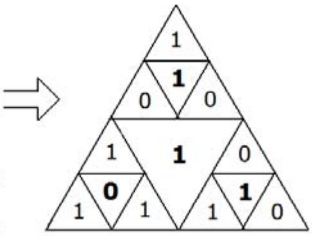


Figure 3-4. Rotate Plaintext

At Last, read the bits from the black triangles as the input order.And that is the ciphertext. It is 100001111. Generally, plaintext is very long, so it need many triangles.

So generally, to all these fractal model,we have the following roles .

1)Translate the key to binary code, and divide it to three segments from left to right. The first segment have four bits, it stand for the fractal model. For example 0000 maybe denote the Cantor set. The second segment stand for the dimension of the fractal graphics. The remain bits is the third segment, it is rotate control code
.

2)Put each bit of the key sequence to the white area(i.e. the holes that have been erased),in accordance with their generating order. For example, if the fractal model is Sierpinski triangle, put the bits in the white triangles, if the fractal model is Sierpinski carpet, put the key bits in the white squares. And the plaintext sequence are put into the black area in a similar way as key sequence. Both of them are put into the model by a clockwise order.

Suppose that the model has been select, and it has m edges when its dimension is 1.For example, Sierpinski carpet has 4 edge, so its m=4,when its dimension is 1,At the same time the Sierpinski triangle has 3 edge, its m=3. Then the model generate a fractal graphics, and its dimension comes to n, its edge increase to m^n. At the same time, the black area in the fractal model is increase to m^{n-1}, and the white area is increase to $\frac{m^{n-1}-1}{m-1}$.You can see that when the Sierpinski carpet has a dimension of 3, the white triangles are 64.

3）Length of the rotate control bits. For Sierpinski triangle, only need one bit to control the sub-graphics rotate. When the value is 0, it rotate 1 round(circle), otherwise rotate 2. But for Sierpinski carpet, one bit is not enough, it at least need 2 bits. Generally, length of the rotate bits is the minimal k, such that $(m-1) \leqslant 2k$ $(k > 0)$[2].

4）Rotation times. Suppose length of the rotate control bits is k, and the rotate control bits is x(in decimal),that the times of rotate is x%m+1 when encryption,is m-x%m-1 when decryption[2].

5) If we put the ciphertext to the input terminal , and change the rotate control role as 4) , we will get the plaintext.

6)This algorithm is a symmetric-key algorithm,it is reversible if the key is known. So this algorithm is not fit for digital signature.

4. IMPLEMENTION

Because of the self-similarity of the fractal graphics, using multi-branches tree to implement this algorithm model is a natural way. For example, we can use ternary tree to implement the Sierpinski triangle. Use the tree node to store the keying bits and the plaintext bits. Leaf nodes have the key bits and the plaintext bits. Non-leaf node only have key bits. The three sub-nodes are have a one - to - one relationship with the three sub-triangles. And the preorder traversal sequence of the tree is just the plaintext sequence and the key sequence.See figure4-1.

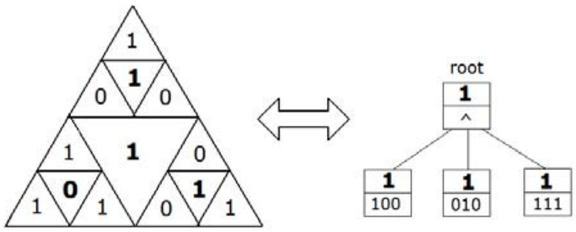


Figure 4-1. Fractal and Ternary Tree

The following is the data structure of the ternary tree.

```
struct TreeNode
{
      struct TreeNode*    blackTriangle[3];
      int                 whiteTriangle[3];
      bool                Isleaf;
      int                 keybit;
} ;
```

For other fractal model we can also define similar structure.

5. ALGORITHM ANALYSIS

The security of this algorithm mainly depends on the following features.

1)The nonlinearity of the fractal. The nolinearity of the fractal make the algorithm difficult to deciphered if do not know the key.

2)Large key space.In this algorithm the key space is very large. For example, if the fractal model is Sierpinski triangle, and its dimension is n, then the key space is 2^x, here x=$(3^{n-1}-1)/2$. If select Sierpinski carpet, the key space will larger. So this algorithm is a strong algorithm to brute force attack.

3) We use multi-branches tree to implement the algorithm. These strengthen the flexibility of the algorithm and make the encryption process more clearly.

This algorithm also has some faults. The biggest one is that it only use permutation, but not use substitution. So some statistic features cannot be hide. And use multi-branches tree to store the sequence also bring some losses of the efficiency.

6. CONCLUSIONS

Nowadays, Fractal geometry has tremendous development. In this paper the new encryption algorithm is based on fractal graphics. It is a symmetric-key algorithm. This algorithm can be used for encrypting data. It is stable and has high efficiency.

REFERENCES

[1] Shamir A. How to share a secret[J]. Communications of the ACM, 1979, pp.612~613.

[2] Wentao Liu, Wensheng Sun, A Cryptographic Algorithm Based on Fractal Theory [J], Information Security, 2009,pp.44~48

[3] Oded Goldreich, Foundations of Cryptography[M]. Beijing: Publish House of Electronics Industry, 2003.1,pp.152~178

[4] Wenbo Mao, Modern Cryptography: Theory and Practice.[M] Beijing: Publish House of Electronics Industry, 2004.5

[5] Karnin ED, Greene JW, Hellman ME. On secret sharing systems[J]. IEEE Transactions on Information Theory, 1983,29, pp.231~241.

[6] Long DY, Wang CJ, Wu D. Encoding and Cryptography[M] BeiJing：Tsinghua University Press，2005.11

Hai Peng is a master candidate of school of computer science and technology. He had received the B.S degree in computer science and technology from Wuhan University of Technology, Wuhan, China, in 2007.Now he goes on further study in his tutor Qingping Guo’s parallel processing and computing lab, and focuses his study on parallel computing and cryptography.

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.

A RISING COMPUTING PATTERN-CLOUD COMPUTING AND SECURITY PROBLEM*

Wang Yufen[1], Guo Xiaojuan[2], Cheng Jifu[3]
[1,2]School of Information Engineering, Henan Technology and Science Institute, Xinxiang, Henan, China, 453003
[3]Library, Henan Technology and Science Institute ,Xinxiang, Henan, China, 453003
Email: [1]wangyufen@hist.edu.cn; [2]gxj1315@126.com

ABSTRACT

From the development history of the computing pattern, it develops from a stand-alone to cluster computing, to grid computing with network development, and to the popular cloud computing. From the concept and characteristics of cloud computing, based on the understanding of cloud computing, the paper introduced the current actuality of cloud computing at home and abroad, pointed out some problems of cloud computing, especially network security, and gave some suggestion methods about network security of cloud computing. Indubitability, in the near future, it will change many aspects of our study, work and life.

Keywords: Cloud Computing, Development Status Quo, Security Problem.

1. INTRODUCTION

In recent years the Internet has made the rapid development and growth. The original Internet system and service design have not solve the following problems: Storage, computer consumption and hardware construction and maintenance costs continue to increase, data centers become went short of space; At the same time, large enterprises must be fully research data resources to support its business practices, to collection and analysis data must be based on a new platform and so on. How to achieve the distributed sharing of resource and computing , how to deal with massive data and services to effectively provide users with convenient and efficient network services, in the context, cloud computing came into being.

2. CLOUD COMPUTING CONCEPT AND CHARACTERISTICS

2.1 Cloud Computing Concept

From the development history, computing model develops from a single calculation to cluster computing, to grid computing with network development, and last to popular cloud computing.

So far, cloud computing has not unitive definition, its App pattern is more variety. Liu Peng defines that cloud computing is an rising model for business computing. Computing tasks will be distributed in resource pool consisting of a large number of computers, so that various applications can access computing power, storage space and a variety of software services as needed. In this paper, the understanding of cloud computing is the development of distributed processing, parallel processing and grid computing, or the business realization of computer science concept. Cloud computing is result of the mixed evolution and risen of the concept which is virtualization, utility computing, IaaS(infrastructure as a service), PaaS (platform as a service)and SaaS (software as a service). It is a new way to share infrastructure. It faces ultra-large-scale distributed environment, its core is to provide data storage and network services.

The basic principles of cloud computing is that user applications do not need to run on the terminal equipment such as personal computers and mobile phones , but to move in the Internet large-scale server clusters. The data which users deal with do not store in the local, but store in the Internet data center. The corporations which provide cloud computing services are in charge of management and maintenance of normal operation of data center, to ensure them have enough computing power and storage space for users. At any time and any place, users can connect to any Internet device.

Kai-fu Lee who is Google global vice president and the president of China compares cloud computing to bank. He said that the people put the money under the pillow at first, after they can save money in banks, it is very security but inconvenient to exchange. Now the bank has developed to inter-bank system, the people can take money in anyone network, or in the ATM machine. Just like the thing, every family not to equip generator, power is bought directly from the power company. Cloud computing is a late-model network App pattern or the changing.

Cloud is a computer groups, each one includes hundreds of thousands of computers, even millions of. The benefit of cloud is that the computer can update at any moment to ensure the cloud longevity. IBM, Microsoft, Google and other IT industry leaders have made an example for cloud computing application. Although some of the applications are only just beginning, but in the future, fast and convenient web applications can be expected.

2.2 Cloud Computing Characteristics

1) Huge scale. Cloud has considerable scale, Google cloud computing has had more than 1 million servers, Amazon, IBM, Microsoft and Yahoo have had hundreds of thousands of ones.
2) Virtualization .Cloud computing supports users to use a variety of terminal at any location for application services. The requested resources are from the cloud, rather than tangible fixed entity. The application runs in the clouds somewhere, but in fact users don’t know or worry about the specific location of running applications. Only with a laptop or a cell phone users can achieve their need through the network services, even including such tasks of supercomputing.
3) High reliability. Using cloud measures that the data multi-copy fault-tolerant and isomorphism computing nodes interchangeable to ensure the service reliability. The use of cloud computing is more reliable than the local computer.
4) Versatility. Cloud computing is not aimed at specific

* project origin: “The 11th five-year plan” layout topic of education science worker in Henan Province, ID 2008- jkghagh- 286.

applications, can be constructed a variety of applications in the cloud sustain, one and the same cloud supports different applications to run at one time.

5) High expansibility .The cloud scale can dynamically flex to meet the need of the growth of applications and users.
6) On-demand service. Cloud is a huge pool of resources, users buy on-demand; Cloud can be charged like running water, electricity and gas.
7) Extremely cheap. As the special fault-tolerant measures, cloud can be constituted by extremely cheap nodes, as the automated centralized management of cloud, so a large number of enterprises don't burden the increasingly high cost of management of data centers, and the versatility of the utilization of cloud make using resources increase more than the traditional system , so that users can fully enjoy the advantages of low-cost cloud.

3. CLOUD COMPUTING DEVELOPMENT STATUS

At present, cloud computing providers can be divided into two categories: Cloud computing basis platform providers and service providers. The first category mainly provides the bottom technologies including hardware platforms, virtual integration, system management and task scheduling and so on, such as IBM, SUN, and Langchao; the second provides the application services, for instance Google provides services of search engine and Google Maps, Amazon provides EC2 which is the flexibility cloud computing platform, as well as Windows Live network storage services.

3.1 Cloud Computing Development Status at Abroad

Google should be the largest users of cloud computing. Google search engine establish on servers which distribute in more than 200 locations and have more than one million , the number of these facilities are the rapid growth, so do the Google earth, Maps, Gmail and Docs. Adopting the Google Docs application, user data will be stored in a location on the Internet, and accessed conveniently through any Internet-connected system. Currently, Google has allowed the third party to run large-scale parallel applications in the Google cloud computing through the Google App Engine. Google that is not conservative is laudable. It has long open up its three magic weapons of cloud computing by publishing academic papers: GFS, MapReduce and BigTable, and set up the courses how to do cloud computing programming in America and China universities.

Amazon provides corporation with computing and storage services using the Elastic Cloud (EC2) and Simple Storage Service (S3). Service charges include storage servers, bandwidth, CPU resources and monthly fees. Monthly fees are similar to monthly telephone rental costs, the according to the capacity, storage server charges and bandwidth to charge, CPU charges based on the time length of computing volume. Amazon has spent less than two years to create a big business- cloud computing. Amazon's registered developers add up to 44 million peoples, and there are a large number of enterprise-class users. Data provided by third-party statistical agencies showed business revenue of Amazon with cloud computing has reached 100 million dollars. Amazon cloud computing is one of the fastest-growing businesses.

IBM launched the blue cloud computing platform on which customers can buy and use in November 2007. It includes a series of automated, self-management and self-repair cloud computing virtualization software, so that applications from the world have access to large-scale distributed server pool and data center runs the calculation in similar Internet environment. IBM is working with 17 European organizations to carry out Cloud computing project, the EU provided 170 million euros as part of the funds, the plan, called RESERVOIR, it's slogan is a "barrier-free resources and service virtualization". August 2008, IBM invested about 400 million dollars to transform cloud computing data center located its North Carolina and Tokyo in Japan. IBM plans to invest 300 million U.S. dollars in 2009 in 10 countries, so that builds 13 cloud computing center.

Keeping up the pace of cloud computing, in October 2008 Microsoft launched the Windows Azure operating system. After Windows replaced DOS, Azure is another disruptive transition- through creating a new Cloud computing platform on the Internet, it lets Windows truly extend from the PC to the "Blue Sky". Microsoft has the world's hundreds of millions Windows users desktop and browser, and now it will connect them to "Blue Sky". The bottom of Azure is the Microsoft global system of basic services, which consist of fourth-generation data centers all over the world.

3.2 Cloud Computing Development Status at Home

In China, the development of cloud computing is also very rapid. May 10, 2008, the China first cloud computing center began to operation, this center was established in Wuxi Taihu New Town Science and Education Industrial Park by IBM. June 24, 2008, IBM set up a second cloud computing center in the Beijing Innovation Center - IBM Greater China Cloud Computing Center; November 28, 2008, Guangdong Electronics Industry Institute and the Dongguan Science and Technology Industry Park Administrative Committee make a contract that Guangdong Institute of Electronics Industry establish cloud computing platform invested 200 million yuan in Dongguan Songshanhu; December 30, 2008, Ali software company which is a Alibaba Group's subsidiary signed 2009 a framework agreement on strategic cooperation with Jiangsu Province government in Nanjing, which is to establish China's first "Cloud Computing Center e-commerce" in Nanjing early in 2009, the first period investment is up to more hundred million yuan; 21ViaNet introduced CloudEx product line, including a complete Internet hosting service "CloudEx Computing Service", based on the online storage virtualization "CloudEx Storage Service" and provided data backup services for personal and corporate in Internet Cloud, etc; China Mobile Research Institute start early in the cloud computing exploration, has completed the test of cloud computing center. China Mobile chairman and CEO Wang Jianzhou thinks that cloud computing and mobile Internet is the future direction of development.

The concept of security of the cloud which China's enterprises create is unique in the international cloud computing field. Cloud security is to monitor anomaly software behavior on the network by a large number of network clients, and to access to the Internet up-to-date information about Trojans and malicious programs, and to push automaticly to the server for analysis and processing, and then distribute viruses and Trojan solutions to each client. The strategy of cloud security is: the more users, the more safe each user will be, because such a large user base, enough to cover every corner of the Internet, as long as a site to be linked to the Trojan horse or a new virus appears, would be intercepted once. Cloud security development is like a gust of wind, Rising, Trend Micro, Kaspersky, MCAFEE, SYMANTEC, JiangMin, PANDA, Kingsoft, 360, iKaka have launched cloud security solutions. Based on cloud-security policy, rising develops 2009 new product, which intercept millions times Trojan horse attack every day, on January 8 which is reached more than 765 times. Trend Micro cloud security has established 5 data centers in the global and the tens of thousands of online servers. It is learned that the cloud security support 5.5 billion times to click inquiries average daily, collect and analyze 250 million samples, database reach 99% hit rate at the first time. With cloud security, Trend Micro is now interdiction the virus infection reach to 10 million times every day.

4. CLOUD COMPUTING PROBLEMS

Despite cloud computing development is rapid, however, is still in its infancy, there is a wide range of issues, this article summarized some problem: network security, shortage of talent and uniform standard. And the network security is most important problem to pay attention for the people. The following network security problems in the cloud computing are introduced , and some suggestions solution are adopted.

4.1 Network Security of Cloud Computing and the Solutions

At present, there are a lot of risks and problems which cloud computing can not avoid, such as safety and confidentiality, access rights risk, as well as privacy and reliability storing remote data.

Cloud computing based on existing distributed network, network computing is considered a node. Once the computer connected to the Internet may become a part of cloud computing. The more and more data that contains personal information are published by the government agencies and public services, such as medical data, voter information and so on. If there is no credible privacy, then the attacker will take advantage of multiple data links between the individuals to obtain private information.

If the users put their data and procedures on other people's hardware, it will be out of control for some sensitive information. For example, an investment bank staff organizes and manages staff social security numbers by Google Spread sheets. In that case, Google have the responsibility protecting this information away from hackers and data leakage, rather than the bank. Looking at the recent security incidents, these thing increase more, such as the network of criminals attack on and data lose, which also brought some problems to the cloud computing. Such as, in 2007, tradesman TJX lost 45 million credit card numbers, the British government lost 25 million taxpayers records.

4.2 Cloud Computing Network Security Solutions

Although with some anxiety, but the cloud computing has the development potential extremely, it may help the enterprise to reduce the IT cost largely, and obviously raise the working efficiency and the flexibility. Once the field found some perfect safety safeguard solution, the cloud computing could obtain the large-scale promotion.
Regarding its security, there are also many solutions.

1) To encrypt the store documents
The encryption technology may carry on the encryption to the document, such only the password can decipher. The encryption enables you to protect the data, even if the data uploaded to others data central which had the distant place .PGP or the correspondence opens source product TrueCrypt that provide enough formidable encryption function: if you use the password which is unable to explain, then nobody can obtain your sensitive information.

2) Pair of email carries on encrypts
PGP and TrueCrypt can encrypt your email before leaving your control area and plays the protective function. But, this will be dangerous for your email, because it still could be the form which is visited by the peeper, and to arrive at your received box. In order to ensure the mail security, might use procedure like Hushmail or the Nutemail and so on, both can use on-line, and encrypt all your mails which receives and dispatches to you automatically.

3) Using prestige good service
Even if you have carried on the encryption to the document, some online activities (involve particularly in on-line processing file, but was not the activity to store documents) are still to protect. This meaning that users were still to consider which service can be used. The experts suggested that uses the fame service, they will not take their name brand to take risk, so the data divulging a secret event will not be occurred, also not shared data to the marketing businessman.

4) Use filter
The companies of Vontu Websense and Vericept provide one kind of system, which the goal is to monitor which data leave from the network, thus prevented the sensitive data automatically. For example, the social insurance number has the unique way to arrange. And the kind of system may be deployed, so that the different user in the company can enjoy the aspect of derived data with the varying degree freedom.

5. CONCLUSIONS

Although cloud computing faced with various problems, but the pace of progress is unstoppable, because it has bright prospects for development: Firstly, cloud computing have stronger computing power, but users do not need to purchase software and to install and maintain it, cloud computing is distributed sharing, and it will change the existing mode of operation of the business; Secondly, cloud computing can reduce the large amount of energy consumption, will make earth healthier; Thirdly, the information is stored in the cloud, will never be lost; Finally, as an educator, cloud computing will promote personal learning environment and the development of informal learning ,reduce education information hardware and software costs, share data in the maximum extent, and avoid the duplication development of education information system in the field of education.

In the near future, cloud computing will soon become the mainstream technology of information life, and show great vitality, and change our study, work and life.

REFERENCES

[1] Liu Peng. Cloud computing definition and characteristic, http://www.chinacloud.cn/show.aspx?id=741&cid=17
[2] Chen Yaying. The future technology trends: Cloud computing, http://www.topoint.com.cn/html/cloud/cloudcom/2009/04/229875.html.
[3] Lin Ye. The future computer computing technology and development, http://tech.it168.com/a2009/0403/270/000000270782_2.shtml.
[4] Liu Peng. Development of Cloud Computing, http://www.cnw.com.cn/server-cloud/htm2009/20090403_171367.shtml.
[5] Zhang Jian. The concept of cloud computing and influence analysis, Telecommunications Network Technology, No.1, pp 15-18, 2009.
[6] Sun Yanzhen. New network application model - of Cloud Computing, Computer Knowledge and Technology,pp126-127
[7] Gao Lanlan. Depth comparative study for cloud computing and grid computing, HAI XIA Science, vol.26, no.2 2009, pp56-57.
[8] Wang Ping, Zhang Jiping. Cloud computing and e-learning, Modern Educational Technology Vo1.18, No.11 2008,pp 81-84
[9] Hu Wei, Zhong Weilian. Cloud Computing on the issue of network security, pp109-111

Wang Yufen (1965-), female, Wenxian, Henan Province, senior experimentalist, engaged mainly in education of computers, wangyufen@hist.edu.cn.

APPLICATION RESEARCH OF IMAGE-BASEDAUTOMATICALLY FOOTPRINT RECOGNITION SYSTEM

Yi Gao
Criminal Technology Department, China Criminal Police College
Shenyang, Liaoning, China
Email: gao_yi7666@sina.com

ABSTRACT

The key to identify footprint is the exact extraction of the characteristics of footprints and the correct identification if footprints images. The system will square the footprints in the database with the footsteps in line with the conditions one by one to calculate the characteristic parameters of the candidate tracks, and then draw the similarity of the footprints on the basis of the accurate entry and retrieval from their characteristic parameters. So we can overcome the two-definitional issues owing to different descriptions from different officers involved and focus on the technical problems of vague on-the-spot footprints enquiry. This paper presents an image-based footprints identification method. Application results show that the image-based automatic footprint identification system has the advantages of fast computing, good effect of feature extraction and efficient recognition.

Keywords: Footprint, Feature Extration, Image-based, Automatic Iidentification.

1. INTRODUCTION

With the development of computer and the technology of figure processing, people may make use of the theory of computer and image processing to identify and analyze footprints. Footprint identification is one of the main technical method for criminal detection. Therefore, how to identify footprint fast accurately and scientifically so as to improve detection efficiently is not only the urgent need to address the issue of footprint identification, but also a new topic in criminal technical development and in practice. Footprint features include local and overall characteristics.

These features can be used to test barefoot footprints, personal analysis and identification, besides it can also be used for shoes, socks footprint test, identity analysis, identification of footwear and identity. In computer footprint identification system, whether we can obtain a complete footprint figure is the basis for feature extraction and the key to identifying the correct figure. Track automatic identification system, based on the platform of information network and figure is mainly applied in investigation technology departments of public security, in the two models in the Council's detachment and the sub-group for footprint application and management, and in support for PUC inquiries between different places. In such a way, we can make" case investigation for case" and "identity for person" by adopting reasonable classification of the sole pattern, combining with the text information of relevant cases, making use of computer for searching on-site footprint in the footprint library, inquiring via the web in different places and comparing on-the-spot footprints with those in footprint database. Main functions are: footprints capture, figure processing, graphics marking, analyzing, input processes, footprint checking, track combined, graphic statistics, user authorization and message tips. This system is different from the previous one which makes manual pattern classification for sole pattern.

The new system, through a simple and convenient figure processing, automatically extract each footprint figure as the index for search query, overcome the overcome the two-definitional issues owing to different descriptions from different officers involved, and solve technical problems of vague footprints on the spot.

2. THE FOOTPRINT INPUT AND THE CATCH CHARACTERISTIC

Footprint input and the characteristic capture are the systematic basic function, the system quote TWATN to develop package's transfer connection, realizing the scanner figure input, it also supports so many ways ,such as digital camera, floppy disk memory, then carries on the trail to separate the background color, the shoes before receiving the mark object surface contact the pattern to take the scenery (white), and making other parts for the background color (black), and carries on to the footprint figure describes or processing, and carries on the background processing after the system program to withdraw the effective characteristic information. and then input database, as shown in Figure 2.1. On the footprint figure demonstrated that footprint district template, the intermediate region for the primitive footprint image, left side upside divides into the system automatic reduction this trail coverage characteristic figure, left side lower part divides into this trail the related text information, right flank the part to carry on each kind of function contact surface which figure characteristic processing and the extraction need.

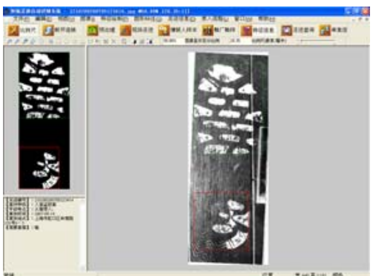


Figure 2.1. Element Distance

Feature extraction defined 3 key points in the footprint: Tip of the toe spot's on peripheral point A, heel spot's under peripheral point B, sole level line's inside vertex C, as shown in Figure 2.2. The complete footprint, each spot must make the corresponding adjustment. After the footprint picture has chosen 3 footprint key points previous time，the footprint system according to the AB two spots (trail middle line) will carry the revolving adjustment on the picture, according to C spot and the line segment AB computation footprint extreme breadth, and gives the default reduction scope. Obviously the ABC three spot selections will complete the footprint correction for direction, the reduction to adjust and to locate the footprint pattern module.

Figure 2.2. Feature extraction interface

Revolving threshold value: ±5. According to the view coordinate system, direction of X axis is from left to right, top-down is the direction of Y axis. A spot in place above, B spot in underneath, therefore the revolving goal causes the angle of line segment AB for the approximate - 90 degrees (-90±5).

You may click on “the characteristic warehousing” the button, When the result is up to the request, the characteristic has already input the database, completed the footprint characteristic input work. when the dialog box presents prompt “to preserve successfully”.

3. FIGURE PROCESSING AND GRAPHICS MARKING

The system has various kinds of figure processing functions and can effectively improve the effectiveness of the footsteps pictures at spots. It mainly provides changes and quality refinement to footstep pictures.

The functions include: luminance contrast, gray variation, reserve color, entire uniform, outburst details, removing noise, subtracting figure, mirror figure and so on. The functions are designed to increase the effective use of footsteps. Aimed at the common reasons which decline the quality of footstep picture, the system choose different figure processing method and get good result in practice. Figure 3.1 shows the figure before processing and after processing. The system can also mark the characteristic information of figure and make further explanation and measure contrast. Figure 3.2 .

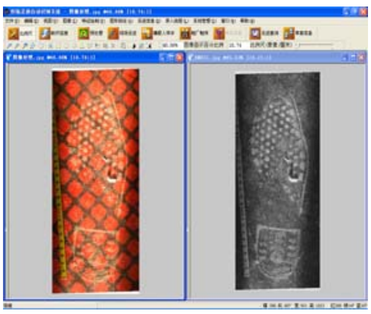


Figure 3.1. The comparitation picture before and after figure processing

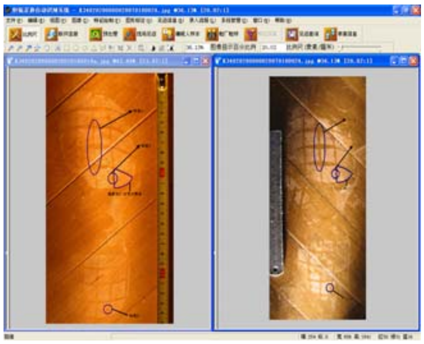


Figure 3.2. Mark the characteristic information of the figure

4. INQUIRY AND CONTRAST OF FOOTSTEPS

Calculate the area of foot on the sole before inquiry. The pattern similarity statistic at heel area A_1,A_2,if the footstep is not complete，the statistic is A_3 the sole area pattern similarity statistic formula:

$$A_i = \sum_{\Omega} \sum_{i \in \Phi} \sigma_i \delta_i \tag{1}$$

$\delta = 1$,including pattern; $\delta = 0$,not including pattern.

A_i： i the pattern similarity statistic within area

Ω_i：the stoical area of soles；i=1，sole area； i=2，heel area；i=3， aim at incomplete footstep.

σ： pattern weight coefficient, calculating method：

σ_i =the ratio efficient of the pattern / the sum of pattern ratio efficient.

In the system, the recall rate and precision rate are over 90 percent. Computer user could use user-defined conditional text to implement multiple component search. Footprint figure is the focus whatever text query or mark pattern search, even if the footmark is no longer existed or the relating information fits the search qualification, it can not be lifted up as inquiry result to reinforce the important degree of footprint figure. Figure 4.1, 4.2

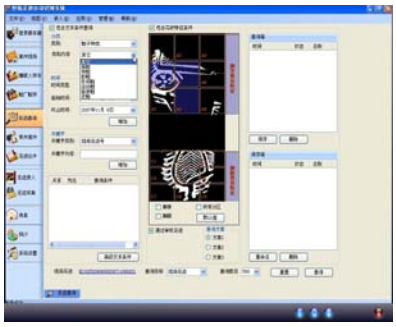


Figure 4.1. Footprint search interface

The system sustains secondary search and starting over which users implement continuous query based on last searching result to reduce your searching range gradually and make the result more precision. Moreover, it is convenient to check and compare the one-to-one resulting footmarks. Query result Figure 4.3.

Figure 4.2. Search result interface

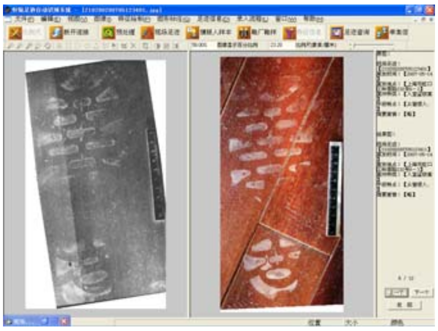


Figure 4.3. The second inquiry results

5. JOINT-FOOTPRINT AND STATISTICS

The system provides “case scene” joint-function according to the basic footmark searching and footmark information and assistant text information. It reflects the footprint system operating condition intuitionally. Figure 5.1,5.2.

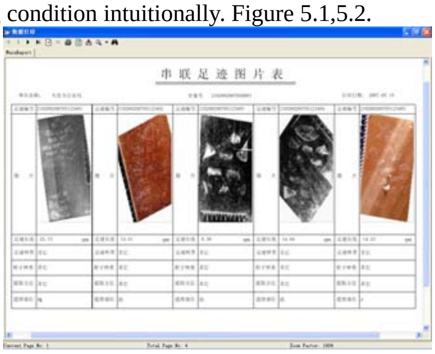


Figure 5.1. Joint-footprint interface

Figure 5.2. Statistics interface

6. MESSAGE MANAGEMENT AND WEB BROWSE PROCEDURE

To make easy contact and communication between city police bureau and local police division, the system provides swift and efficient photograph download and convenient web browse under the condition of low network flux. It also has the function of exchange visits of technical personnel between different cities to play the most value of suspect footprint reverse check and joint-footprint cases. Figure 6.1,6.2.

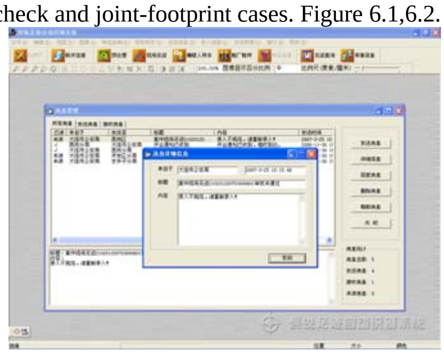


Figure 6.1. Message management interface

Figure 6.2. Photograph browse interface

7. SYSTEM APPLICATION AND SCHEDULE

The system is the network edition which uses investigational division of public security department of the province or local city as constructional unit. It makes use of existing public security network as hardware basic and takes advantage of the data-base server in the investigational division of public security department of the province or local city. Moreover, the system sets up the client in the technology development division of the county public security division. After finishing a series of record, query, compare and joint work, police station officers could uncover the suspect target by footprint sample instrument and management of suspect sample combining reverse query of B/S pattern from the existing footmark database.

REFERENCES

[1] Yang Shu, Shi Li-Min, Gao Li-qun. The the automatic extraction algorithm of toe shape feature. Chinese People's Public Security University (Natural Science Edition), 2005, (03).

[2] Ao qi. The application of the footprint test in computer[J]. People's Public Security University Journal, 2000, 1:29-32.

[3] Zheng Yongan. the weighted self-adaptive image fusion algorithm of the regional characteristics based on wavelet transformation [J]. Computer Engineering and Applications ,2004,6:94-96.

[4] Wang Xin-huai, GAO Shu-Hui, Ban Maosun, Guo Wei. Footwear records management and computer-assisted retrieval system [J]. Police technology, 2001, (05).

[5]Deskiewicz,K.J;schallamach pattern on shoe outsole acknowledged by court in footwear identification Schallmach;Journal of Forensic Identification.-2000,50(1)1-4.

[6]Yu. P. Pattern recognition based on morphological shape analysis and neural networks [J]. Mathematics and Computers in Simulation, 1996, (40): 577~595.
[7]Geradts,Z.;Bijhold,J;new developments in forensic image processing andu pattem recognition;Journal of the Forensic Science Society2001,41(3) 159-166
[8]Zhang Qiang; GAO Li-qun; Guo; barefoot plantar and automatic extraction and self-adaptive feature recognition model [J]. Criminal technology, 2008, (01).

Gao Yi, male, (1978－), the Han nationality, Master, Criminal Technology Department, China Criminal Police College, Lecturer, majoring in teaching and researching of criminal technology.
Address: Criminal Technology Department, China Criminal Police College, No.83 Tawan Street, Huanggu District, Shenyang, Liaoning
Postal Address：110035
Email：GAO_YI7666@SINA.COM

COOPERATION OF FIREWALL AND INTRUSION DETECTION SYSTEM BASED ON C\S MODEL

Zhang nanping[1], Tan wei[1], Guotao Qi[2]
[1]College of Computer Science&Technology, Wuhan University of Technology
Wuhan, Hubei 430070, China
[2]The College of Henan Procuratorial Profession
Email:eric_0206@163.com

ABSTRACT

With the continuous development of the network system, network security gets more and more attention. As a special equipment which strengthens access control, firewall prevent the unlawful entry of external users, protects the internal network environment, and thus become the mainstream of security products. Intrusion Detection System (IDS) collects information from the system and network initiatively, and analyze them to find out whether current system has been invaded or attacked, which make IDS being more and more popular. Based on the analysis of firewall and intrusion detection technical features, this paper presents an issue which makes firewall and IDS work together effectively to make up for their deficiencies, and finally improve the network security.

Keywords: Network, Security, Firewall, Intrusion detection, Cooperation.

1. INTRODUCTION

With the continuous development of the network system, people get much convenience of daily life and work. Meanwhile, our network system creates so many opportunities to hackers [1]. As a result, a variety of network attacks continue to occur. Being one of the most effective tools of defending attacks from hackers, firewall becomes the mainstream of security products. However, due to the firewall itself, there are many shortcomings in firewall with passive defense mechanism. Intrusion Detection technology is a detection technique which developed rapidly. It monitors network invasion and anomaly, and is considered to be the second gateway firewall. Yet Intrusion Detection System (IDS) cannot do very well in resisting network attacks because its lack of measures to those detected attacks. Therefore, it is becoming a hot topic of current research that how to combine firewall and IDS to make up for each.

2. FIREWALL AND INTRUSION DETECTION

2.1 Firewall Overview

A firewall is a piece of software or hardware that helps screen out hackers, virus, and worms that try to reach Intranet. It makes a protective structure between the internal network and external networks, private networks and public network. Thus a firewall is a communication channel, which in accordance with enterprise security policy to control (allow, refuse, monitor, record) the accessing network behavior.

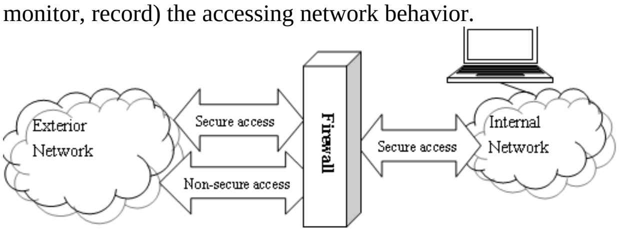


Figure 1. Firewall filter data flow

As shown in Figure 1, Firewall filters the network information flow coming from internet. Only data stream which is consistent with firewall security policy can be allowed to pass through. Therefore, the typical firewall has the following three basic characteristics:

1) Any data stream between internal network and the exterior network must pass through the firewall.
2) Only those data stream which conforms to security policy can pass firewall.
3) The firewall itself should have the very strong anti-attack immunity.

At present, the firewall technology in the access control functions have been more perfect, to a certain extent, guarantee the security of network access. On the other hand, the firewall still has some insufficiencies, mainly has the following several aspects:

1）Firewall cannot defend the access which is authorized or already exist in the network internal attack.
2）Firewall cannot repair the frail management measure and the problematical security policy.
3）Firewall cannot defend those attack and threat which do not undergo firewall.

2.2 Intrusion Detection Overview

According to the current definition given by United States Agency for International Computer Security Association, Intrusion Detection checks whether there is any breach of network security strategy and the phenomenon of attacks, by collecting and analyzing information coming from the computer network and some key points in computer system.[2]

Intrusion Detection System (IDS) is composed of the software and the hardware. As shown in figure 2, based on characteristic description stored in characteristic description storehouse, IDS detect network intrusion and anomaly by analyzing network packets and system log[3]. Therefore, a basic IDS needs to solve two aspect problems: First, how to extract the full and reliable description of behavioral characteristics of the data; second, how to determine the nature of a behavior efficiently and accurately, according to the existing characteristics.

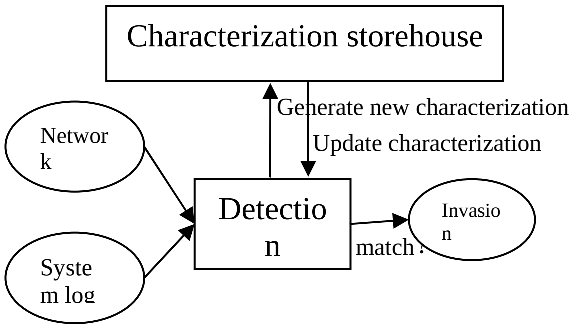


Figure 2. Intrusion Detection Model

As an initiative safety protection technology, Intrusion Detection technology provides the real-time protection against to internal attack, exterior attack and incorrect manipulation [5]. After more than 20 years research and the development, the intrusion detection technology has already obtained the rapid development, but at present it also has many insufficiencies, mainly being the following several aspect questions:

1) The main existing Intrusion Detection System only monitors the network behavior passively, even if had discovered the invasion, itself also lacks the effective active defense measure.
2) The existing intrusion detection technology in aspects of the rate of false alarm, rate of missing report and examination accuracy cannot entirely as desired.
3) The examination efficiency is low, and the latency is big, which cannot meet the increasing bandwidth requirements of the testing.
4) Lacking of deep level analysis to warning, the existing intrusion detection system produce a huge volume of the alarm.

3. COOPERATION OF FIREWALL AND INTRUTION DETECTION SYSTEM

The firewall and the intrusion detection have separately represented the static and dynamic way in tradition network security technology. How to make “static” and “dynamic” work together has become a new development direction of network security recently.

3.1 Design Objective of Cooperation

In accordance with the purpose and technical requirements of cooperation, cooperative system needs to have the following basic characteristics:

1）Validity
In view of the concrete invasion behavior, the response measure adopted by system should be able to prevent the invasion effectively, which is the main goal of cooperation.
2）Timeliness
The system should be able to take the effective response measure promptly.
3）Rationality
The response should be received in exchange for the biggest security goal by the minimum price. Any response is unnecessary when the price is bigger than the loss due to the continual attack.
4）Security
On protecting network and computer from attack, IDS’s own security is the most basic request.

3.2 System Design

3.2.1 Cooperative Ways Outline

Currently, there are two ways of realizing cooperation-tight integration and interface opening.

Tight integration means embedding the IDS to firewall. Intrusion detection data is no longer the original source of network packets, but the data flow through the firewall. All packets passing through the firewall not only must accept firewall's examination rule validation, but must accept the IDS's inspection. Actually, tight integration synthesizes two systems together. However, because IDS itself is a very huge system, both implementation difficulty and overall performance can receive the very tremendous influence.
Interface-opening is another way of realizing cooperation. The firewall or the IDS opens an interface for opposite part. Both sides communicate with each other by using a certain protocol. This way is quite flexible, which does not affect the firewall and the IDS respective performance.

3.2.2 System Design

In consideration of performance and complexity, this article chooses interface-opening to realize cooperation. And in this foundation, this article proposes cooperation architecture based on C\S pattern. In this architecture, firewall runs as a server and IDS as a client. Both sides communicate with each other basing on “request/response” pattern.

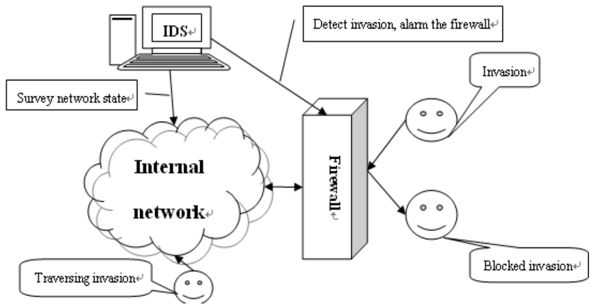


Figure 3. network structure of cooperation

As indicated in figure 3, firewall filters a part of invasion attack through the security policy which establishes in advance, while the other invasion enter internal network. When IDS detects these traversing invasions, it alarms the firewall and transmits related information immediately. After receiving alarm, firewall adopts response measure immediately, such as cut off the connection. And then, firewall adjust the security policy according to the security event, preventing this invasion event occurs again.

This kind of C\S structure makes firewall and IDS unchanged in network location and system structure, which maximize the functional characteristics of their respective advantages. On the other hand，because of C\S pattern mechanism, only when IDS needs firewall handling security event, IDS sends out connection request. And firewall response after handling event. This mode saves valuable network resources, and also reduces the cost which linkage takes.

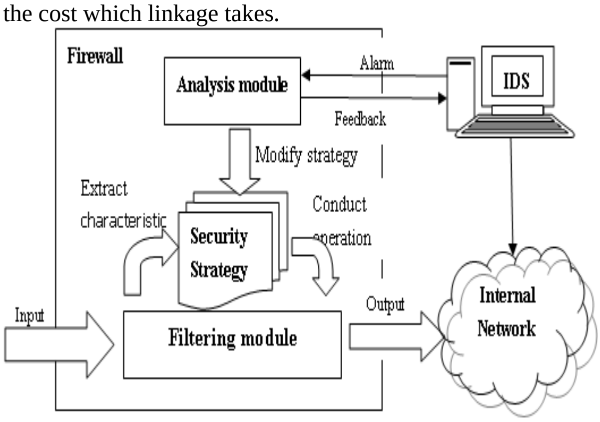


Figure 4. The cooperation model

3.2.3 Security Analysis Module

In order to reduce coupling degree, we choose to embed a security analysis module (SAM) into firewall, which is indicated in figure 4. The responsibility of SAM is as below:

1) In accordance with the agreed protocol, SAM communicates with IDS to provide service to IDS.
2) SAM interacts with original module of firewall. By analyzing security event, SAM detects and fixes bug of security strategy, guiding implementation of access

control.

When IDS detects network invasion or anomaly, it alarm SAM immediately, and transmit security event information. After receiving related information, SAM analyzes the reported event and response to it, and then modifies security strategy accordingly. On basis of modified security strategy, firewall renews its own control behavior, cutting off the invasion connection and thus stopping the further development of attack.

By embedding SAM, firewall makes the original module unchanged, which greatly reduces the complexity for firewall to achieve linkage with IDS. In addition, as we know that firewall access control implementation is based on security strategies that it followed. By analyzing security event and then modifying strategies, SAM enhanced the firewall capability of preventing internal network from attack. This makes up the vulnerability that firewall cannot modify problematical security strategy, increasing firewall’s adaptive ability.

After handling security event, SAM sends feedback, including the handling result, to IDS, which makes IDS enabled to track the event and record logs. More important, if IDS discovered that this security event's processing exist problem, it can seek the higher level way of solution by alerting to the administrator console.

3.3 Information Exchange of Cooperation

In order to exchange data orderly, it is necessary for each part of system to follow some pre-agreed protocol. This protocol defines the sequence of events as well as data exchange format clearly. Figure 5 has indicated the whole process of event handling. On discovering invasion, IDS send alarming request to SAM as a role of client. After receiving alarming request, SAM sends back response and prepare for receiving related information. IDS transmit security event information and wait for feedback subsequently. After analyzing event and then taking steps, SAM sends feedback to IDS finally.

On the aspect of exchange data types, there are mainly 4 types for exchange:

1) heartbeat：To keep IDS and firewall contact normally, IDS send a heartbeat request to firewall periodically. After receiving heartbeat request, firewall send back a heartbeat response immediately, indicating that the connection is normal.
2) Alarming request：When IDS detect network invasion or anomaly, it alarm the firewall by send alarming request immediately. After receiving alarm, firewall send back response and prepare for receiving security event information.
3) Security event information：This is the information need to be analyzed by SAM. Security event contained in a variety of information, and the general types of events including security alarming, security alarm event identification, attacking time, attacking type, attack description, the attacked address, port, protocol, event description and so on.
4) Feedback：After analyzing event and then taking steps, SAM sends feedback to IDS. The general feedback includes security alarm event identification, security event handling type, response type, event handling result, handling description and so on.

3.4 Security of Cooperation System

Regarding cooperation system，its own security must be paid a special attention. Once intruder discovers the existence of IDS or related cooperation system, IDS or cooperation system probable be the first target to be attacked. In such situation, the IDS should be firewall's key protection object.

1) The IDS should be transparent. IDS should be invisible in both internal network and external network except for firewall. Firewall filters all packets that try to access IDS, prevent IDS from attack.

2) The communication between IDS and firewall should be encrypted. In order to guarantee the confidentiality and consistency of communication, SSL protocol should be adopted to realize p2p encryption and authentication. SSL is the abbreviation of Secure Socket Layer. It works between the transport layer and the application layer, which can provide an effective way to guarantee consistency and confidentiality of data.

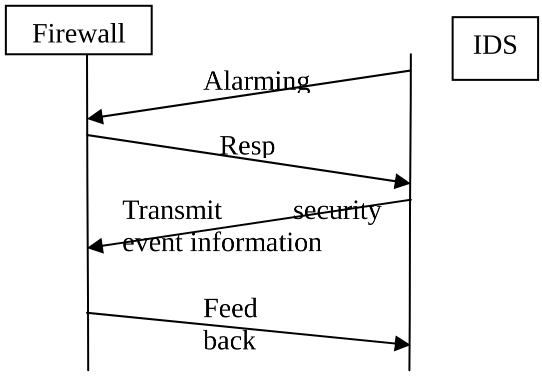


Figure 5. Security event processing flow

4. CONCLUSIONS

Based on the analysis of characteristic of firewall and Intrusion Detection System (IDS), this article proposes an approach to make firewall and IDS work together, making up between “static defense” and “dynamic monitoring” to obtain the higher network information security degree. Further more, this article has given a concrete solution which unifies firewall and IDS. This solution maximizes the functional advantage of firewall and IDS. Finally, security of cooperation system itself has been discussed. The further work is unifying firewall, IDS and network management system to provide a much safe, comfortable network environment.

REFERENCES

[1] Mo Zu-qin. Intrude Detecting Technique Summarizing. Metrology and Measurement Technique [J] 2005,32(12):p.24-25

[2] Li Yu-jun. Firewall and Intrusion Detection Based on the Linkage Mode. Computer Knowledge and Technology [J] 2009-03-021 p.560-562,575

[3] Wang Lei. Research and Implementation of Interaction Response Strategy with Firewall and Intrusion Detection System [D] HARBIN Institute of technology.2008

[4] Wang Dei-zheng. Studies on Network Intrusion Detection Technology Based on Protocol Analysis. Popular Science & Technology [J] 2009-01:p.13-14

[5] MA Tong-wei; et al. On Invasion Detection Technology of Network Security. Journal of Henan Mechanical and Electrical Engineering College. 2007, 15(3):p.21-22

ON SOFTWARE IMPLEMENTATION OF THE UNDERLYING FIELD MULTIPLICATION IN $GF(2^p)$ OF ELLIPTIC CURVE CRYPTOSYSTEMS

Minglong Qi, Qingping Guo, Luo Zhong
College of Computer Science & Technology, Wuhan University of Technology
West Mafangshan Campus, 430070 Wuhan China
Email: mlki01@sohu.com

ABSTRACT

In this paper, we have presented, improved and implemented some algorithms of the underlying binary extension field multiplication over $GF(2^p)$ $p = 163, 233$ of elliptic curve cryptosystems. We have used irreducible trinomials as the field element generator, and carried out the modulus reduction algorithm for these specific irreducible trinomials. Implementation and numeric experimental timings show that, if we divide a native machine word in equal parties called "windows", and pre-compute these small polynomial multiplications, then timings for the field multiplication operation can be considerably reduced.

Keywords: Binary Extension Field, Field Multiplication, Elliptic Curve Cryptosystem

1. INTRODUCTION

In cryptography, there are two kinds of encrypting and decrypting schemes: symmetric key cryptosystems and asymmetric cryptosystems, the last is synonym to public key cryptosystems. In the public key cryptographic schema, public key is used to encrypt plaintexts while the private key associated to the public one is used to decrypt ciphertexts. The most famous public key systems are, here cited as examples, RSA[1], ElGamel[2], Elliptic Curve Cryptography(ECC)[3,4], etc. Intractability of some modern number theory forms the basis of the security of these public key cryptosystems: hardness of factorizing a big integer as a product of two primes for the security of the RSA; hardness of discrete logarithm problem (DLP) over a finite field being primordial for the tranquility of the ElGamel; while the hard difficulty of the Elliptic Curve Discrete Logarithm Problem (ECDLP) in the additive abelian group formed by the points of an EC over a finite field is essential for the security of ECC.

What is the DLP over a finite field? Suppose now that $(G, \bullet)$ is a multiplicative cyclic group of order n with generator g , i.e., $G = \langle g \rangle$. For an arbitrary integer x in the interval [1, n-1], it is easy to compute $y = g^x$. But the inverse problem, i.e., given the generator g and an arbitrary group element y , it is very hard and even impossible to find x satisfying the identity from the view of computational complexity. This is called a discrete logarithm problem.

From a mathematical view, an elliptic curve is an affine variety over an algebraic closed field. The underlying mathematical background of the EC theory is very sophisticated. Roughly speaking, an EC is a simple cubic curve in the Cartesian plan $R \times R$. For computer security uses, we often substitute the underlying field R^2 by a finite field such as $GF(p)$, or $GF(p^n)$, or $GF(2^p)$. The group formed by all the points of the EC is an additive group, given Q, and P, it is very hard to find an integer n such as $Q = nP$, whereby Q, and P are two arbitrary points of the EC. This is the so-called elliptic curve discrete logarithm problem.

To efficiently implement (setup) an ECC, efficiency of the underlying finite field arithmetic operation is crucial for the performance of the EC cryptosystems, namely the field multiplication and inversion are two operations consuming very much computational times. An effort spent ameliorating these last is never usefulness. In this paper, we present, implement and ameliorate some algorithms of field multiplication in $GF(2^p)$ $p = 163, 233$.

The remainder of the article is organized as fallows: in section 2, we present and ameliorate some algorithms of the field multiplication in $GF(2^p)$ $p = 163, 233$; in section 3, we implement these algorithms in ANSI-C and give numeric experimental timings; in section 4, we give a short discussion about the efficiency of the algorithms according to our numeric experimental results.

2. BINARY FIELD MULTIPLICATION ALGORITHMS IN $GF(2^p)$ $p = 163, 233$

From elementary Galois Theory, we know that, binary extension fields $GF(2^p)$ or F_{2^p} is isomorphic to the finite field $F_2[x]/(f)$, where $f(x)$ is an irreducible polynomial in $F_2[x]$ of degree p . This field may be viewed as a vector space of dimension p over F_2, of which an arbitrary field element may be represented by a polynomial of degree at most $p-1$ with the form $\sum_{i=0}^{p-1} a_i x^i = a_{p-1}x^{p-1} + a_{p-2}x^{p-2} + \ldots + a_1 x + a_0$ where $a_i \in F_2$ $(i = 0 \cdots p-1)$ so that $(x^{p-1}, x^{p-2}, \cdots x, 1)$ forms a basis for the vector space. For example, with irreducible polynomial $f(x) = x^3 + x^2 + x + 1$, the finite binary field F_{2^3} or $GF(2^3)$ has eight elements that are respectively $\{0, 1, x, x+1, x^2, x^2+1, x^2+x, x^2+x+1\}$. It is easy to see that, there is a unique correspondence between the polynomial representation of a field element and a point in the space $(GF(2))^p$, a bit string of length p of form such as $(a_{p-1}, a_{p-2}, \cdots a_2, a_1, a_0)$. For example, the point $(1,1,0)$ in the space $(GF(2))^3$ is an alternative form of the element $x^2 + x$ in the finite binary field $GF(2^3)$.

Bitwise operations such as XOR, bit shift etc. on a field element $(a_{p-1}, a_{p-2}, \cdots a_2, a_1, a_0)$ might be substituted by machine word level operations at certain moment in order for improving efficiency of an algorithm. In such situation, it is preferable to regroup the bit string of the field element in an array of native machine-word. In this paper, we always use the machine-word array for representing an arbitrary field element. Let W be the length of a native processor word that generally is a multiple of 8 (in this paper, we adapt that $W = 32$), and let $L = \left\lceil p/W \right\rceil$, then the bit string of the field element may be replaced by a machine-word array $(A_{L-1}, A_{L-2}, \cdots A_2, A_1, A_0)$. The most-right word has $s = W \times L - p$ bits unused. Let U and V be two words of an element in the field $GF(2^p)$ in the forms of machine-word array, the following standard notation is used to denote operations on the words: $U \oplus V$ standing for bitwise exclusive-or; $U \& V$ denoting bitwise AND; $U >> i$ for right shift of U by i positions with the i high-order bits set to 0; $U << i$ for left shift of U by i positions with the i low-order bits set to 0. From the programming view in C, we can define these parameters in a head file; frequent practice is similar as in table 2-1:

Table 2.1 Definition of Parameters for $GF(2^{163})$ in C

```
#ifndef FF_ARITHMETIC_H_
#define FF_ARITHMETIC_H_
typedef unsigned int WORD
#define W    (sizeof(WORD)*8)
#define DIMENSION 163
#define         L              ((DIMENSION-1+
sizeof(WORD)*8)/(sizeof(WORD)*8))
… …
#endef
```

How to perform a left shift operation on a field element in form of a native machine-word array? If the most left bit of a word of the field element is equal to 1, after the left shift operation, this bit is gone into the most right bit of the word just at the left of the word in consideration. This consideration gives us the next algorithm 2.1:

Algorithm 2.1 Left Shift Operation on a Field Element

INPUT: A field element $A = (A_{L-1}, A_{L-2}, \cdots A_2, A_1, A_0)$.

OUTPUT: The same field element after a left shift operation.

1. $t_1 \leftarrow 0, t_2 \leftarrow 0$
2. For i from 0 to $L-1$ do
   2.1 If $A[i] \,\&\, 0x80000000 \neq 0$ then $t_2 \leftarrow 1$ Else $t_2 \leftarrow 0$ End If
   2.2 $A[i] \leftarrow A[i] << 1$
   2.3 If $t_1 \neq 0$ Then $A[i] \leftarrow A[i] \oplus 0x1$ End If
   2.4 $t_1 \leftarrow t_2$
   End For

The efficiency of polynomial multiplication in the underlying finite field is the dominant performance constraint of an EC cryptosystem, especially in projective representation [5]. So that any effort spent ameliorating this operation is well spent. Because a field element in the $GF(2^p)$ can be represented by a polynomial of degree at most $p-1$, multiplication of any two field elements is reduced to the multiplication of the two representative polynomials. The classical algorithm for polynomial multiplication is called "right-to-left shift-and-add field multiplication", which is based on the next consideration:

$$a(x)b(x) = \sum_{i=0}^{p-1} a_i b(x)$$

$$= a_{p-1}x^{p-1}b(x) + a_{p-2}x^{p-2}b(x) + \cdots$$

$$+ a_1 x b(x) + a_0 b(x)$$

The term $b(x)x^i$ is nothing other than the vector representation of $b(x)$ after i times of left shift operation, that is, $b(x)x^i \cong B << i = (B_{L-1}, B_{L-2}, \cdots B_1, B_0) << i$. If we iterate on exponents from right to left, then in each iteration, if the corresponding bit a_i is not null, we can do a left shift operation on B and add it to a vector of length $2L-1$ that is the product of $a(x)b(x)$. See Algorithm 2.2.

Algorithm 2.2 Right-to-left shift-add Field Multiplication

INPUT: $a(x) = (A_{L-1}, A_{L-2}, \cdots A_1, A_0)$ and $b(x) = (B_{L-1}, B_{L-2}, \cdots B_1, B_0)$

OUTPUT: $c(x) = a(x)b(x) = (C_{2L-1}, C_{2L-2}, \cdots C_1, C_0)$

1. For i from 0 to $p-1$ do
   1.1. $j \leftarrow i/W$ /*calculate the word index in the array A from the bit index */
   1.2. $k \leftarrow i\%W$ /*calculate the bit position in the word $A[j]$ */
   1.3. If $A[j] \& (1 << k) \neq 0$ Then $C \leftarrow C \oplus B$
   1.4 $B \leftarrow left\_shift(B)$
   End For
2. Return C

We denote a truncation of a field vector in Matlab style: $C(j:) \cong (C_{2L-1}, C_{2L-2}, \cdots C_{j+1}, C_j)$, and $C(j:) \oplus B$ means $C_{j+k} \leftarrow C_{j+k} \oplus B_k \;\; k = 0 \cdots L-1$.

Consider how to compute a term Bx^{jW+k} from Bx^k. The fact is that, the first term may be calculated by simply left shifting j words of the second, or alternatively appending j words of zeros at left of the vector B. Based on this consideration, we have an algorithm called "Right-to-left combo method for polynomial multiplication". See Algorithm 2.3.

Algorithm 2.3 Right-to-left Combo Method for Polynomial Multiplication

INPUT: $a(x) = (A_{L-1}, A_{L-2}, \cdots A_1, A_0)$ and $b(x) = (B_{L-1}, B_{L-2}, \cdots B_1, B_0)$

OUTPUT: $c(x) = a(x)b(x) = (C_{2L-1}, C_{2L-2}, \cdots C_1, C_0)$

1. $C \leftarrow 0$
2. For k from 0 to $W-1$ do
   For j from 0 to $L-1$ do
   If $A[j]\,\&\,(1 << k) \neq 0$
   Then $C(j:) \leftarrow C(j:) \oplus B$
   End For
   If $k \neq W-1$
   Then $B \leftarrow left\_shift(B)$
   End For
3. Return C

Another version of the combo method for polynomial multiplication is called "left-to-right combo method for polynomial multiplication". We process the most right word of the polynomial $a(x) = (A_{L-1}, A_{L-2}, \cdots A_1, A_0)$, that is A_0, bit by bit and from left to right, say $x^k B$ $k = W-1 \cdots 0$, and $x^{jW+k} B$ $j = 0 \cdots L-1$. Given k and loop over $j = 0 \cdots L-1$ once done, we left shift the product vector C. It is a bit more difficult to understand this new version, but after rearranging the product formula, the algorithm seems straightforward self-clear:

$$c(x) = \cdots((a_{(L-1)W+k} x^{(L-1)W} B + \cdots a_{0W+k} B)x +$$

$$(a_{(L-1)W+k-1} x^{(L-1)W} B + \cdots a_{0W+k-1} B))x^{k-1} + \cdots$$

for $k = W-1 \cdots 0$. For a k given, the inner loop is doing

$InnerLoop(k) = a_{(L-1)W+k} x^{(L-1)W} B + \cdots a_{0W+k} B$, and $(InnerLoop(k))x$ corresponding a left shift operation of the product vector C. See Algorithm 2.4.

Algorithm 2.4 Left-to-right Combo Method for Polynomial Multiplication

INPUT: $a(x) = (A_{L-1}, A_{L-2}, \cdots A_1, A_0)$ and $b(x) = (B_{L-1}, B_{L-2}, \cdots B_1, B_0)$

OUTPUT: $c(x) = a(x)b(x) = (C_{2L-1}, C_{2L-2}, \cdots C_1, C_0)$

1. $C \leftarrow 0$
2. For k from $W-1$ to 0 do
   a) For j from 0 to $L-1$ do
   b) If $A[j]\,\&\,(1 << k) \neq 0$ Then $C(j:) \leftarrow C(j:) \oplus B$
   c) End For
   d) If $k \neq 0$ Then $C \leftarrow left\_shift(C)$
   e) End For
3. Return C

Remark that Algorithm 2.3 should be faster than Algorithm 2.4, because left shift operation is doing on a vector of length less long.

If we divide a machine word of the vector A into equal parties called "windows" by certain authors [6, 7], and pre-compute the products of the polynomials represented by these windows with the vector B, we can considerably reduce computational times for polynomial multiplication. For example, let $WWidth = 4$ be the length of the windows, and let $N = W/WWidth = 32/4 = 8$ be the number of the windows of a word, then the vector A can be represented by

$$A = (A_{L-1}, A_{L-2}, \cdots A_1, A_0)$$
$$= ((A_{L-1}^{N-1}, A_{L-1}^{N-2}, \cdots A_{L-1}^{1}, A_{L-1}^{0}) \cdots$$
$$(A_0^{N-1}, A_0^{N-2}, \cdots A_0^{1}, A_0^{0}))$$
$$= \begin{bmatrix} A_0^{N-1} & \cdots & A_0^{0} \\ \vdots & \vdots & \vdots \\ A_{L-1}^{N-1} & \cdots & A_{L-1}^{0} \end{bmatrix}$$

Denote

$$TVector\ \ wb[L][N]; wb[j][n]$$
$$= A_j^k \otimes B\ \ j = 0 \cdots L-1, n = 0 \cdots N-1$$

whereby $TVector$ represents the type of a field element vector, and

$$A_j^n \otimes B =$$
$$(0_{L-1}, 0_{L-2}, \cdots, 0_1, 0x0000000(A_j^n)) \otimes B\,.$$
$$= t(x)b(x)$$

After this preparation, we can give an algorithm for field multiplication, called "Left-to-right combo method with windows of width $WWidth$". See Algorithm 2.5.

Algorithm 2.5 Left-to-right Combo Method with Widows of Width $WWidth$

INPUT: $a(x) = (A_{L-1}, A_{L-2}, \cdots A_1, A_0)$ and $b(x) = (B_{L-1}, B_{L-2}, \cdots B_1, B_0)$

OUTPUT: $c(x) = a(x)b(x) = (C_{2L-1}, C_{2L-2}, \cdots C_1, C_0)$

1. $TVector\ \ wb[L][N]$
2. Compute $wb[L][N]$
3. $C \leftarrow 0$
1. For k from $N-1$ down to 0 do
   a) For j from 0 to $L-1$ do
   b) $C(j:) \leftarrow C(j:) \oplus wp[j][k]$
   c) End For
   d) If $k \neq 0$ Then $C \leftarrow left\_shift\_WWidth\_bits(C)$
   e) End For
2. Return C

Remark that Algorithm 2.4 is a special case of Algorithm 2.5 with the width of windows equal to 1. In Algorithm 2.5, on the vector of the product should be performed the number equal to the width of the windows times of left shift operation, that is, $C \leftarrow Cx^{WWidth}$.

For the binary extension field $GF(2^{163})$, NIST in FIPS 186-2[8] recommended an irreducible polynomial $f(x) = x^{163} + x^7 + x^6 + x^3 + 1$, that permits fast modulus reduction for polynomial multiplication in the field. In [9, 10], the authors have shown that use of an irreducible trinomial can speed up even future the reduction operation. The irreducible trinomial is under a form as $f(x) = x^m + x^t + 1$ with $t \le \lfloor m/2 \rfloor$. In this paper, we adapt the irreducible trinomial $f(x) = x^{163} + x^{74} + 1$ for the finite field $GF(2^{163})$, and $f(x) = x^{233} + x^{74} + 1$ for the finite field $GF(2^{233})$. For working out an algorithm of modulus reduction for $GF(2^{163})$, and similar for $GF(2^{233})$, we should observe carefully reduction of monomials of degree greater than $p(p = 163, 233)$. For example, the monomial x^{288} can be expressed in

$$x^{288} = x^{163}x^{125} =$$
$$(f_{163}(x) - x^{74} - 1)x^{125}$$
$$\equiv x^{199} + x^{125} (Mod\ f_{163}(x))$$

This gives us an inspiration that a word of higher-order can be expressed as linear combination of several words of lower-order. As an example, we shall carry out this kind of linear combination for the word $C[9]$ of the vector $C = (C_{2L-1}, C_{2L-2}, \cdots C_1, C_0)$ to be reduced.

$$C[9] = \sum_{i=288}^{319} c_i x^i$$
$$= (c_{319}, c_{318}, \cdots c_{289}, c_{288})(x^{319}, \cdots x^{288})^T$$
$$\equiv \overrightarrow{C_9}(x^{230} + x^{156}, \cdots x^{199} + x^{125})^T$$
$$\equiv \overrightarrow{C_9}(x^{230}, \cdots x^{199})^T \oplus \overrightarrow{C_9}(x^{156}, \cdots x^{125})^T$$
$$\equiv (C[9] >> 25)(x^{255}, \cdots x^{224})^T$$
$$\oplus (C[9] << 7)(x^{223}, \cdots x^{192})^T$$
$$\oplus (C[9] >> 3)(x^{159}, \cdots x^{128})^T$$
$$\oplus (C[9] << 29)(x^{127}, \cdots x^{96})^T$$
$$(\text{mod } f_{163}(x))$$

Based on above observation, we can establish an algorithm in Algorithm 2.6 for fast reduction modulo $f(x) = x^{163} + x^{74} + 1$.

Algorithm 2.6 Fast Reduction Modulo $f(x) = x^{163} + x^{74} + 1$ in $GF(2^{163})$

INPUT: A binary polynomial $c(x)$ of degree at most 324.

OUTPUT: $c(x) \bmod f(x)$.

1. For i from 10 down to 6 do
   $T \leftarrow C[i]$.
   $C[i-6] \leftarrow C[i-6] \oplus (T << 29)$
   $C[i-5] \leftarrow C[i-5] \oplus (T >> 3)$
   $C[i-3] \leftarrow C[i-3] \oplus (T << 7)$
   $C[i-2] \leftarrow C[i-2] \oplus (T >> 25)$
   End For
2. $T \leftarrow C[5] >> 3$
3. $C[0] \leftarrow C[0] \oplus T$
4. $C[2] \leftarrow C[2] \oplus (T << 10)$
5. $C[3] \leftarrow C[3] \oplus (T >> 22)$
6. $C[5] \leftarrow C[5] \,\&\, 0x7$
7. Return $(C[5], C[4], C[3], C[2], C[1], C[0])$.

3. NUMERICAL EXPERIMENTAL TIMINGS

All the algorithms for field multiplication with reduction discussed in the section 2 were implemented in ANSI-C with compiler MinGW GCC 5.5 running on a Pentium IV based PC. A thousand of field elements in $GF(2^p)$ $p = 163, 233$ was randomly generated, and timings were calculated averagely over 300,000 field multiplications with reduction of any two arbitrary field elements randomly taken from this set of the elements. The numeric results are shown in Table 3.1. From which we can see that, the most fast software implementation is due to the algorithm of "left-to-right combo method with windows" with pre-computation, the slowest algorithm is the "right-to-left shift-add" method, and two others called "combo method" are faster than the "right-to-left shift-add" method and have same order of timings. But the left-to-right combo method is slightly slower than the right-to-left combo method.

Table 3.1 Timings for Field Multiplication Over $GF(2^p)$ $p = 163, 233$

ALGORITHMS	TIMING FOR $GF(2^{163})$ (ms)	TIMING FOR $GF(2^{233})$ (ms)
Right-to-Left Shift-Add	60.623	97.343
Right-to-Left Combo	10.57	15.780
Left-to-Right Combo	11.04	15.930
Left-to-Right Combo-Win Without Pre-computations	38.437	64.893
Left-to-Right Combo-Win With Pre-computations	0	0

4. CONCLUSIONS

In this paper, we have presented, improved and implemented some algorithms of the underlying binary field multiplication in $GF(2^p)$ $p = 163, 233$ for elliptic curve cryptosystems. We have demonstrated that use of irreducible trinomials and the algorithm of "combo-windows" with pre-computation can speed up dramatically the underlying field multiplication.

ACKNOWLEDGMENTS

Sincere thanks for helpful and inspiratif discussions with Professor GUO, Dean of the Key Laboratory of Parallel Computing, and Professor Zhong, general director of the College of Computer Science and Technology of WHUT. However, the originality and the authenticity of this technical report are entirely under the responsibility of Dr Qi. The implementation C source code is in evoluation. In that if you might be interested, please email to us.

REFERENCES

[1] R. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems", Communications of ACM, Vol. 21, 1978, pp.120~126.

[2] T. ElGamel, "A public key cryptosystem and a signature scheme based on discrete logarithms", IEEE Transactions on Information Theory, Vol. 31, 1985, pp. 469~472.

[3] N. Koblitz, "Elliptic curve cryptosystems", Mathematics of Computation, Vol.48, 1987, pp. 203~209.

[4] V. Miller, "Uses of elliptic curves in cryptography", Advances in Cryptology-Crypto '85, LNCS 218, 1986, pp.417~426.

[5] N. P. Smart, "A comparison of different finite fields for elliptic curve cryptosystems", Computers and Mathematics with Applications, Vol. 42(1-2), July 2001, pp. 91~100.

[6] P. Ning, and Y. L. Yin, "Efficient Software Implementation for Finite Field multiplication in Normal Basis", Pending US Patent Application. Provisional patent application field in December 1997.

[7] C. Koc, and T. Acar, "Montgomery multiplication in $GF(2^k)$", Designs, Codes and Cryptography, Vol.14, 1998, pp.57~69.

[8] National Institute of Standards and Technology, Digital Signature Standard, FIPS Publication 186-2, February 2000.

[9] R. Schroeppel, H. Orman, S. O'Malley and O. Spatscheck, "Fast key exchange with elliptic curve system", Advance in Cryptology, Proc. Crypto '95, LNCS 968, D. Coppersmith, Ed., Springer-Verlag, 1995, pp. 43-56.

[10] D. Hankerson, A. Menezes and S. Vanstone, "Guide to Elliptic Curve Cryptography", Springer-Verlag, 2004.

THE GAMBLING ANALYSIS ON CREDIT RISK OF BANK IN INVESTIGATION STAGE WITH ASYMMETRIC INFORMATION— US SUB-PRIME CRISIS: IMPLICATIONS FOR BANKS' INVESTIGATION STAGE

Ye Shujun, Meng Qian
School of Economics and Management, Beijing Jiaotong University
Beijing, China
Email: shjye@bjtu.edu.cn

ABSTRACT

Information asymmetry is the reason for the risks of adverse selection in banks' investigation stage. This paper establishes a gambling model to analysis the risk of adverse selection, pointing out that banks in investigation stage can not use the probability of review raised as a solution to asymmetric information. The banks can get more information to improve the efficiency of the review, but the existing mortgage loans did not improve efficiency. Only setting up a system to enhance the information exchange can prevent the risk of adverse selection effectively. Ignoring the investigation and over-emphasising on loan collateral are the reasons for US sub-prime crisis, which is a warning to the loan security of China's bank.

Keywords: Credit risk, Game theory, Adverse selection, Asymmetric information, US sub-prime crisis.

1. THE IMPLICATIONS OF THE US SUB-PRIME MORTGAGE CRISIS FOR THE SAFETY OF CHINA'S BANK LOANS

U.S. sub-loan crisis leads to financial crisis in the world and the world economic recession has taken place. The cause of the economic crisis in the United States is the sub-prime home loans. Sub-prime mortgages are provided to the people with bad credit, low income, high debt ratio and higher probability of default. There are obvious risks with the loan itself. The reason for the bank loans is that banks have the housing as the mortgage and expect the price to be up. Even if the borrower does not repay, the bank can sell the housing to make up for the loss. Driven by the interests and to expand customer groups and markets, lending institutions create new varieties of loans for the borrowers who do not meet the conditions. At the same time, a lot of marketing work even be outsourced to an independent brokerage firm. Lending institutions give up the control of review on customer information. This provides an opportunity to fraud for the lack of professional ethics people.

Through the analysis of the causes of the crisis, China's banks can put forward a warning to avoid the credit risk: First, the investigation work must be done carefully and it can prevent the occurrence of adverse selection risk; Second , banks can not emphasis on collateral too much.

2. ANALYSIS FOR THE RISK OF ADVERSE SELECTION WITH ECONOMICS OF INFORMATION

The economics of information believes that the risks arise from uncertainty. Accordingly, the credit risk of banks in the credit markets was due to the uncertainty and loss in the use of funds. Uncertainty is a state of incomplete information essentially. So the amount of information increasing will reduce uncertainty and risk.

Banks do not understand the applying companies' information fully before the contracts signed, including the earnings and risk of the loan, the ability to repay, the financial condition of enterprises and so on. Banks are in a state with disadvantage information, but companies are in the information superiority with a better understanding of their situation. Information is asymmetric. There are companies with good credit and companies with bad credit. Banks can not distinguish because they don't have enough information. Banks have to make the lending rates according to the average level of risk of companies in the loan market. This rate is higher than the rate which the good companies are willing to bear, but lower than the rate which the bad enterprises are willing to bear. The bad enterprises may be in order to get business loans to whitewash its own and take risks. In this way, good companies often have to leave the loan market. Banks may choose bad enterprises which provide a higher rate as qualified candidates. It is the phenomenon that bad enterprises expel good companies. Banks' adverse selection leads to increased credit risk and decreased credit quality, resulting in providing less loan. The less loans banks provide, the more intense competition enterprises do. Banks can't distinguish between good and bad companies, thus a vicious circle. The key to the question is to let banks know more information of the applying companies in order to change the bank's information disadvantage and reduce credit risk.

3. THE STATIC GAMBLING ANALYSIS OF ADVERSE SELECTION RISK IN THE INVESTIGATION STAGE WITH ASYMMETRIC INFORMATION

At present, China's commercial banks are still not with a good quality of the loan. According to the five-level classification of non-performing loans at the end of 2007, the total number of non-performing loans is 1.3 trillion yuan and the non-performing loans rate is 6.2 percent. It doesn't meet the requirement in the "Basel II" that advanced international banks' non-performing loan ratio should be maintained below 5%. As a result, China's commercial banks have to do a good job in the survey in order to select the good credit borrowers in the market and prevent the occurrence of credit risk.

The banks' survey plays an important role in decreasing the information asymmetry between banks and borrowers. It can avoid providing loans to the customers with bad credit and reduce credit risk.

There are two types of companies in the loan market. One is with good credit that they have a better ability to operate and a stronger will to repay. And the other is with bad credit that they have worse are less ability to operate and a weaker will to repay. In order to get loans, companies may have their own application for a cosmetic and different enterprises have different whitewash cost. Therefore, good and bad companies both have the potential to whitewash, but they are with different cost. The companies' strategy is 'whitewash' or 'no whitewash'. Companies do not know whether they will get the loans, but they know the probability.

In order to avoid the risk, commercial banks will review the material to decide whether or not to lend the loans. Therefore, banks' strategy is 'review' or 'no review'. Banks do not know the applying company is good or bad. They only know the probability distribution. Banks do not know whether the companies make a cosmetic in their material either. They also only know the probability to make a cosmetic. Once banks find that the company whitewashes its applying materials, the company will be punished.

A good company will repay the principal and interest on time surely, but a poor company will not. So even if the good company was found to whitewash, it still get the loan but with punishment. Lending to the bad company, banks will suffer losses. So once the bank knows the applying company is a bad one through reviewing, it will not provide the loan.

Assumptions: the amount of loans is W; the loan's interest rate is r; a good yield company's investment rate of return is Rg; a poor corporate's investment rate of return is Rb; the cost of bank's review is M; the cost of whitewash is F; the loss of being punished is L; the bank's review probability is p; the probability of good company is q; the probability of whitewash is n.

List benefit matrix after analysis the banks' and enterprises' benefit under a variety of strategies.

Table 3-1: Benefit Matrix

Bank \ Company	Good company	
	Whitewash	No whitewash
Review	Wr-M，W（Rg-r）-F-L	Wr-M，W（Rg-r）
No review	Wr，W（Rg-r）- F	Wr，W（Rg-r）

Bank \ Company	Bad company	
	Whitewash	Whitewash
Review	-M，-F-L	-M，-F-L
No review	-W，W（1+Rb）-F	-W，W（1+Rb）-F

Banks make an adverse selection, which is, although banks have reviewed companies' applications for loans, they still make the wrong decision to lend loans to poor business. The reason is that banks can get more benefit here than good companies: $p\ (Wr-M) + (1-p)\ Wr < -W$, $p> [W\ (r+1)] / M$. $W(r+1)$ is the total benefit for banks; M is the cost of review; $[W\ (r+1)] / M$ can be seen as banks' benefit per cost. P is the probability of review, '$p> [W\ (r+1)] / M$' suggests that when the probability is greater than the benefit per cost, banks will make an adverse selection and lend to bad company.

This appears to be inconsistent with common sense that the greater the probability of the bank's review is, the less mistakes banks make. But we need to know that the probability of banks' review and the probability that banks can make a right decision are not the same. Here comes a question that whether the review is effective. It is proved that the existence of asymmetric information between banks and companies makes it difficult to judge the quality of the business effectively and accurately. Now '$p> [W\ (r+1)] / M$' can be understood: the higher the probability is, the higher cost of the review is. When it is beyond the benefit per cost, banks will look for the company which provides higher loan interest rates to lend. Actually in the market, the loan rates that the good companies provide are below average rate, while the bad companies are willing to take risks and provide higher loan rate. If the banks' review is not effective, banks will make an adverse selection and lend to the bad. At the time, the good have to leave the loan market and the loan market will be full of more and more poor business. The result is the quality of banks' loans will decline. If banks is always increasing the probability of the review rather than raising the quality of the review, it will enter the vicious circle mentioned above. Therefore, improving the quality of the review is the key to the question.

Since China's banks' non-performing loans rate is high, banks review each loan in practice. The review probability is hundred percent. At this point, the cost is so much that banks tend to lend loans to enterprises which provide a higher rate and attention should be paid at this time to prevent the occurrence of adverse selection. In order to understand the real situation of enterprises and reduce the information asymmetry between banks and enterprises, China's banks' investigation Stage including three stages: accepted, investigation and approval. Accepted stage includes: customer application - qualification - guest to submit material - a preliminary review. Staffs review the qualifications of customers and the application materials in accordance with relevant laws, regulations and credit policy to decide whether to accept the customer's credit application. In the qualification stage, there is a number of restrictive conditions to control risk. Preliminary review includes basic information of customers, credit information and information about the collateral. In investigation stage, it needs to evaluate clients, business and collateral. In the approval stage, credit approval officers review the applying materials in accordance with national guidelines, policies, laws, regulations and credit policy and analysis the economic and commercial feasibility of the project. According to the credit risk and revenue of the business, the officers decide whether to approve.

We can see in our country banks evaluate three aspects in the investigation process: customer survey, business survey and collateral survey. But in practice the banks pay more attention to the collateral, they ignore the cash flow factor and the whole value of the enterprise.

4. NEGATIVE IMPACT OF OVER-EMPHASIS ON COLLATERAL IN THE INVESTIGATION

U.S. releases a large number of sub-prime mortgage. It reflects banks are over-reliance on collateral under the macro background with rising house prices. The model has a built-in defect: the value of the collateral goes up and down with the volatile market. Once the market goes down, the value of the collateral will be greatly devalued. So the borrowers default. China's commercial banks are generally not to provide loans without collateral. Usually the companies provide real estate, securities, machinery and equipment as collateral. When companies can not return the loans, the banks would gain ownership of the collateral as compensation. However the liquidity of China's financial market is not high, collateral is always not changed into cash timely. At the same time the value of the collateral will change. All of these uncertainties will lead to banks' losses. Sometimes companies will not give collateral to the bank when they default. Even though banks appeal to resolve this problem, they often will not get the results. Banks have to pay legal fees and cost so much time and it will increase the opportunity cost. As a result, although our country builds a mortgage loan system, the quality of banks' credit asset has not improved. Mortgage loan system is only a way to make up for the loss for banks. Although the occurrence of mortgage loans is prior to risk, it does not play the role of early warning and avoiding the risk. So it does not reduce the information asymmetry between banks and enterprises and help banks identify the good or bad companies accurately.

Commercial banks should not emphasis on collateral too much, because the stability of the borrower's cash flow is more important. The decision whether lend or not should be based on the customer's ability of creating cash flow and personal credit primarily but not on raising collateral requirements only. Cash flow is the first source of repayment and collateral is the second source of repayment. The new Basel Capital Accord emphasis: the cash flow is the first source to repay bank loans; the commercial banks should be in the proper use of the terms of collateral requirements in the loan contract and should not one-sided emphasis on the capabilities of risk prevention and loss recovering; banks should pay more attention to the function that collateral can help banks resolve the problem of adverse selection. As a result, the bank should reduce the loan collateral requirements and strengthen its market competitiveness by increasing the ability of information screening which help to judge whether an enterprise is with the ability to repay.

5. MEASURES FOR BANKS TO AVOID ADVERSE SELECTION RISK

The reason why adverse selection risk arises is the information asymmetry between the bank and company. The important solution is to promote the exchange of information between the two parts.

In the investigation stage, investigators need to communicate with the company deeply so as to verify financial data and get the information about the enterprise's production, operation and management. Banks should make a comprehensive analysis with a large number of data in order to get a valuable conclusion which is helpful to make decision.

The establishment of a new type of relationship between banks and enterprises will help banks have a full understanding of the companies and ownership some rights about the financial control. For example, part-time exchange of directors is the good experience. According to Britain and the United States experience, part-time directors play a very active role in the supervision of the company's financial and the scientific of the operational decision-making.

In addition, the establishment and improvement of the credit system is an effective method. It is to establish a credit file for every enterprise. The file is open to all the banks and the other companies in the loan market which help banks make up their lack information of applying companies and reduce the cost of getting information. It makes a single game change into a repeated game. Banks can avoid lending to poor credit company with the raising adequacy and transparency of information.

REFERENCES

[1] Zhou Xueping，2008. US sub-prime crisis: Implications for domestic banks' operation and management [J]. CHINA MONEY5，45-47.

[2] Liang Dapeng, Qi Zhongying2004. Research about asymmetric information and the "adverse selection" in credit [J]. TECHNOECONOMICS & MANAGEMENT RESEARCH4，67-68.

[3] Zhao Rui, Zhou Xirun2006. The gambling analysis on commercial bank credit risk [J]. CHINA SCIENCE AND TECHNOLOGY INFORMATION18，223-224.

[4] Song Ying2006. The gambling analysis on commercial bank credit risk [J]. FINANCE AND ECONOMY6，74-75.

[5] Jiang Haijun, Hui Xiaofeng2006. Collateral's impact for bank credit risk and social welfare [J]. ZHEJIANG FINANCE,5，14-15.

[6] Simon C. Parker2007. Entrepreneurial learning and the existence of credit markets[J]. Journal of Economic Behavior & Organization，37–46.

Ye Shujun is a finance professor, doctoral tutor and member of Chinese Finance Association, now dean of Finance Department of School of Economics and Management, Beijing Jiaotong University. Got the degrees of bachelor, master and doctorate separately from Xi'an University of Technology, Tsinghua University and Beijing Jiaotong University. 1997, engaged in academic exchanges in Duquesne University of Pittsburgh, the United States; 2001, went to Germany to survey the railway's investment and financing. The main research orientation are: financial institutions, theory and development strategies, financial risk management theory and evaluation, microstructure theory on financial market. And has taken charge and participated more than 30 scientific research projects from National Natural Science Foundation of China, China Insurance Regulatory Commission, the Ministry of Commerce, the Ministry of Railways and so on. Published more than 40 academic articles on specialized academic journals, some of which can be searched by international information search client. Published a monograph, edited and published 8 teaching materials and 1 translation.

Meng qian graduated from the Finance Department of Beijing Jiaotong University in 6.2008, enrolled in 9.2008, major in finance, is a graduate student in Beijing Jiaotong University now. Research interests are in risk of commercial banks and personal financial services. Published a paper 'The impact of financial crisis to commercial banks' personal financial services and the response measures' in New Finance in 6.2009.

THE EXPLORATION OF COMPUTER ACTIVE DEFENSE STRATEGY

Tang chuhua, Guo Qingping
The Wuhan University of Technology
Wuhan,430063,HuBei
Email: Tangchuhua@126.com

ABSTRACT

This article is to discuss issues of computer security defense in perspective of active defense. In this paper, the main strategy of design are: Three kinds of technigy-digital signature, behavioral surveillance, feedback system of verification and scoring, constitute the three defense lines of active defense of computer system to fight against virus attacks. It will successively avoid defects-lag and passiveness. The traditional virus detection is over-reliance on "Eigenvalue Scanning" technology, which has led to those defects. Active defense strategy will be able to greatly strengthen security of computer system.

Keywords: Active Defense, Digital Signature, Behavioral Surveillance, Feedback System.

1. RESEARCH BACKGROUND AND CURRENT SITUATION

1.1. Research Background

Traditional anti-virus technology can be used to prevent known viruses, but can not resist tens of thousands of new viruses threats in a day, which resulting in that the user's information security is not effectively protected.Therefore, it is necessary to design a new strategy for redesigning Anti-Virus architecture to determine the unknown viruses. So that it is more effectively to resist the virus attack to computer systems and maintain the user's information security.

Active defense of computer is the goal that anti-virus experts have been pursuing.Generally speaking, the "pre-emptive" is much better than "never too late to mend".With the emergence of active defense, the computer has a certain intelligence and immunity. As we all know, the traditional antivirus software's detecting viruses principle is collected the signature features that have appeared into the database, and then use this database to analyze the characteristics of the same type of viruses.[1] Virus killing tools couldn't do anything on new viruses.Today in particular, the "wisdom" of viruses makers is more and more superb, and makers make their own "work" have new features to combat virus killing tools based on the "eigenvalue scanning" technology[2]. For example, some new virus can be achieved in adding shells automatically, and locating and modifying the characteristics of the viruses.That may make virus killing tools couldn't do anything any more. Once a great number of viruses have similar characteristics, it would be a disaster for security defense.

1.2.Status of Current Research

In response to the current grave situation of anti-virus, experts at home and abroad in active computer defense technology has taken the initiative to carry out research and receive certain results in the corresponding areas.[3]

For example, experts of Kaspersky anti-viruses focus their emphases on research of heuristic technology. Heuristic Anti-viruses technology refers to "the ability of self-discovery" or "knowledge and skills that use some methods to determine things". Antivirus software can analyze the logical structure of codes of documents to figure out if it contains any malicious program features, or proactively running codes to determine whether there is any malicious behaviors in a virtual security environment. In the field, the former is called *static code analysis* and the latter is called *dynamic virtual machine*.

Chinese scientists also haved started fruitful research in active defense. In the international field, "Micro-point" anti-virus experts take the lead in creating the anti-virus active defense system with "simultaneous monitoring, dynamic protection"[4]. According to "determining law of independent analysis of program behavior" theory and the definition that computer viruses generally followed in the field, adopting dynamic simulation technology, analysising program behavior, determining the nature of the programmes to realize active defense to new viruses. Kingsoft's anti-virus experts adopting technology of monitoring behavior and *bootclean snatching killing* to identify and kill viruses actively before that they destroy users's system.

2. ARCHITECTURE OF ACTIVE DEFENSE

In this paper, the design of the digital signature, the process behavior monitoring, verification and score feedback system constituted three functional modules of active defense architecture. These are shown in Figure 1.

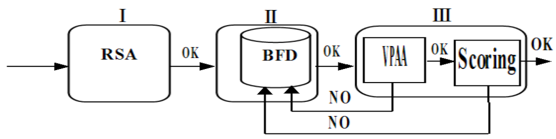


Figure 1. Sructure of Active Defense System

1) Digital signature modules. For a large number of viruses disguised themselves as legitimate process to run on a computer system, this article have designed a digital signature (RSA) to complete authenticating software, which it will make illegal procedures be out of computer system.

2) Process behavior monitoring module. Getting rid of over-reliance on "Eigenvalue scanning" technology and monitoring all behaviors of the processes running on the computer. With the help of Behavioral Characteristics Database(BFD), determining processes is more effective in detecting and killing the viruses.

3) Validation and scoring feedback system module. Completing the Verification of Process Access Authority(VPAA) and quantifying scoring the behavior of the processes, then return processes behavior characteristics to database. In this way, we can not only prevent the case of ultra vires accesses, but also go further to enrich the characteristics database and improve the intelligence of determination.

3. DIGITAL SIGNATURE AUTHENTICATION

In this paper, digital signature (in this article, making use of RSA algorithm) is introduced into to determinate the behavior of viruses in defense system in order to achieve the certification function to software[5]. In order to achieve certification and license functions to legitimate software, we need three representatives of this process: the software vendors *Producer*, User , software certificate authority Center CC that software vendors and users trust to participate in work. The relations between them in Figure 2

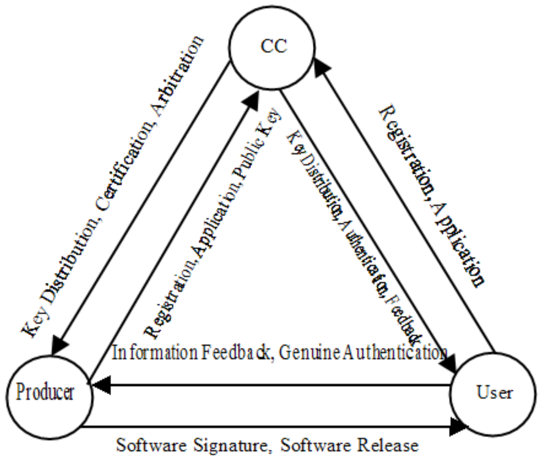


Figure 2. Authentication Model

In this figure, the Certification Center's task is to distribute secret keys(KD) to software Producer and software users, authenticate manufacturers' signatures , apply and certificate access authority. Also it informe manufacturers and users certification results. Users not only use their own private key to decrypt and authenticate the legality of software vendors, but also check the results of the authentication and certification center. It will ensure that the software really comes from legitimate manufacturers. Concrete realization of the process is as follows:

1) Certification Center generate keys and sequence <a,b,e,n>, then publish them. ***a*** is private key for software vendor achieving digital signature , and ***e*** is public key that used to decrypt the digital signature.

2)Software vendors apply for authorization certificate to the CA for releasing software. Software vendors can be eligible for release software only after get the certificate. The goal of designing this process is to be used as an important evidence in ruling when users and software vendors in the event of a dispute and take charge of arbitration in resolving economic disputes of the two sides'. At the same time, the certificate that contain important API function calling information and information about terminal vector also can be used as basis for certification of access authority in latter step. Software vendors are now able to apply their own important information to Certification Center : ｛PID, S, b, K1, K2, ...｝. In this case , PID is the unique software vendor identity; S is signature scheme logo that a software vendor have received from the Certification Center(in this article, using the RSA signature scheme), ***b*** is certificate password in the software vendor register certificate. Then, authentication Certification Center audit the registration information that software vendors provide, then award software vendors digital certificates to allow software vendors to release software：

$$L=[PID||S||b||K_1,\ K_2,\ \ldots\ldots]^{a} \bmod n$$

In this expression,"||" stands for juxtaposition computing; "[]" stands for the integer value calculated by the ASCII code.

(3) Completing signatures on software SW that vendors have released as follows:

$$\text{Sign}（i，SW）=[PID，SW，L]$$

(4) software users implement validating software obtained, which is to check software running in computer whether is actually belong to legitimate software vendors. This process is like this: When the software users have obtained packets signed, in following step, according to the expression:

$$[PID||S||b||K_1,\ K_2,\ \ldots\ldots]=L^{e} \bmod n$$

We have succeeded to restore unique software vendor identity, signature scheme, and private key. If decrypted data are completely in line with information that software vendors published, it show that software running in computer indeed arise from the releases that legitimate producers provide.

(5) Users inform certification center the results of verification, and attach its own digital signature to them.

(6) As a third party, Certification Center inform both manufacturers and users of judging reults of success of software verification.

Safety of Digital Signature (RSA) authentication scheme is based on the large number decomposition, which is admitted mathematical puzzle. Guarding the entrance into the user's system by the digital signature certification, the viruses into the system to would be significantly reduced, thereby it can effectively reduce viruses' attacks to system.

4. BEHAVIOR MONITORING AND DETERMINING

Of course, only relying on digital signatures to protect system security is not reliable, because there is no encryption algorithm to ensure that their will always be forever safe. Practice has proved that viruses behavioral surveillance is effective to monitor and control operation of the processes. In this way, we needn't to leave all the hope on security of algorithm. First of all, we need to kown something about basic principles of viruses behavior monitoring and common behavioral characteristics of viruses.

4.1 Basic Principles

By analyzing , summarizing, and summing up the law of the viruses, combined with anti-virus experience that experts determine viruses, we can extracte these into identification rules knowledge database, simulate that the mechanism experts found new viruses. In the hlep of a large number of probes appearing in the operating system and dynamic surveillance programs, we monitor the operation of the various procedures and analyze their behavior semantics. Then comparing with a series of action semantic rules database of programs, and automatically completing determination computer viruses. When an unknown program attempts to run and modify the critical data in operating system, the behavior will be detected by take active defense system. Then the behavior will trigger an early warning function in defense system , and inquire users whether allow relevant operation to continue, or directly interrupt operation.

4.2 Behavioral Characteristics and Achievement of Behavior Determination

Viruses' behavior characteristics are behavior semantics that virus process demonstrated in running. Viruses usually have common behavioral characteristics are as follows [6] :

1. Intercepting operation that caused by Interrupt Vector INT13H. In booting time, under normal circumstances, the system generally will first call the vector INT13H guide to normally start the operating system. The viruses may prohibit the operation of reading and writing disk by INT13H, the system will not start normally or modify the codes in vector to guide the operating system into the virus code area then start runing.
2. Rewriting or tampering documents in types of .COM or .EXE in order to achieve the purpose of hiding the true goal and carrying out malicious codes[7].
3. Utilizing automatically-run technology on CD-ROM. Primarily by ***autorun.ini*** files, the technology will run a program when users open files in a disk.
4. Viruses occupy a great deal of memory,and computer memory that is occupied couldn't be covered by user's programs and data.
5. Stealing access authority that only operating system has.

In addition to above characteristics, the viruses have many other behavior characteristics. We need bring all the behavioral characteristics into database of behavior semantics for determining the viruses, combined with relatively mature viruses characteristics database to determine the viruses. In practice running, monitoring and determination on viruses' behavioral characteristics is feasible. According to the basic knowledge of operating systems principle, we know that a procedure or a process is necessarily call the API function of operating system through interfaces if it want to achieve their function. Therefore, as long as we monitor which specific API functions have been called in by processes, we can known what kind of functions that processes will implement. The realization process of determining viruses' behavioral characteristics is as shown in Figure 3.[8]

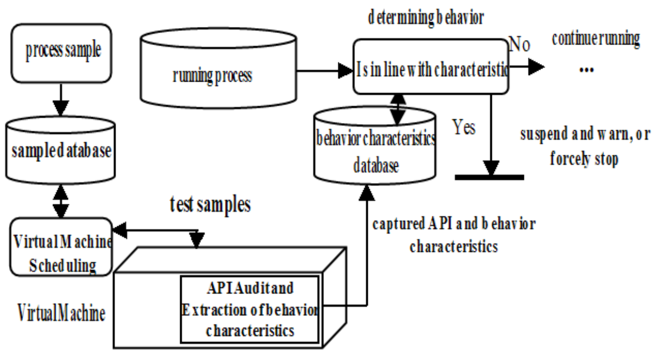


Figure 3. Determining Viruses Behaviors Model

Designing behavioral surveillance program of determination as follows:

1) The process samples come into database. Widely collect samples of processes and load the samples into the ***samples database***.

2) Virtual machine scheduling**.** Loading a group of processes samples into virtual machine to start test run, then recovering the initial state of the virtual machine and loading the next batch of test samples again after the virtual machine complete calling API functions and extracting task of the behaviors characteristics.

3) API audit and extraction of behavioral characteristics. This process is completed by the virtual machine, and it will record the call API calling and analysises its behaviors semantics.

4) determining behaviors. The process that is running in the system and the behavior semantic in characteristics database are compared. If it show that the one match to another, the current process is indentified as virus; Otherwise, process pass.

The above strategy of behavioral surveillance and behavioral determination is not only can found the known viruses, but also can relatively accurately forecast unknown viruses, which have greatly improve the ability that computer system resist viruses' attacks.

5. FEEDBACK SYSTEM

In order to strengthen ability of monitoring and determining process behavior, we can design access authority authentication system on process behavior and scoring feedback system, to make it as the third defense line of computer' active defense. Thus, we can further improve intelligent of surveillance and determination on process behavior. In this step, because the system can send out early warning for possible risk of processes, and record behavior characteristics of new high-risk process at the same time. It is possible for adding characteristics these into characteristics database in next viruses' surveillance and determination step. Feedback system model is as shown as Figure 4.

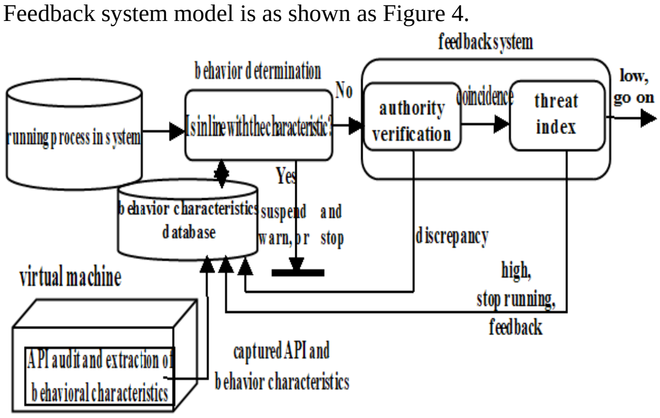


Figure 4. Model of Validation and Scoring Feedback

5.1 Verification of Access Authority

The first action is to verify access authority. As we all know, all of the operating system have set corresponding access control authority for a process. For example, a user process can not modify or delete the core files in operating system. Before software vendors formally release software, software vendors has applied some important access authority(K1, K2, ...), thus, the software will be able to carry out the corresponding operation. If the procedures that software vendors have released will call operating system' commands that is the core codes of 0 layer space of commands, having uppermost access authority that only operating system has, the operating system will trace back to digital signature of software in access authority. Only the software vendors that signature is on behalf of have this access authority, the operating system allow process continue accessing data and completing its task. Otherwise, operating system terminate it, and report risk warning of unauthorized accesses to users, as well as return their behaviors to characteristics database determining viruses' behaviors[9]. The realization of process is shown in Figure 5:

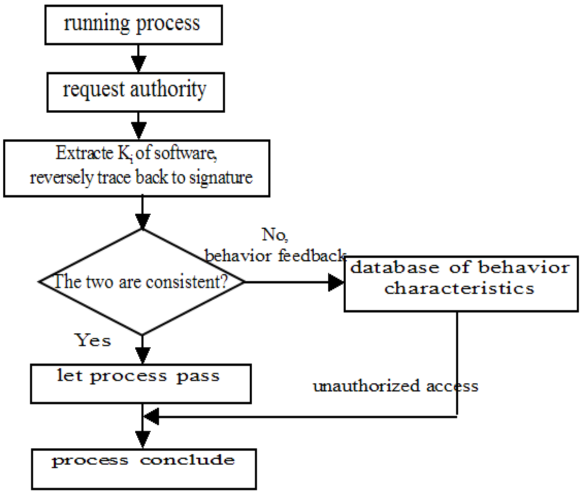


Figure 5. Verification of Access Authority

5.2 Assign Quantizing Marks

In this paper, the last resort designed for the active defense strategy is to carry out ***assigning quantizing marks*** for risk index of all processes running in the computer. The basic idea of this process is: giving a specific coefficient to the different process in accordance with their dangerous extent of practical acts. When the process running in system has too high sum of the risk index, it can be illustrated that current process is rather risk. It must inform operating system of adopting timely measures to prevent it from continuing running. For example, the risk index of operating system disk format is endowed with 10 scores; the risk index of stealing access authority that only operating system has is endowed with 8 scores; the risk index of spreading viruses is endowed with 6 scores; the risk index of garbage information into the disk is endowed with 2 scores. At the same time, we assume that sub-threshold of risk index of the process is 7 scores. If a particular process want to spread viruses into the other disks and input garbage information into them, it will be impossible to complete this illegal task. It is attributed to that the risk index of process have reached 8 scores, which is greater than threshold of the risk index. Similarly, the process that format operating system disc or to steal access authority of operating system won't be completed. In order to make it more intuitive and user-friendly for uses to operate anti-virus tools, this module can also be designed into different sections according to specific risk grade. For example, if the risk index is below 3 scores, the running process will be classified into green safe section and operating system let process pass; if the risk index is between 3 scores and 6 scores, the running process will be classified into yellow risk-warning section and operating system warn users remind users to pay attention to risk and respond to specific situation; if the risk index is above 7 scores, the running process will be classified into red high-risk section and operating system instantly terminate process and return records of process behaviors to behavioral characteristics database.

6. SUMMARY AND FURTHER WORK

The active defense technology that has discussed in this paper only is initial anti-virus technology exploration. If we want to perfect it, it is necessary that everyone needs to pay harder efforts in this field. At present, some well-known anti-virus software in the market, such as Kaspersky, Rising, Kingsoft have already begun such as unknown virus scanning, network behavior monitoring modules in their anti-virus products. These precisely are the initial attempt of active defense technology in practical application. Of course, the actual results should be in practice in further step.

In the long run, the defense technology in computer field must be an important development direction. However, active defense technology has a lot of problems at present. One of the most important problem is that intelligence is at a low level, resulting in relatively high false alarm rate. Therefore, in the next step, we should make further efforts in optimizing rules database of viruses' behaviors and improving determining viruses algorithm. With the development of artificial intelligence technology and algorithm theory, I believe that one day we can take the initiative in scanning and killing viruses to isolate risk outside system before the viruses caused actual damage.

REFERENCES

[1] Chen guisheng, Zhangzhe, The Analysis of Detection Method of Computer Virus, Journal of Shangqiu Vocational Technical College,Vol.6,No.2, Apr2007,pp39~40.
[2] Overview of Micro-point Active Defense Software, http://shop.micropoint.com.cn/product/index.htm.
[3] Li Guo, The Analysis and Comparison of Computer Virus Detection Technology, Information Design, TP393.08,2007, pp108.
[4] Overview of Micro-point Active Defense Software, http://shop.micropoint.com.cn/product/index.htm.
[5] Jin Maoshun, A Digital Signature-based Computer Virus Detection Scheme, Computer and Digital Engineering, Vol.26,No.2,1998,pp38~39.
[6] Chen guisheg, Zhangzhe, The Analysis of Detection Method of Computer Virus, Journal of Shangqiu Vocational Technical College,Vol.6,No.2， Apr2007,pp39~40.
[7] Li Xia, Guo Xueli, A Immune Principle-Based and Preliminary Exploration of Computer Virus Immune System, Computer Engineering and Application, TP393.08,2003,pp95.
[8] Chen Liang, Zhen Ning, Win32API-based Detection of Unknown Viruses, Computer Application, Vol.28, No.11, Nov.2008, pp2829~2830.
[9] Chen Lijun, The New Ways of Computer Immune Technology, Journal of Beijing University, Vol.34, No.5, Sep1998, pp581~587.

Tang Chuhua is a master candidate in school of computer science and technology, Wuhan university of technology, focuses his study on parallel computing and multimedia.

THE RESEARCH OF LAN VIRUS DETECTION MODEL BASED ON IMMUNE SYSTEM

Shouming Xu ,Qingping Guo
Department of Computer Science and Technology, Wuhan University of Technology
Wuhan , Hubei 430063, China
Email: ming_53@163.com; qpguo@whut.edu.cn

ABSTRACT

With the development of the computer virus, there are more and more problems in independent system. So we should take measures to solve the problems. And the research of LAN virus detection model based on immune system is generated and developed. In this paper, firstly we simply introduce the definition of the computer virus, virus detection and two algorithms about virus detection. Then we analyze the existing virus detection techniques and the biological immune theory. Finally we design a model of LAN virus detection based on the immune theory, and show the work flow.

Keywords: Computer Virus, Virus Detection, Biology Immune, Negative Selection, Clone Selection

1. INTRODUCTION

With the development of the Internet, the computer virus is developed quickly and spread into many fields. And the threat of the computer virus is increasing. The traditional virus detection can not defend the attack of the virus, especially the unknown virus. The Computer Virus Immune System, which simulates the biology immune technology, is the way to solve the problem. [1]The LAN virus detection model based on immune system is a kind of Computer Virus Immune System. With the help of the power server, we can solve the problem in server node and then deliver the virus disease to the client .In the process, we use the negative selection algorithm and clone selection algorithm.

2. THE DEFINITION OF THE VIRUS AND VIRUS DETECTION

2.1 The Definition of the Virus

There are two definitions on the computer virus: the first is the Fred Cohen’s definition, which is very authority; the second definition is defined by the Chinese law.

Fred Cohen’s definition: A program that can infect other programs by modifying them to include a possibly evolved copy of itself. Because of the characteristics of infection, the virus can undermine the integrity of the message in the information flow transition. Fred Cohen emphasized that virus is not the program making use of the defects of the operating system. It is norm user program, which only use the formally operate.

The definition of the virus by Chinese law: A set of codes or instructions, which is programmed or inserted into programs to destruct of computer functions and dates, affect the proper use of computer, and have the ability of self-replication. It is a strict definition, only contains malicious programs with the capacity of self-replicating.

2.2 The Definition of the Virus Detection

Virus detection is an algorithm in theory: Determine the existence of an algorithm, which can confirm another program is infected by virus. A more formal definition given by Cohen is widely accepted by people. Define an algorithm A, If and only if all of the program P, A(P) can finish in finite steps. And If P is infected, the results are true. We can determine that A can detect one virus V. As the same, A can detect an array of virus S can be defined as follow: if and only if all program P ,A(P) can finish in the finite steps and A(P) are true only when P are infected by y which is contained by the array of s. At the same time, Cohen draw a conclusion .Whether a program is infected or not is equal to the Turing Problem is not sure.[2]

Giving a random algorithm about the virus detection, you should find a program P. And it must satisfy the follow condition:
If A(P) then finish ,else outbreak
In this situation, A can not find the right result.
If A return TRUE (P are infected), then P are not outbreak (P are not infected).
If A return FALSE (P are not infected), then P are outbreak (p are infected).

There is not exist an algorithm can check out the entire virus. FALSE must exist. FALSE negative (loss many virus), FALSE positive (regular program virus positive) and no results are exist.

What we can do is to decrease the rate of the FALSE negative and FALSE positive.

3. THE BIOLOGY IMMUME THEORY OF THE COMPUTER

Biology immune system has some advantages which the computer system doesn't have. For example, the strong study capacity, identification, memory and feedback mechanism .Computer immune system simulates the biology immune system, it's an intelligent system.[3]

The primary theory of the biology immune system is that it can identify the "Self" and the "Non-self". Self is its own cells. "Non-self" is pathogens, toxicity organic matter mutant and Aging cells. The Immune cells, which can eliminate the damage to the self, can response to the "Non-self". The core problem of the Immune system is that how to definition of the "Self" and "Non-self". "Self" and "Non-self" have different definition in different domain. In the computer fields, the self represents to the legal program, such as the normally visit and the registered user and so on. But the Non-self represents to the illegal program, such as the network intrusion and the other exceptional. The identification of the Self and Non-self is carried out by a serious of the immune modules and the immune mechanism. [4]Now we corresponds the biology immune technology and the computer virus immune

technology
.
The corresponding between the biology immune system and CVIS (computer virus immune system) as Figure 3-1:

Biology Immune system	CVIS
Self	Normal system activity
Antigen	Computer virus or intrusion
B cell and T cell	detector
Antigen identify	Appetency calculate
Memory cell	Memory detector
Self tolerance	Negative selection
Cell clone	Clone selection

Figure3-1. Biology Immune System and CVIS

The primary problem in the computer virus immune system is that how to identify the Self. In computer security detection system, the complexity of the Self and Non-self is the same .The legal application program is not much easier than the illegal one. It demands that we should identify the reasonable Self. It can separate the legal program from the illegal one. It can transform to 0 and 1 problem. In virus detection the Self sometimes recognize as the Non-self and vice versa. The fine Self can reduce the two errors. It is impossible to avoid the error completely based on the biology immune system ,what we can do is to select the appropriate algorithm to improve the efficiency and the accurate .The negative selection and clone selection is the most important algorithm in immune system.

4. NAGATIVE SELECTION ALGORITHM

Define the normal system set is S, Abnormal set is N, and the total set is A. Ideally, $S \cup N = A, S \cap A = \Phi$.
Negative selection algorithm is proposed by Forrest. The principle: Create a detection string in random, which is compared with the S. If it matches any element in S, delete the string .Otherwise, we save the string as the detector to detect the exceptional. The aim of the algorithm is to find a detector set R, which is not the same with S but should match with the N. We can describe the algorithm as follows:
(1) We select a serious of strings from the protected data set whose length is L, all the strings constitute the S, That is to say, S is self. The formal data in that come from the protected file or the call process transfer.
(2) The created R is a set .The length of the R is L based on S .The string of R doesn't match with the S.
(3)Comparing the detector R with the detected set S' is to find the difference. If one of the detector R matches with the S', that is to say that S' are infected.
The process is as follows: Figure 4-1 and Figure 4-2

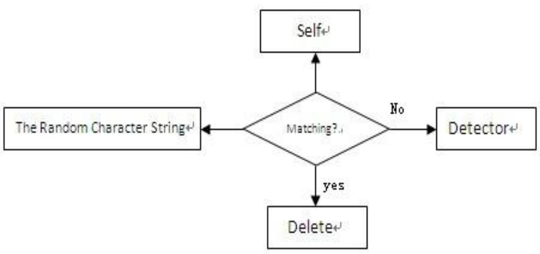


Figure4-1. The Process of Creating a Detector

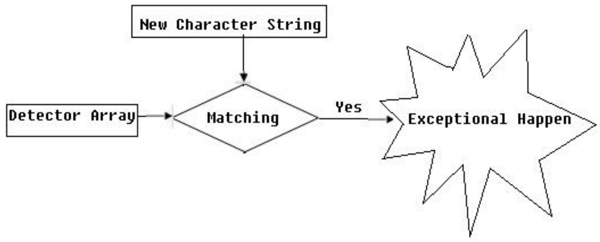


Figure4-2. Find the Abnormal

5. CLONE ALGORITHM

The clone theory in biology Immune system is formed in 1958 by Bumet. But the clone theory in computer field is proposed by Decastro and Vonzuben in 2001.Clone selection algorithm is similar to the principle of nature selection. The cells, which have higher affinity with antigen, can generate more offspring. In the process of cell division, individual cells experience a process of mutation. Then the next generation has higher appetency with antigen .We can draw a conclusion: the parent cells with higher antigen have less change in mutation.
We can describe the classical clone selection algorithm as follows:
(1)Initialize the antibody set P. Repeat the next step until fulfilling the conditions.
(2)Repeat the next steps for each antibody mode in Ag:
Calculate the appetency with P
Choose the individual cell with higher appetency, clone this based on appetency .The more appetency, the less mutate. And then add the individual to antibody P
(3)Add individual with the highest appetency to memory cell M.
(4)Update the P, Create the antibody mode instead of the individual in (2)

The Clone selection related to the calculation of the appetency and the match method of the strings. The match has two ways: Full match and partial match. If two strings have the same length and the same chars, we call it full match. But in the computer system, most of them are partial match. So we have some regular method, such as Hamming rule and R rule.

6. THE VIRUS MODEL BASED ON IMMUNE SYSTEM

6.1 The Description of the Model

We analysis the same points of the biology Immune system and the computer virus system .Then introduce the mode's work flow:

In this model, it has four steps: found the virus, deliver the virus sample, analyze and produce the antidote. The two steps are finished in the client node.[5][6] The last two steps are finished in Server.
The workflows like figure 5-1:
(1)Find the virus: Innate immune system in this model use abnormal behavior and heuristic analysis method to find the virus, and use the disinfect technology to reconstruct the infected file.
(2)Deliver the virus sample: Once an unknown virus is found by the heuristic technical, get and compress the virus sample, add with the version of the virus and the anti-virus program, some related data to form a comprehensive data.
(3)Analyze and find the antidote: When the analysis centre in

adaptive immune system received a virus sample, it first records the address of the deliver machine and other relevant information. Then, decrypt the virus sample and detect the virus with the newest feature code database .If It can identify the virus, send the corresponding antidote to the infected machine. If it can not identify the virus, we should put up the adaptive identification .First of all, we should simulate a reasonable circumstance to the virus sample with the information along with the virus sample. Run the virus and get the detail information, this is a trap program. The virus runs, when the trap program is created. We can monitor the virus activity by a monitor program. Then separate the data from the code, we find the way by Automatic Sequencing.

(4)Give out the antidote: Once the eliminating method is developed, the message is delivered to the client machine, and save it to the feature code .Update the feature code in the entire machine.

This is the process of solving the unknown virus.

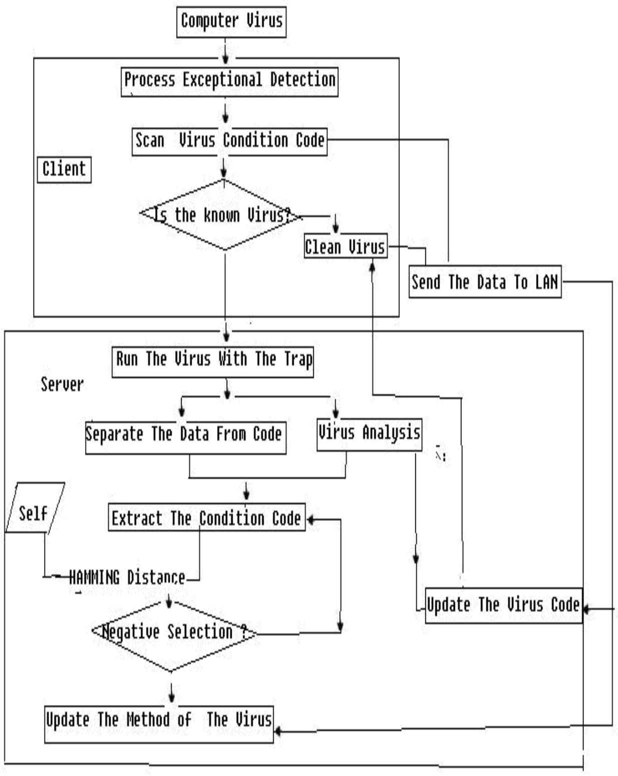


Figure 6-1. The Workflow of the Virus Detection Model

6.2 The Algorithm Code

The model has two parts. One part is the intrinsic immune system, which is built in client node. The other part is the adapting immune system, which is built in server node. The pseudocode as follows:

```
Scanknowvirus() //scan the owned virus disease
Cleanvirus() //clean the virus
Refresh() //refresh the virus disease
Buildtrap() //send the new virus to trap program
Package //the virus sample package
Cure //the virus sample character code and clean method

Intrinsicimmune() //the intrinsic immune system
{while(memory program regular); //detection abnormity
  Scankownvirus(); //check the owned virus disease
  If ture
  cleanvirus(); //clean the virus
  else send(package); // send the relative message package to
                        //server
while(1)
  {if receive(cure)
{cleanvirus();
  Refresh();
}
exit
}
}
Adaptingimmune() //adapting immune system
{while(1)
{if receive(package);
  Scankownvirus();
If true
Cleanvirus();
Else
{buildtrap();
extract the character code；
While(Self tolerance) //transfer the Negative Selection
                      //algorithm
Reextract; // Clone Selection algorithm
}
Else
{send(cure); //send to the client node and the linker node
Refresh();
}
Exit;
}
}
}
```

7. CONCLUSIONS

The LAN virus detection system has high efficiency in identifying the virus and unknown virus. Especially, used in the Distributed computing, the speed and stability of the virus detection is improved considerable. The technology should have a wilder future.

REFERENCES

[1] S.Forrest, B.javornik,R.E.Smith, A.S.Pereison, Using genetic algorithms to explore patern recognition in the immune system ,Evolutionary Computation,vol.1(3),1993

[2] Hofmeyr, Somayaji. Computer Immun-ology. Communications of ACM,1997,40(10)88-96

[3] L.Ferenc.Reduction of general virus detection problem 2001

[4] D.M. Chess and S.R. White. An Undetectable Computer Virus. Virus Bullettin Conference,2000

[5] S. Hofmeyr, S. Forrest. Architecture for an Artificial Immune System Evolutionary Computation,2000

[6] Hu Jun-Hua. A new model and implementation of network intrusion detection.Hu nan Da Xue bao,2006,33(6):119-122

AN ECONOMETRIC ANALYSIS OF COMMERCIAL BANKS' OPERATIONAL RISK BASED ON THE INCOME MODEL

Yue Ni, Shujun Ye
School of Economics and Management, Beijing Jiaotong University
Beijing, China
Email: 08120624@bjtu.edu.cn

ABSTRACT

The operational risk that China's banking industry faces is becoming more prominent as the financial environment is continuous deepening, as one of the three major risks of commercial banks with market risk, credit risk alongside, operational risk management in China is not mature enough yet. The paper will compare the different measurement models on the operational risk, then select the income model to carry out empirical analysis between China's state-owned banks and joint-stock banks. The samples are Bank of China and the Industrial Bank. The Empirical conclusion shows that Bank of China's operational risk accounts for 7.9%of the total square deviation, while Industrial Bank's is 3.2%, it means that Bank of China is facing greater operational risk than the Industrial Bank. On this basis raises the improvement of the operational risk management.

Keywords: Operational risk, Risk management, Income model, Risk measurement, Risk capital.

1. INSTRUCTION

With the accelerated process of financial globalization, as well as the increasing complexity of financial products and financial technology, the financial environment that Chinese commercial banks faced is deepening, in addition to the strengthening of credit risk and market risk control, operational risk management has got more attention, the new Basel has incorporated operational risk into the regulatory capital category management include credit risk and market risk. As a result of the special developing background of Chinese banking industry, there are serious deficiencies on structure, management, quality of personnel, as well as internal governance structure. The operational risk widely exists in our banking system, and the proportion in China's commercial banks is much larger than that of the international peer level, once deterioration, it will bring the banks great losses.

The definition of operational risk can be summed up in a narrow sense and a broad sense. The former believes that the process of control system or operator error and negligence may cause operational risk, and it may increase the risk of potential losses. The definition excludes the reputation, legal, personnel and other aspects of operational risk; therefore the scope of coverage is narrower. The latter definition includes all kinds of risks in addition to market risk and credit risk. This paper will use the latter definition.

In recent years, though big and serious cases not happened frequently in domestic commercial banks, almost all of them had connection with operational risk. And compared with credit risk, market risk, operational risk has a more important place in the financial status. However, the operational risk management in China is still in the early stage, the ways to control it include the establishment of internal control system, qualitative analysis, and to take administrative control. Judging from the current situation, the problems of the operational risk management reflect from the following aspects: (1) Not enough data has accumulated. According to the new Basel agreement, the commercial banks need to prepare corresponded capital for the operational risk in the basis of measurement, but no matter which measurement method is applied, there is the plight of less data. (2) Lack of awareness of operational risk management. In the actual management, not good at analyze the root causes of operational risk, lack of corresponding preventive and control measures, even without a special department to manage it. (3) System construction and supervision are not enough. It was difficult for the grass-roots system to accurately place the line or the implementation process, so the head office has a blind area in the internal controls of the branches because of the domestic financial information-based, electronic construction lag behind. So that big and serious cases frequently happened in the banking industry. Therefore, the research on operational risk measurement and control has important practical significance.

2. A COMPARATIVE ANALYSIS OF OPERATIONAL RISK MEASUREMENT MODELS

With more and more emphasis on operational risk, operational risk measurement model is also on the continuously developing process, in accordance with the different starting point of each measurement, the measurement models can be divided into "top-down" and "bottom-up ".

The Top-down model uses the financial indicators and earnings volatility as the measure variables of operational risk, for it is easier to collect the required data and the difficulty of the calculation is smaller. To get all the risky data with this method, and exclude the quantify of the market risk and credit risk to work out the operational risk, then the operational risk capital can be worked out. However, it's difficult for the method to tell out exactly different type of business hides what kind of the risk, therefore it's not easy to achieve the incentive effects of evaluating the performance of various types of business, revenue management and risk management. The top-down model mainly includes income model, the Basic Indicator Approach (BIA), standardized Approach (SA), operational risk of the CAPM method and the fluctuation rate model.

The Bottom-up model is a means that based on in-depth study on the operating in various business sectors of financial companies and the loss event of various operational risk, differentiate risk as the type of loss or business, then have statistics measurement step-by-step. Therefore, Bottom-up method has higher demanding and strict limitations to internal loss data. The Bottom-up model includes Internal Measurement Approach (IMA), Loss of Distribution Approach (LDA), Score Card Approach (SCA) and so on.

Because most of China's financial institutions are faced with the problem of scarcity of data, especially the low-frequency and high-impact data is difficult to obtain. Therefore, China's operational risk measurement method can only take the

Top-down model at present, and the Bottom-up model will be a long-term objective. The paper will use the basic principles of the income model, carry out calculations of banking operational risk, then get a general understanding of operational risk that China's banks faced and make some suggestions.

3. OPERATIONAL RISK MEASUREMENT BASED ON THE INCOME MODEL

Income model uses the net profit of a enterprises as the target variable, and a number of external risk factors as explanatory variables, including market factors, industry factors, as well as credit factors, takes the square deviation that can't be explained by these external factors of the net profit as the fluctuations in income caused by operational risk. These variables can be gained from macroeconomic and annual bank reports, and will not involve the internal data of banks. The paper will improve the interpretable variables, and have an inspection of the operational risk on the basis.

All tables, equations and figures should fit within the space allowed for text and separated from the text by one lines of space above and below the table. They must be numbered consecutively with Arabic numerals.

3.1 Variables and Samples Selection

In view of the structure of China's banking industry, and to make the empirical more representative, the paper will select two banks in the state-owned banks and joint-stock banks as the samples, Bank of China and the Industrial Bank which developed rapidly, with SPSS software, make use of the two banks financial data during the period of 2000-2007, and analyze the operation risk factors.

For dependent variable selection, the target variables is the net profit of the two banks, recorded as profit; for independent variables, the paper will take a moderate improvement of previous models to have a comprehensive analysis on the factors that may have impacts on the earnings of China's commercial banks, and consider primarily from the following points:

1) Market factors: Considering the two aspects, ① economic growth, economic development have a great influence on the proceed of banking industry, so real GDP (GDP / CPI) growth rate can be take as a measure of economic growth, recorded as gc; ② stock market index, as listed companies, the income of a bank is affected by the overall market conditions, so select the closing price of Shanghai Stock Exchange (SSE)Composite Index at the end of the year as representative of the stock market index, and recorded as index.

2) Industry factors: deposit and loan spreads, bank is a kind of companies that takes liabilities as a business operation, and its income is closely linked to the cost and income of capital. Therefore, select China's one-year deposit and loan spreads as a measure, recorded as LD.

3) Credit factors: uses non-performing loan ratio of banks to measure, inspect the quality of bank assets and the credit level, recorded as bl.

3.2 Model Structure

According to the analysis of the variables above, build an income model as follows:

$$profit = b_1 gc + b_2 index + b_3 LD + b_4 bl + c \quad (1)$$

In which b_i represents the sensitivity of the risk factors, c is a constant item.

Table 1. Net Profit and Operational Risk Factors of the Banks

year	net profit of Bank of China（100 million）: profit	net profit of the Industrial Bank（100 million）: profit	real GDP growth rate %: gc
2000	45.80	3.20	8.37
2001	79.08	4.82	8.24
2002	95.09	4.77	9.17
2003	287.07	9.85	9.88
2004	209.32	11.01	9.72
2005	274.92	24.65	10.22
2006	418.92	37.98	10.54
2007	562.29	79.90	11.35
SSE Composite Index: index	deposit and loan spreads %: LD	non-performing loan ratio %（Bank of China）: bl	non-performing loan ratio %（the Industrial Bank）: bl
2073.48	3.60	27.20	7.37
1645.97	3.60	27.51	4.14
1357.65	3.33	22.49	3.13
1497.04	3.33	15.92	2.49
1266.50	3.33	5.12	2.50
1161.06	3.60	4.62	2.33
2675.47	3.60	4.04	1.53
5261.56	3.33	3.12	1.15

Sources: annual reports of Bank of China and the Industrial Bank, the State Statistical Bureau, Shanghai Stock Exchange

3.3 An Empirical Analysis

Firstly, the paper has correlation tests between the selected variables and the dependent variable, build linear regression analysis respectively between the real GDP growth rate, SSE Composite Index, deposit and loan spreads, non-performing loan ratio and the net profit of the banks, the specific results are in the following tables, Tables 2 and 3 respectively show the results related to data processing of Bank of China and the Industrial Bank.

Table 2. Single-Factor Linear Regression Analysis Table of Bank of China

Risk factors	R	R Square	Sig.（F）
real GDP growth rate %	0.960	0.921	0.000
SSE Composite Index	0.743	0.553	0.035
deposit and loan spreads %	0.250	0.062	0.550
non-performing loan ratio %	0.828	0.685	0.011

Unstandardized Coefficients

Constant	Sig.（t）	B	Sig.（t）
-1319.219	0.000	161.650	0.000
39.871	0.669	0.098	0.035
1321.513	0.467	-310.231	0.550
435.352	0.001	-13.728	0.011

Table 3. Single-Factor Linear Regression Analysis Table of the Industrial Bank

Risk factors	R	R Square	Sig.（F）	Unstandardized Coefficients: Constant	Sig.（t）	B	Sig.（t）
real GDP growth rate %	0.855	0.731	0.001	-182.632	0.012	21.128	0.007
SSE Composite Index	0.907	0.822	0.002	-14.941	0.118	0.017	0.002
deposit and loan spreads %	0.177	0.031	0.674	133.929	0.616	-32.296	0.674
non-performing loan ratio %	0.623	0.389	0.099	47.744	0.021	-8.351	0.099

If take real GDP growth rate as an example, the correlation coefficient in Table 2 is R = 0.960, R^2 = 0.921, shows that this variable has high relevance with Bank of China's net profits. At the same time, Sig.= 0.000 <0.05 in F test, regression equation is adopted, and the constant term and coefficients are passed inspection, and get the fitting equation: profit1 =- 1319.219 +161.650 * gc. Similarly, the correlation coefficient in Table 3 is R = 0.855, R^2 =0.731, correlation is relatively high. And the relevant data also passes the F test and t test, the fitting equation: profit2 =- 182.632 +21.128 * gc.
Similarly analyze the variable of SSE Composite Index, Bank of China's SSE Composite Index passed the test, R^2 is 0.553, relevance has declined than the real GDP growth rate; the Industrial Bank's is 0.822, higher than the corresponding data of Bank of China. But seen from the coefficient, its impact is not great, this may be due to the existence of China's securities market has the phenomenon of split share structure, and the share of banking stocks in the stock market is still small. But it also can be seen that joint-stock commercial banks represented by small and medium-sized banks influenced much more by the market in China; at the same time, with the maturing of China's securities market and banking stocks' continuing develop and grow, the relation between the banks and the stock market will be more closely.

For the deposit and loan spreads, the results show that the two models in the analysis not pass the test, it shows that net interest income may do not significantly contribute to the net profit of China's commercial banks, this may because of the effects of other incomes such as intermediary business income in addition to net interest income influence. As China's financial liberalization and the accelerating pace of innovation, traditional commercial bank's loan-dependent survival model will also gradually change.
The results of non-performing loan ratio of the two models in the goodness of fit test are different: Bank of China's R^2 = 0.685, the Industrial Bank's R^2 = 0.389 and not pass the F test. That shows Bank of China's net profit is affected more significantly by non-performing loan ratio, while to the Industrial Bank has no significant effect. This may be related to the development context of the two kinds of commercial banks, as one of the four state-owned Banks of China, Bank of China firstly had a heavy burden of non-performing loans, with the continued divestiture of non-performing assets; non-performing loan rate had a significant change since 2004. But for the Industrial Bank, since its inception, the operation structure is well, and the non-performing loan ratio has been strictly controlled, the ratio has been constantly low and therefore does not have a significant impact on net profit.

Through the above test analysis for each variable, build a multi-variable linear regression model of net profit and the variables which have significant relationship with it from a macro perspective. Bank of China's independent variables include real GDP growth rate, SSE Composite Index and non-performing loan ratio; the Industrial Bank's are the real GDP growth rate and SSE Composite Index.
As can be seen from the regression results, in the multiple linear regression analysis of Bank of China, although the goodness-of-fit is 0.931 and passes the F test, but the coefficients of SSE Composite Index and non-performing loan ratio don't pass the t test, so they have to be removed . After calculating, the final regression equation is: profit1 =- 1319.219 +161.650 * gc.

Table 4. The Multiple Linear Regression Analysis Results of Bank of China

Risk factors	R	R Square	Sig.（F）
real GDP growth rate %	0.965	0.931	0.001
SSE Composite Index			
non-performing loan ratio %			
Constant			

Risk factors	Unstandardized Coefficients	
	B	Sig.（t）
real GDP growth rate %	128.370	0.041
SSE Composite Index	0.035	0.150
non-performing loan ratio %	-0.665	0.893
Constant	-1061.317	0.134

The results of multiple linear regression of the Industrial Bank is better, the goodness-of-fit is 0.968, indicates the good correlation, and also pass the F test and t test, the regression equation is: profit2 =- 118.286 +11.904 * gc +0.012 index .

Table 5. The Multiple Linear Regression Analysis Results of the Industrial Bank

Risk factors	R	R Square	Sig.（F）
real GDP growth rate %	0.984	0.968	0.000
SSE Composite Index			
Constant			

Risk factors	Unstandardized Coefficients	
	B	Sig.（t）
real GDP growth rate %	11.904	0.005
SSE Composite Index	0.012	0.002
Constant	-118.286	0.003

For the parameters from the perspective of economic significance: the coefficient of gc is much greater than 0, and has a greater impact on the net profit. In the model of Bank of China, each 1% increase in gc can cause Bank of China's net profit grows 16.165 billion yuan; and In the model of the Industrial Bank, each 1% increase in gc may cause the Industrial Bank's net profit grows 1.1904 billion yuan. That shows the economic growth can make great positive effect in the creation of bank's profit, and the impact on the state-owned commercial banks is greater than joint-stock commercial banks. On one hand, it's because the size of the assets of state-owned commercial banks are much larger than that of small and medium-sized joint-stock commercial banks, on the other hand, it also shows that state-owned banks still with strong economic policy quality currently, the change of shareholding system is continuing in the process. In the Industrial Bank model, the coefficient of SSE Composite Index is greater than 0, it means it has a positive effect to net profit of the bank, but in the Bank of China model, its effect is not significant, this may be because China's joint-stock commercial banks have been into the securities market earlier. But even with positive effect, the influence is small, each 1% increase in index cause only net profit grow 0.0012 million yuan, which may be related to banking stocks' share is small in China's securities market.

For the model's explanatory power: R^2 shows that the square deviation's extent of the dependent variable that can be explained by the model, the closer its value to 1, the stronger that the model can explain. In the above model, the part of the square deviation that can not be explained by the model is considered caused by the operational risk. In Bank of China's model, R^2 = 0.921, it means that 92.1% of the square deviation can be explained by the model, operational risk accounts for 7.9%of the total square deviation; in the Industrial Bank model, R^2 = 0.968, it means the model can explain the 96.8% of it, and operational risk accounts for 3.2%. This shows that the operational risk in the Industrial Bank is lower than Bank of China, only about half of its. For international practice, the industry generally believes that the operational risk can account for 10% -20% in the total proportion, though the operation risk of the two banks are in a reasonable level, operational risk is sudden and uncontrollable, it needs to see the gaps between the state-owned banks and joint-stock banks in operational risk management and to improve the risk control.

3.4. The Measurement of Operational Risk Capital

Because the fluctuations in income that arising from operational risk are as follows:

$$\delta^2 = \sigma^2 * (1 - R^2) \tag{2}$$

It can assume that the volatility of income subject to normal distribution, then 3.1 times of the standard deviation that under 99.9% of the confidence interval is operational risk, that is:

$$OpRisk = 3.1\delta \tag{3}$$

By calculating the available estimate operational risk capital of the two banks are as the following table:

Table 6. The Computation of Banks' Operational Risk Capital

	Bank of China	The Industrial Bank
The total square deviation of net profit σ^2	224522.9	4828.657
R^2	92.1%	96.8%
corresponding variance of operational risk δ^2	17737.309	154.517
the corresponding standard deviation of operational risk δ	133.181	12.430
the estimated value of operational risk under confidence level of 99.9% $OpRisk$	412.863	38.535

It can be seen that in the 99.9% confidence level, the estimated value of the operational risk of Bank of China is 41.2863 billion yuan, and the Industrial Bank's is 3.8535 billion yuan, less than 10% of Bank of China. So though the relative value of the ratio of operational risk for the two banks has no apparent gap, the bank's operational risk capital gap magnifies the gap, operational risk can not be neglected.

4. THE CONCLUSIONS AND SUGGESTIONS

4.1. Conclusions

After the above analysis, the paper gets the following conclusions:

Firstly, the income model can explain the operational risks to a certain extent. After the empirical analysis, it can be seen that the operational risk Bank of China faced is higher than that of the Industrial Bank, especially to the operational risk capital, the gap becomes more apparent. This is in line with the operational risk cases happened so far, it shows that the state-owned commercial banks similar to Bank of China still exist certain deficiencies on risk control.

Secondly, the impact of market factors is greater than the industry and credit factors to income. At present, China's commercial banks are focus on credit risk management which is much higher than the degree of importance of market risk and operational risk, so it is necessary to strengthen attention to market factors, and improve the income level.

The third is the lack of model accuracy. For the current research on operational risk has very few historical data, and models can quoted is limited, it can only use the income model which required less internal data of banks to have a general study, the accuracy needs to be strengthened. But through the analysis on measurement of operational risk, it is possible to reflect the problems on the operational risk of China's commercial banks faced to a certain extent.

4.2. Suggestions

Therefore, we should pay attention to the following aspects in future operational risk management:

Firstly, to improve the internal control, and pay attention to the operational risk. Many institutions of state-owned banks, the chain for operation of internal control bodies is too long and some other problems affect operational risk to a certain extent. Facts have proved that bank's operational risk mostly come from internal fraud and external fraud, so it's need to unify and improve the overall standard of business practices and the internal controls of banking system, improve relevant laws and regulations, strictly enforce business operational rules, to resolve operational risk from the source.

Secondly, to strengthen the operational risk data collection, and establish the database. Operational risks are low-frequency and high-risk events, and require high level data of loss, but related data are all the internal data of banks. So banks can strengthen the collection of internal risk in order to provide more accurate tools for risk measurement and control. Therefore, it needs to speed up the establishment of relevant databases, and prepare for the accurate measurement of operational risk.

The third is the establishment of a risk assessment system. At present, the research of risks in China is mostly on the qualitative research phase, lack of the supporting of quantitative research. And risk identification, measurement, prevention is a comprehensive qualitative and quantitative study with interaction process, therefore, a risk assessment system can promote the combination of risk management study on qualitative and quantitative, and contribute to the development of risk management in China .

REFERENCES

[1] Liu Gui-rong and Zhao Yan, An Empirical Analysis of Commercial Bank Operational Risk Based On the Income Model, Journal of East China University of Science and Technology: Social Science.J., March,2008, pp54-58.

[2] Wang Jun-peng, An Empirical Study of Operational Risk of commercial banks based on the income Model, Contemporary economics.J., November,2007, pp146-147.

[3] Dai Gui-xia, The Research on Measuring to Operational Risk of Commercial Bank, Taxation and Economy.J, March,2007, pp41-45.

[4] Wu Hai-jun, Appficafion of income model in operational risk measurement for Chinese commercial bank, Economic＆Trade Update.J., June.2007,pp112-113.

[5] Zhang Xue-tao and Tong Jing, The Empirical Analysis of Commercial Banks'Operational Risk and the Calculation of Venture Capita, The Theory and Practice of Finance and Economics.J.,Vo1.27. No.141.May.2006, pp33-37.

[6] Ruan Ming-zhen, A Number of Countermeasures on Operational Risk of China's Commercial Banks, New Finance.J., January,2006, pp61-62.

[7] Dong Yan-wen, Reflection on the Measuring of Operational Risk in Chinese Commercial Banks, Business Economics and Administration.J.Vol.167. No.9. September.2005,pp58-62.

[8] Wang Xu-dong, The New Basel Capital Accord and The Quantify Management of Operational Risk on Commercial Banks, The Financial Forum.J., February.2004, pp57-61.

[9] Li Hui-min and Ding Xin-wei, The Impact of New Basel Capital Accord Risk Measures on China, The Theory and Practice of Finance.J., November,2003,pp42-44.

[10] Fan Xin and Yang Xiao-guang, The Measures and Status Quo of Operational Risk Management, Securities Market Herald.J., June.2003, pp64-69.

Ni Yue a graduate student of School of Economics and Management, Beijing Jiaotong University, enrolled in 2008 with the specialty of finance, and graduated from the University with bachelor degree of Finance in 2008. The main research orientation is the application of finance, especially on the risks management of commercial banks. During graduate study session, participated several scientific research projects leaded by tutor Professor Ye. The paper "The Economic Effects of China's International Capital Flows and An Empirical Study On Its Influencing Factors" won the first prize in the 19th "Huiguang Cup" post-graduate academic and cultural festival paper contest of Beijing Jiaotong University, 2009.

Ye Shujun is a finance professor, doctoral tutor and member of Chinese Finance Association, now dean of Finance Department, Beijing Jiaotong University. Got the degrees of bachelor, master and doctorate separately from Xi'an University of Technology, Tsinghua University and Beijing Jiaotong University. 1997, engaged in academic exchanges in Duquesne University of Pittsburgh, the United States; 2001, went to Germany to survey the railway's investment and financing. The main research orientation are: financial institutions, theory and development strategies, financial risk management theory and evaluation, microstructure theory on financial market. Has taken charge and participated more than 30 scientific research projects from National Natural Science Foundation of China, China Insurance Regulatory Commission, the Ministry of Commerce, the Ministry of Railways and so on. Published more than 40 academic articles on specialized academic journals, some of which can be searched by international information search client. Published a monograph, edited and published 8 teaching materials and 1 translation.

RESEARCH OF ACTIVE PREVENTION SYSTEM BASED ON KEY RESOURCES

Xi Li-ya[1], Zhang Zhi-zhong[2] , Bai Ping[3]
[1] Dept. of Computer & Electronic,Huazhong University of Technology, Wuchang Branch
Wuhan 430064, Hubei, China;
[2] Information & Technology Division, Hankou Bank
Wuhan 430015,Hubei,China
[3] Business Administration College OF WUSE
Wuhan 430077, Hubei, China
Email: panhao@whut.edu.cn

ABSTRACT

The article introduces key resources and the idea of active prevent into the design of security system, it proposes the concepts of active prevent system entity(APSKR Entity) which bases on key resources and APSKR engine, designs a security model of active prevent system based on key resources and has a detailed design of Security Command Central.

Keywords: Key resources, Active prevent, Security APSKR.

1. INTRODUCTION

With the vast number of Internet users and growing maturity of network security technology, users become more aware of the importance of network security. In order to defend against network attacks, many users have installed Anti-virus, Firewall or Intrusion Detection Systems. However, those preventing intruders can not stop the pace of intruder. The common problem faced by the Internet users is there are not effective returns in the huge investment in security. There is an analysis of the security measures (such as: Firewalls, Intrusion Detection Systems, etc.) used by the common customers, which has shown that most security measures are carried out during the outlying guard. They guard against network attacks passively with the pre-configured strategy, which can only defense the current attacks. Nowadays, with the continuous advancement of network security technologies, network attacks have also shown the new unknown characteristics, which disturb the unresting internet world. "Zero-day Attack" with the dynamic characteristics is the root of our security problem. We make our major focus on a potential、new target of attacks so that we can find a way to destruct them. When the threats and crises bought by the attack like Zero-day Attack, we should take the initiative measures to protect the system and protect the system resource actively [1].

Having Studied all network attacks, we find that the ultimate goal of them is to obtain or damage the key resources of the system [1]. An effective secure solution is to protect the key resources at first. In fact, the most important key resources are the sensitive data and the key documents, key equipment and critical processes, which often reside in the host and the server. As a result, the design of Active Prevention System is based on the key resources.

2. ACTIVE PREVENTION SYSTEM BASED ON KEY RESOURCES

A single security protection technology is not sufficient to protect the security of a system. In order to maximize the protection of a system, a wide range of security methods should be used in a cooperative way.

We ultimately arrange the passive safety tools, active security protection tools, host's protections based on key resources, and construct a active prevention systems based on key resources. The systems use the active security measures to take evidence of the intent attack, generate the safety event and finally at the edge of security gateway (firewall) to block the attack from the network security border, which is a success paradigm to prevent the unknown attack.

Active prevention system based on key resources is a complete security system. It mainly utilizes a host security technology, authentication technology, access control technology and security management technology etc. Its overall structure is shown in Fig. 1.

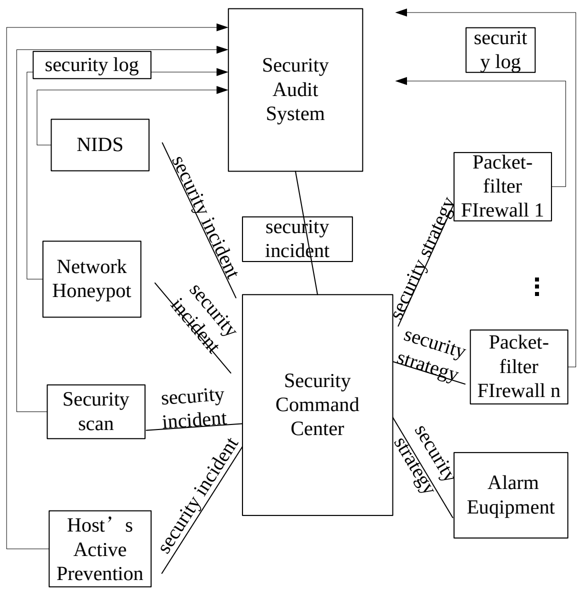


Figure1. Active prevention system based on key resources.

In the active prevention system, Security Command Center is a bridge to connect all security processing units, which plays a command and coordination role. These security units which are situated in the left side of Security Command Center include Intrusion Detection System, Network Traps, Security Scanning and Host's Protection System based-on Key Resource, etc. They are the monitoring and protection entities of the system and mainly used to monitor the security status of the current situation and to prevent all kinds of unknown attacks. The security units located in the right side of Security Command Center are the executing entities. They are used to carry out the policies of network security and to cut off the known attacks, such as the simple packet-filter firewall or the alarm equipment sending safety alarm to Security Command

Center. In an active prevention system, all the security units are not necessarily needed and we can make different selection according to our actual needs.

Security Command Center is the core of the entire security model, which realizes the centralized security management. When it received the security incident from the security processing units (such as Intrusion Detection System, host protection initiatives based on the key resource, etc.), it matches those events with the security policy that have been pre-defined. If matching successfully, Security Command Center starts the corresponding security measures in accordance with the matched security policies.

In the entire security model, security audit system is designed as an independent part. It coordinates and complements together with other safety processing units (such as: firewalls, intrusion detection systems, vulnerability scanning systems, etc.) to protect the overall network security and is an important part in the safety system. Safety Audit mainly relates to collection the safety processing unit logs and analyzes the safety strategies based on time and space and send the results to the Security Command Center in the form of security incidents.

In the Active Prevention System based on key resources, there are a variety of network security processing units. These security incidents are different in configuration of the safety processing units, security strategy and management, maintenance and log. We have adopted the form like MIB (Management Information Base) in the SNMP to organize the information uniformly and standard to realize the flexibility and scalability of the system management. At the same time, in order to protect the security of the control information being transmitted between security processing units, we use security technologies such as identify authentication between the communication entities, confidentiality and integrity between the information of exchange.

The Authentication methods includes the shared key based on a simple three-way handshake (such as: Chap agreement), the shared key based on trusted third party certification (e.g. Kerberos Protocol) and the public key-based authentication system (e.g. RSA). Here we select the shared key based on a simple three-way handshake as an example.

Confidentiality is to use the symmetric encryption of the shared key and symmetric encryption of the public-key cryptography [2][3].

Integrity is mainly use [4]:

(1) based on the digital signature of RSA: m+E(m,Private);
(2) encoded MD5: m+MD5(m+k) OR m+MD5(m+k)+E(k,Private);
(3) MD5 have the RSA digital signature: m+E(MD5(m),Private).

PS: m represents message; private represents private key; k represents shared key; E(m, Private) represents encode message by private key; MD5(m+k) represents generate message digest code of message and shared key based on MD5 algorithm.

3. THE REALIZATION OF THE ACTIVE PREVETION SYSTEMS BASED ON KEY RESOURCES

The active prevention systems based on key resources mainly includes the host's active prevention based on key resources, the simple packet-filter firewall that distributed on network security border, the centralized Security Command Central etc. security processing and coordination units.

3.1 APSKR entity

Based on the principle of simple, safe, efficient and convenience, we abstract each safety processing unit and design a Active Prevention System based on Key Resources entity in light of the SNMP v3 Protocol, the structure of the entity is shown in Fig. 2. APSKR entity is made up by the APSKR engine and the application program above the APSKR engine. APSKR engine is the base of the security entity, the message processing subsystem mainly responsible for the accordance of the information exchange between the security entities; security subsystem is responsible for the communication security between the security entities; access and control subsystem is responsible for the access control.

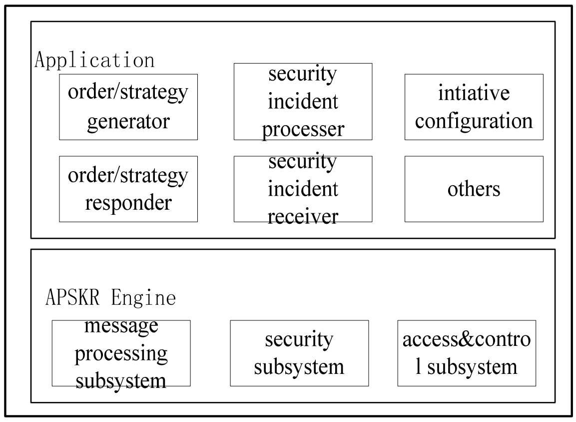


Figure2. APSKR entity.

We arrange and organize the Firewall APSKR entity, HAP APSKR entity and APSKR entity to construct a active prevention system based on the key resources, which is shown in Fig. 3.

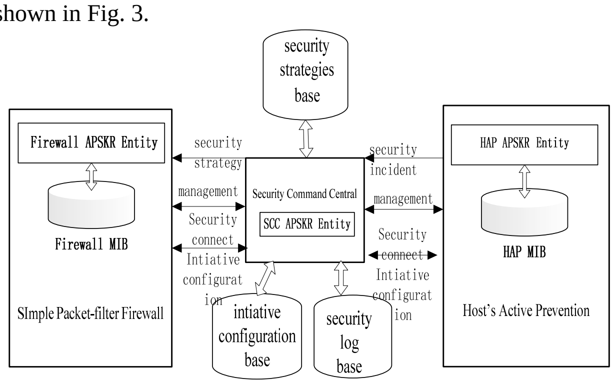


Figure 3. The specific realization of APSKR.

In the entire system, Security Command Central maintain the security strategies base, configuration base and security log base, and it is responsible for host's active prevention based on the key resources, the identification between the simple packet-filter firewalls, the construction of the secure connection, initial configuration; it also is responsible for the management of the HAP and packet-filter firewall; receives the security incidents from the HAP, generate security strategy in accordance with the security strategy base and send the security strategy to the packet-flitter firewall. The centralized design, simple, accordance, visual and comprehensive security incident management function of the SCC, effectively decrease the operate difficulty of the system and the management expense, and provide a strong support to construct a complete security management system.

In the system, the packet-filter firewall distributed around the security domain border, filter the internet data by the prepared initialized security strategy. On the other hand, HAP protects the key resource in the host, monitoring and scanning the security violation activities that violate the security strategy. When HAP detects a security violation actability, it stops the progresses related to this activity. At the same time, the security incident generator in the HAP will generate a security incident that include the specific information about this activity and send the incident to SCC. When SCC receives the security incident, the strategy generator will match the security incident. If match succeed, SCC will generate corresponding security strategy and send the security strategy to the packet-filter firewall to block the new and sudden attack; if match failed, SCC will send alarm to the security alarm equipment and remind the security management person to handle it and to protect the safety of the system.

3.2 The implementation of SCC

SCC is made up by the several parts: APSKR engine, security incident receiver, security incident processer, order/strategy generator and initial configuration. The process of the system handling the internet security incident is shown in Fig. 4.

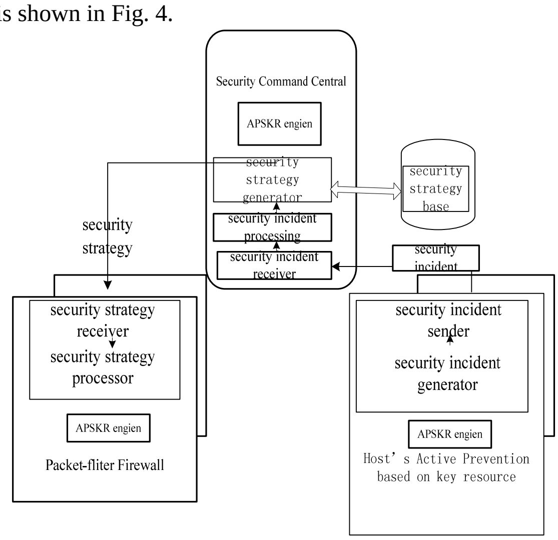


Figure 4. The procedure to handle network security incident.

The process SCC handle initial configuration and integral management of the APSKR entity is shown Fig. 5.

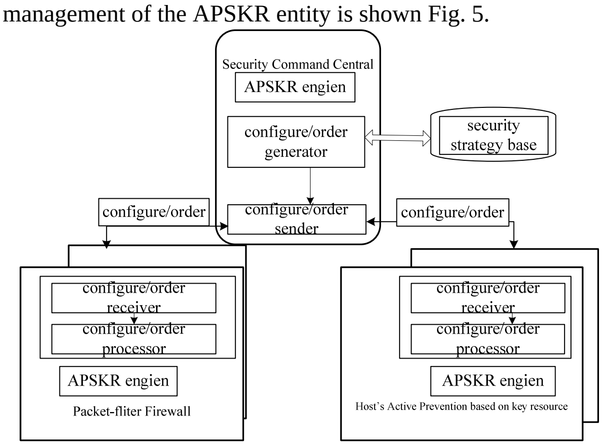


Figure5. The procedure to initialize configuration and integrate management.

The process SCC handles the security incident is shown in Fig. 6. In the entire process, the security incident receiver of the SCC locate at the port to monitor the security incident form HAP。Receiving a alarm incident, SCC will immediately start a new process to handle this incident; then returns to the monitor situation in order to receive next security incident. The new handle process will extract incident abstract (such as the time incident happens, origin address, destination address, attack type etc.), then the abstract will be write into the security log base. Next, according to the attack type in the security incident, strategy generator will match the attack type with the same type strategy in the whole security strategy base, in order to determine security type it triggered. If it finds the matched security strategy, it will read the security activity in the security strategy. The security activity means a security rule that includes origin address, destination address and port number to the firewall. So SCC will send the security rule to the destination firewall to update the security rule of the firewall and to block the internet attack at the internet security border. If match failed, it will send the security incident to the security administrator at the remote end by the means of RPC (Remote Procedure Call) etc, in order to remind the administrator to pay attention to this incident or to handle this incident by using artificial intervention method.

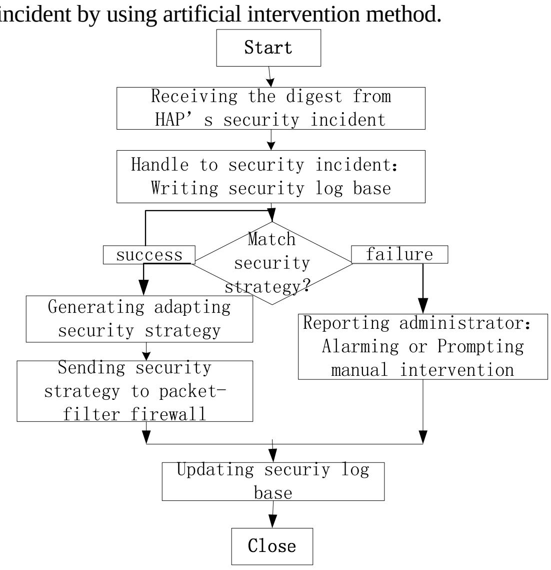


Figure 6. The procedure to handle the security incident.

4. CONCLUSION

This essay starts from the deficiency in current security prevention measures, and it makes a detail analysis in "Zero-day Attack", reexamine the new type of the internet attack and propose a solution of active prevention system based on key resources. In the essay, it explains each security processing unit and Security Control Center in the system, constructs an application framework of prevention system based on key resource with active prevention mind to set up a safer prevention system and to provide certain theoretical basis and technical support.

REFERENCES

[1] Ying Xiang-rong. Importance of Active Defense System under New Trends based on Network Attacks [J].Beijing: Computer Security,2003.7.(29), 53-55.

[2] Feng Deng-guo. Computer Communication Network Security [M].Beijing: TsingHua University Press.2004.

[3] Ramachandran,J. Design of Security Architecture[M]. Hu Jun. Beijing: China Machine Press,2003.

[4] PETERSON,L.L., DAVIE,B.S .Computer networks: a systems approach, fourth edition[M]. Beijing: China Machine Press, 2007.

ON SOME ATTACKS OF DISCRETE LOGARITHM PROBLEMS OVER FINITE FIELDS

Minglong Qi, Qingping Guo, Luo Zhong
College of Computer Science & Technology, Wuhan University of Technology,
Mafangshan Campus West, 430070 Wuhan China
Email: mlki01@sohu.com

ABSTRACT

The intractability of the discrete logarithm problem over finite fields is the basis of the security of recently mainstream public key cryptosystems such as ElGamel, DSA, Diffie-Hellman and ECC (Elliptic Curve Cryptosystems), etc. In this paper, one has presented, illustrated and analyzed some attack algorithms on the DLP over finite fields, namely Shanks's giant-steps baby-steps algorithm, Pollard field splitting ρ-algorithm, and Pohlig-Hellman Algorithm. One equally implemented these algorithms for the finite field $GF^*(3049)$ using a high-level programming language.

Keywords: Discrete Logarithm Problems, Finite Fields, Shanks's Giant-steps Baby-steps Algorithm, Pollard ρ-algorithm, Pohlig-Hellman Algorithm.

1. INTRODUCTION

Since the work of Differ and Hellman[1] about key-agreement protocol, researches in the domain of the public-key cryptosystems have been dramatically stimulated and developed. Public-key cryptosystems have many of advantages against symmetric or private-key cryptosystems: easily key generating and management, providing more robust security, easily being implemented, smaller key size, etc. A common feature of the modern public-key cryptosystems is that, their security depends on intractability of some specific number theory problems. RSA[2] based on hardness of big prime integer factorization, ElGamel[3] based on the intractability of the DLP over finite fields, and ECC[4, 5] based on the intractability of the Elliptic Curve DLP over finite fields. Finding efficient algorithms of solving the DLP over finite fields becomes the mission of first priority both for crypto analysts and hackers to breaking some public-key systems. Among all algorithms found until today, one cites the most efficient ones: Shanks's giant-steps baby-steps algorithm [6], Pollard ρ-algorithm [8], Pohlig-Hellman algorithm [9], index-calculus method [10, 11], Gaussian integer method [12, 13], number field sieve method [14], etc.

What is the DLP over a finite field? One gives a generalized definition of the DLP over the finite field $GF(p^n)$:

Definition 1.1 let $GF^*(p^n) \equiv GF(p^n) \setminus \{0\}$ be the multiplicative group of the finite field in consideration, and let g be a primitive root of the group. Given an element $y \in GF^*(p^n)$, find an smallest integer $x \in [0, p^n - 2]$ such that $y \equiv g^x (\text{mod } p^n)$, is called the discrete logarithm problem over the finite field $GF(p^n)$. Algorithms breaking the DLP over finite fields are called attacks on the DLP over finite fields.

In this paper, one presents, illustrates and analyzes some of these attacks on the DLP over finite fields. One is limited to consider Shanks's algorithm, Pollard ρ-algorithm, and Pohlig-Hellman algorithm. Other attacks need a very sophisticated mathematical background, so that they are out of the scope of the present short note. The remainder of the article is organized as follows: in section 2, above cited attacks will be explained and implemented in a high-level programming language for $GF^*(3049)$; in section 3, one will conclude his short introductory note.

2. ATTACKS ON THE DLP OVER FINITE FIELDS

2.1 Shanks's Baby-Step-Giant-Step Algorithm

In [6, 7], Shanks proposed a relatively efficient and practical algorithm for solving the DLP over finite fields. If the order of the multiplicative group of the finite field in consideration is equal to n, then the complexity of the algorithm is about $O(\sqrt{n})$. The key observation Shanks has made is that, every element of the group can be represented in a basis such as $\{1, m\}$ with $m = \lfloor \sqrt{n} \rfloor + 1$, so that if $x \in GF^*(p)$ then $x = x_1 m + x_2$ with $0 \le x_1 \le m$ and $0 \le x_2 \le m$. Suppose that g is a primitive root of $GF^*(p)$, such an element of the group that for every element in the group y, there is a non negative integer that lies in the interval $[0, p-2]$ such that the next identity holds $y = g^x (\text{mod } p)$. Substituting x with $x = x_1 m + x_2$ and rearranging the formula, one has $yg^{-x_2} = g^{x_1 m}$, whereby constructing two sets: the first being $GSteps = \{g^{x_1 m} \mid 0 \le x_1 \le m\}$ (the giant steps) and the second being $BSteps = \{yg^{-x_2} \mid 0 \le x_2 \le m\}$ (the baby steps), one could expect that at certain moment, two sets must have an identical entry corresponding an index i in the first set and an index j in the second set, we finally find the logarithm of y: $\log_g y = im + j$.

Remark 2.1.1 Gauss referred to the discrete logarithm as the index of a number. In any textbook of elementary number theory, one can find the following formulae:

$AB \equiv C \ (\text{mod } p) \Leftrightarrow$

$\log_g A + \log_g B \equiv \log_g C \ (\text{mod } p-1)$

$A/B \equiv C \ (\text{mod } p) \Leftrightarrow$

$\log_g A - \log_g B \equiv \log_g C \ (\text{mod } p-1)$

Remark 2.2.2 When computing the set BSteps, one can exploit identity $yg^{-x_2} \equiv yg^{-x_2+p-1} \pmod p$.

Notice that Shanks's algorithm is easy to be implemented with any high-level programming language. One can use the fast algorithm of exponentiation modulo a prime for generating all group elements. In this paper, for modular exponentiation operation, we give a recursive one in **Algorithm 2.2.1**.(In Visual Basic style)

Algorithm 2.2.1 Fast algorithm for modular exponentiation (recursive version)

INPUT: g as a primitive root, p as modulus, n as exponent.
OUTPUT: powMod $\leftarrow g^n \pmod p$.

```
Function powMod(ByVal g As Integer, ByVal n As Integer, ByVal p As Integer) As Integer
n = n Mod (p-1)
If n < 0 Then n = n + p - 1
If n = 0 Then powMod = 1
Else If n = 1 Then powMod = g Mod p
Else If n = 2 Then powMod = g ^ 2 Mod p
Else If n > 2 And n Mod 2 = 1 Then
tg = g * g Mod p
powMod = g * powMod(tg, (n - 1) / 2) Mod p
ElseIf n > 2 And n Mod 2 = 0 Then
tg = g * g Mod p
powMod = powMod(tg, n / 2) Mod p
End If
End Function
```

Example 2.1.1 Let the finite field in consideration is $GF^*(3049)$. It is easy to check $g = 11$ is a primitive root. One want to compute $\log_{11} 2009$, that is to find a x such as $2009 \equiv 11^x \pmod{3049}$.

Solution:

The basis $= \{1, m\} = \{1, \lfloor \sqrt{3049} \rfloor + 1\} = \{1, 56\}$ **;**

$GSteps = \{g^{mi} \mid 0 \le i \le m\} = \{11^{56i} \mid 0 \le i \le 56\} =$ **{** (0,1),(1,2031),(2,2713),(3,560),(4,83),(5,878),(6,2602),(7,745),(8,791),(9,2747),(10,2536),(11,855),(12,1624),(13,2375),(14,107),(15,838),(16,636),(17,1989),(18,2783),(19,2476),(20,955),(21,441),(22,2314),(23,1225),(24,3040),(25,15),(26,3024),(27,1058),(28,2302),(29,1245),(30,974),(31,2442),(32,2028),(33,2718),(34,1568),(35,1452),(36,629),(37,3017),(38,2086),(39,1605),(40,374),(41,393),(42,2394),(43,2108),(44,552),(45,2129),(46,517),(47,1171),(48,81),(49,2914),(50,225),(51,2674),(52,625),(53,991),(54,381),(55,2414),(56,42)}

$BSteps = \{2009 \times 11^{-j} \mid 0 \le j \le 56\} =$ {(0,2009),(1,737),(2,67),(3,1392),(4,2344),(5,1599),(6,2640),(7,240),(8,299),(9,2799),(10,1086),(11,2039),(12,2680),(13,798),(14,2290),(15,2980),(16,1934),(17,453),(18,2813),(19,2196),(20,754),(21,2286),(22,485),(23,1430),(24,130),(25,289),(26,1135),(27,2875),(28,2756),(29,2468),(30,2719),(31,3019),(32,1106),(33,2318),(34,2151),(35,2413),(36,2714),(37,2187),(38,476),(39,1152),(40,2045),(41,1849),(42,1554),(43,1250),(44,668),(45,2001),(46,1845),(47,2108),(48,746),(49,345),(50,2526),(51,784),(52,1180),(53,1216),(54,2328),(55,766),(56,624)**}.**

One find that in the set of giant steps, when $i = 43$ and in the set of baby steps when $j = 47$, there is an identical entry equal to 2108, so one can conclude that $\log_{11} 2009 = 47 + 43 \times 56 = 2455$, that is, $2009 \equiv 11^{2455} \pmod{3049}$.

In [7], Buchmann et al. have improved the Shanks's algorithm with a run time faster of order $O(\sqrt{x} + \log \sqrt{x})$ using a table of $O(\sqrt{x})$ entries.

2.2 Pollard ρ-Algorithm

In [8], Pollard, in exploiting the cyclic intrinsic nature of the multiplicative group of finite fields, proposed an algorithm of solving the DLP. Starting by constructing a sequence of group elements by some way probabilistic, one could expect that this sequence must be repetitive because of the cyclic nature of the group. The remainder problem is: how should –it be the generating term of the sequence? Since one want that the sequence should be rapidly "converse", so that the genetic term of the sequence must behavior as giant-steps as in Shanks's algorithm. Because one wants to compute $x = \log_g y \Leftrightarrow y \equiv g^x \pmod p$, why not to take terms $y^\alpha g^\beta$ as the generating term of the sequence? He has proceeded as fallows:

Step 1: Dividing the group in three or more distinct subsets, say G_1, G_2 and G_3, such as $G_1 \cap G_2 = \Phi, G_2 \cap G_3 = \Phi, G_1 \cap G_3 = \Phi$.

Step 2: Defining the generating term as $\gamma_i = y^{\alpha_i} g^{\beta_i}$.

Step 3: Constructing the sequence of group elements according to the next iterative formulae:

$$\alpha_{i+1} = \begin{cases} \alpha_i & \gamma_i \in G_1 \\ 2\alpha_i \pmod{p-1} & \gamma_i \in G_2 \\ \alpha_i + 1 \pmod{p-1} & \gamma_i \in G_3 \end{cases}$$

And

$$\beta_{i+1} = \begin{cases} \beta_i + 1 \pmod{p-1} & \gamma_i \in G_1 \\ 2\beta_i \pmod{p-1} & \gamma_i \in G_2 \\ \beta_i & \gamma_i \in G_3 \end{cases}$$

α_0 and β_0 are randomly taken from the interval $[1, p-1]$.

Step 4:Since G is cyclic there will eventually be two elements, γ_i and γ_{i+T} for some $i \ge 0$ and $T > 0$ such that $\gamma_i \equiv \gamma_{i+T} \pmod p$. By consequence:

$$y^{\alpha_i} g^{\beta_i} \equiv y^{\alpha_{i+T}} g^{\beta_{i+T}} \Rightarrow$$

$$g^{\beta_i - \beta_{i+T}} \equiv y^{\alpha_{i+T} - \alpha_i} \Rightarrow$$

$$g^{\beta_i - \beta_{i+T}} \equiv g^{x(\alpha_{i+T} - \alpha_i)} \Rightarrow$$

$$\beta_i - \beta_{i+T} \equiv x(\alpha_{i+T} - \alpha_i) \pmod{p-1}$$

Step 5: Applying the Extended Euclidian Algorithm to compute x : if $\gcd(\alpha_{i+T} - \alpha_i, p-1) = 1$ then $x \equiv (\beta_i - \beta_{i+T})(\alpha_{i+T} - \alpha_i)^{-1} \pmod{p-1}$,else if $\gcd(\alpha_{i+T} - \alpha_i, \beta_i - \beta_{i+T}, p-1) = d$, then

$z \equiv$

$(\beta_i - \beta_{i+T}/d)(\alpha_{i+T} - \alpha_i/d)^{-1}$

$(\bmod\ p-1/d)$

$x \equiv z + k(p-1/d)\ \ 0 \le k < d$

Example 2.2.1 Let the finite field in consideration is $GF^*(3049)$. It is easy to check $g = 11$ is a primitive root. One want to compute $\log_{11} 2009$, that is to find a x such as $2009 \equiv 11^x \pmod{3049}$ using Pollard ρ-algorithm.

Solution:

a) Dividing the group $GF^*(3049)$ in three subsets, $G_1 = [1, 1016], G_2 = [1017, 2032], G_3 = [2033, 3049]$

b) Defining the generating term as $\gamma_i \equiv 2009^{\alpha_i} 11^{\beta_i} \pmod{3049}$

c) According to the iterative formulae to computing the tuple $(i, \gamma_i, \alpha_i, \beta_i)$. Using the Extended Euclidian Algorithm, and the fast exponentiation algorithm of generating group element, it is straightforward to implement Pollard rho-Algorithm. Computational results for this example are shown in **Table 2.2.1**.

Table 2.2.1 Computational results $\gamma_i \equiv 2009^{\alpha_i} 11^{\beta_i} \pmod{3049}$ with the up bound $m = \lfloor \sqrt{q} \rfloor + 1 = \lfloor \sqrt{3049} \rfloor + 1 = 56$

$(i, \gamma_i, \alpha_i, \beta_i)$	$(i, \gamma_i, \alpha_i, \beta_i)$
(0, 2943, 2010, 124)	(29, 1474, 638, 2336)
(1, 476, 2011, 124)	(30, 1788, 1276, 1624)
(2, 2187, 2011, 125)	(31, 1592, 2552, 200)
(3, 74, 2012, 125)	(32, 745, 2056, 400)
(4, 814, 2012, 126)	(33, 2097, 2056, 401)
(5, 2856, 2012, 127)	(34, 2204, 2057, 401)
(6, 2535, 2013, 127)	(35, 688, 2058, 401)
(7, 985, 2014, 127)	(36, 1470, 2058, 402)
(8, 1688, 2014, 128)	(37, 2208, 1068, 804)
(9, 1578, 980, 256)	(38, 2626, 1069, 804)
(10, 2100, 1960, 512)	(39, 864, 1070, 804)
(11, 2133, 1961, 512)	(40, 357, 1070, 805)
(12, 1352, 1962, 512)	(41, 878, 1070, 806)
(13, 1553, 876, 1024)	(42, 511, 1070, 807)
(14, 50, 1752, 2048)	(43, 2572, 1070, 808)
(15, 550, 1752, 2049)	(44, 2142, 1071, 808)
(16, 3001, 1752, 2050)	(45, 1139, 1072, 808)
(17, 1136, 1753, 2050)	(46, 1496, 2144, 1616)
(18, 769, 458, 1052)	(47, 50, 1240, 184)
(19, 2361, 458, 1053)	(48, 550, 1240, 185)
(20, 2054, 459, 1053)	(49, 3001, 1240, 186)
(21, 1189, 460, 1053)	(50, 1136, 1241, 186)
(22, 2034, 920, 2106)	(51, 769, 2482, 372)
(23, 646, 921, 2106)	(52, 2361, 2482, 373)
(24, 1008, 921, 2107)	(53, 2054, 2483, 373)
(25, 1941, 921, 2108)	(54, 1189, 2484, 373)
(26, 1966, 1842, 1168)	(55, 2034, 1920, 746)
(27, 2073, 636, 2336)	(56, 646, 1921, 746)
(28, 2772, 637, 2336)	

From **Table 2.2.1,** one can find that when $i = 14$ and $i = 47$, $\gamma_{14} \equiv \gamma_{47} \equiv 50 \pmod{3049}$ and $\begin{matrix} \alpha_{14} = 1752, \beta_{14} = 2048, \\ \alpha_{47} = 1240, \beta_{47} = 184 \end{matrix}$,

$$\gamma_{14} \equiv y^{\alpha_{14}} g^{\beta_{14}} \equiv y^{\alpha_{47}} g^{\beta_{47}} \equiv \gamma_{47} \Leftrightarrow$$

$$g^{\beta_{14}-\beta_{47}} \equiv y^{\alpha_{47}-\alpha_{14}}$$

so we have $\Leftrightarrow g^{\beta_{14}-\beta_{47}} \equiv g^{x(\alpha_{47}-\alpha_{14})}$

$$x = \log_g y \Leftrightarrow$$

$$\beta_{14} - \beta_{47} \equiv x(\alpha_{47} - \alpha_{14}) \pmod{q-1}$$

$$\Leftrightarrow 1864 \equiv -512x \equiv 2536x \pmod{3048}$$

Using the Extended Euclidian Algorithm, on can compute the greatest common divisor of 1864, 2536, and 3048, that is, $\gcd(1864, 2356, 3048) = 8$, so the above congruent equation

$$233 \equiv 317z \pmod{381} \Leftrightarrow$$

can be reduced as $z \equiv 233 \times 317^{-1} \pmod{381}$, from which

$$\Leftrightarrow z \equiv 169 \pmod{381}$$

$$x \equiv \log_{11} 2009$$

one obtains a formula for $\equiv z + k \times 3048/8$ with

$$\equiv 161 + 381k \pmod{3048}$$

$k \in [0, 380]$. Trying different values of k, one finds that when $k = 6$, $x \equiv 2455 \pmod{3048}$ is the discrete logarithm we have wanted to compute.

2.3 Pohlig-Hellman Algorithm

The strategy employed by Pohlig and Hellamn [9] for breaking the DLP over a finite field is of reducing the DLP over a larger finite field to the DLPs over several subfields of smaller order. The order of such a subfield is a primer factor of the order of the finite field in consideration. After having found these DLPS over the subfields, one can construct the solution of the DLP by applying the Sunzi Theorem or the Chinese Remainder Theorem, the name often used by western scientist. One begins by announcing the famous CRT.

Theorem 2.3.1 Sunzi Theorem or Chinese Remainder Theorem (CRT)

Let $m_1, m_2, \cdots, m_k$ be the positive integers that are pair wise cop rime. Then, for any given integers $b_1, b_2, \cdots, b_k$, there exists an integer x satisfying the system of simultaneous congruence :

$x \equiv b_1 \pmod{m_1}$

$x \equiv b_2 \pmod{m_2}$

$\vdots$

$x \equiv b_k \pmod{m_k}$

Let $M = \prod_{i=1}^{k} m_i$, $M_i = M/m_i$, and M_i^{-1} be the inverse of M_i modulo m_i, then the smallest solution of the above simultaneous congruent system is $x \equiv \sum_{i=1}^{k} M_i M_i^{-1} b_i \pmod{M}$

In establishing the Pohlig-Hellman algorithm, one needs the Little Fermat Theorem:

Theorem 2.3.2 Little Fermat Theorem
Let p be a prime and a is a positive integer, then there exists next congruence: $a^{p-1} \equiv 1 \pmod p$

Considering the DLP over $GF^*(q)$ q being a prime, one at first finds the prime factorization of the multiplicative group order as $\left|GF^*(q)\right| = q-1 = \prod_{i=1}^{k} p^{n_i}$. For example, for $GF^*(3049)$, $\left|GF^*(3049)\right| = 3049-1 = 3048 = 2^3 \bullet 3 \bullet 127$. Let g be a generator of the filed in consideration, and $y \in GF^*(q)$, one wants to find the discrete logarithm $x = \log_g y$, that is, finding an element x such that $y \equiv g^x \pmod q$. This DLP can be reduced into k subprobelms, each of which will be solved by finding n_i discrete logarithms in a subgroup of order p_i. $\forall i, 1 \le i \le k$, Try to solve congruent equation $x \equiv x_i \pmod{p_i^{n_i}}$. In order for doing this, one takes x_i as an element of the finite field $GF(p_i^{n_i})$, and expands it as a polynomial in p_i, that is, $x_i = \sum_{j=0}^{n_i-1} x_{i,j} p_i^j$, $0 \le x_{i,j} \le p_i - 1$. Remark that $\forall i$, $\exists c_i \in \mathbf{Z}$ $x = x_i + c_i p_i^{n_i}$. Applying the **Theorem 2.3.2** that gives $g^{q-1} \equiv 1 \pmod q$, one can derive the following identities for computing the coefficients $x_{i,j}$:

$$y^{\frac{q-1}{p_i}} \equiv g^{x \bullet \frac{q-1}{p_i}} \equiv g^{(\sum_{j=0}^{n_i-1} x_{i,j} p_i^j + c_i p_i^{n_i}) \bullet \frac{q-1}{p_i}} \equiv (g^{\frac{q-1}{p_i}})^{x_{i,0}} \pmod q \quad (eq.\ 2.3.1)$$

$$(y \bullet g^{-x_{i,0}})^{\frac{q-1}{p_i^2}} \equiv g^{(\sum_{j=1}^{n_i-1} x_{i,j} p_i^j + c_i p_i^{n_i}) \bullet \frac{q-1}{p_i^2}} \equiv (g^{\frac{q-1}{p_i}})^{x_{i,1}} \pmod q \quad (eq.\ 2.3.2)$$

$$(y \bullet g^{-x_{i,0} - x_{i,1} p_i})^{\frac{q-1}{p_i^3}} \equiv g^{(\sum_{j=2}^{n_i-1} x_{i,j} p_i^j + c_i p_i^{n_i}) \bullet \frac{q-1}{p_i^3}} \equiv (g^{\frac{q-1}{p_i}})^{x_{i,2}} \pmod q \quad (eq.\ 2.3.3)$$

$$\vdots$$

$$(y \bullet g^{-\sum_{j=0}^{n_i-2} x_{i,j} p_i^j})^{\frac{q-1}{p_i^{n_i}}} \equiv g^{(x_{i,n_i-1} p_i^{n_i-1} + c_i p_i^{n_i}) \bullet \frac{q-1}{p_i^{n_i}}} \equiv (g^{\frac{q-1}{p_i}})^{x_{i,n_i-1}} \pmod q \quad (eq.\ 2.3.4)$$

Remark that the order of the cyclic group $< g^{\frac{q-1}{p_i}} >$ is precisely equal to p_i, so that solving the congruent equation $x \equiv x_i \pmod{p_i^{n_i}}$ is equivalent to find the n_i discrete logarithms $x_{i,j}$ $\forall i$ $1 \le i \le k$ and $0 \le j \le n_i - 1$ over the cyclic group $< g^{\frac{q-1}{p_i}} >$ by recursively employing the equations 2.3.1-2.3.4. Once one gets k solutions of the congruent system $x \equiv x_i \pmod{p_i^{n_i}}$ for $1 \le i \le k$, one will get the solution of the global discrete logarithm $x = \log_g y$ by applying the CRT, as shown in **Theorem 2.3.1**. According to the above analysis, one is ready to describe the Pohlig-Hellman algorithm:

Algorithm 2.3.1 Pohlig-Hellman Algorithm

1) INPUT: q, g, y.

2) OUTPUT: $x = \log_g y$.

3) Factorizing the order of the field: $q-1 = \prod_{i=1}^{k} p^{n_i}$

4) Expanding x_i in $x_i = \sum_{j=0}^{n_i-1} x_{i,j} p_i^j$, $0 \le x_{i,j} \le p_i - 1$

5) Using Eq(2.3.1)-Eq(2.3.4)and the Shanks's Giant-steps Baby-steps Algorithm or the Pollard rho-algorithm to solve n_i discrete logarithms.

6) Using the CRT to find $x = \log_g y$.

Example 2.3.1 Let the finite field in consideration is $GF^*(3049)$. It is easy to check $g = 11$ is a primitive root. One wants to compute $\log_{11} 2009$, that is to find a x such as $2009 \equiv 11^x \pmod{3049}$ using the Pohlig-Hellman algorithm.

Solution:

1) Factorizing the order of $GF^*(3049)$ into prime factors: $\left|GF^*(3049)\right| = 3048 = 2^3 \bullet 3 \bullet 127$

2) Trying to solve the following simultaneous congruent system:

$$\begin{aligned} x &\equiv x_1 \pmod{2^3} \\ x &\equiv x_2 \pmod{3^1} \\ x &\equiv x_3 \pmod{127^1} \end{aligned}$$

3) Expanding x_1 into $x_1 = x_{1,0} + x_{1,1} \bullet 2 + x_{1,2} \bullet 2^2$ with $0 \le x_{1,i} \le 1$ and applying Eq(2.3.1)-Eq(2.3.4), one obtains three discrete logarithm

problems over $<11^{\frac{3048}{2}}>$: $2009^{\frac{3048}{2}} \equiv (11^{\frac{3048}{2}})^{x_{1,0}} \pmod{3049}$, $(2009 \bullet 11^{-x_{1,0}})^{\frac{3048}{2^2}} \equiv (11^{\frac{3048}{2}})^{x_{1,1}} \pmod{3049}$, and $(2009 \bullet 11^{-x_{1,0}-x_{1,1} \bullet 2})^{\frac{3048}{2^3}} \equiv (11^{\frac{3048}{2}})^{x_{1,2}} \pmod{3049}$. By using the brute force algorithm (because of smallness of the group order), one gets $x_{1,0} = 1$, $x_{1,1} = 1$, $x_{1,2} = 1$. So finally one is being able to find $x \equiv x_1 \equiv 7 \pmod{2^3}$.

4) Similar to step 2), one could find that $x \equiv x_2 \equiv 1 \pmod{3^1}$, and $x \equiv x_3 \equiv 42 \pmod{127^1}$.

5) Using the CRT to solve the simultaneous congruent system:

$$\begin{aligned} x &\equiv 7 \pmod{2^3} \\ x &\equiv 1 \pmod{3^1} \\ x &\equiv 49 \pmod{127^1} \end{aligned}$$

, one successfully recovers the same answer as in the precedent examples, that is, $x \equiv \log_{11} 2009 \equiv 2455 \pmod{3048}$.

Example 2.3.2 Let the finite field in consideration is $GF^*(4729)$. It is easy to check that $g = 17$ is a primitive root. One wants to compute $\log_{17} 2009$, that is to find a x such as $2009 \equiv 17^x \pmod{4729}$ using the Pohlig-Hellman algorithm. By employing our software developed in a high level programming language as Visual Basic 6.0, one just needs to input the prime factors of the group order ($4728 = 2^3 \bullet 3 \bullet 197$), the primitive root, and $y = 2009$ to finally find that, $\log_{17} 2009 \equiv 4553 \pmod{4728}$.

3. CONCLUSIONS

In this article, one has presented, illustrated and analyzed some attacks of the discrete logarithm problem over finite fields. One equally implemented these attack algorithms for the finite field $GF(3049)$. It is worth to emphasize that it should be a continuous effort to find again more efficient algorithm or mathematical theory, in order to break the DLP over finite fields, of which the intractability becomes again and again more difficult!

REFERENCES

[1] W. Diffie, and M. Hellman, “New directions in cryptography”, IEEE Trans. Info. Theory, Vol. 22, 1976, pp. 644~654.

[2] R. Rivest, A. Shamir, and L. Adleman, “A method for obtaining digital signatures and public-key cryptosystems”, Communications of ACM, Vol. 21, 1978, pp.120~126.

[3] T. ElGamel, “A public key cryptosystem and a signature scheme based on discrete logarithms”, IEEE Transactions on Information Theory, Vol. 31, 1985, pp. 469~472.

[4] N. Koblitz, “Elliptic curve cryptosystems”, Mathematics of Computation, Vol.48, 1987, pp. 203~209.

[5] V. Miller, “Uses of elliptic curves in cryptography”, Advances in Cryptology-Crypto ’85, LNCS 218, 1986, pp.417~426.

[6] D. Shanks, “Class number, a theory of factorization and genera”, In Proc. Symposium Pure Mathematics, American Mathematical Society, Vol. 20, 1970, pp. 415~440.

[7] J. Buchmann, M. Jacobson, and E. Teske, „On some computational problems in finite abelian groups“, Math. Comp., Vol. 66(220), 1987, pp. 1663~1687.

[8] J. M. Pollard, “Monte Carlo methods for index computation (mod p)”, Math. Computation, Vol. 32, 1978, pp. 918~924.

[9] S. C. Pohlig and M. E. Hellman, “An improved algorithm for computing logarithms over $GF(p)$ and its cryptographic significance”, IEEE Trans. Info. Theory, Vol. 24, 1978, pp. 106~110.

[10] L. M. Adelman, “A subexponential algorithm for the discrete logarithm problem with application to cryptography”, Proc. Of the 20th Annual IEEE Symposium on Foundations of Computer Science, 1979, pp. 55~60.

[11] D. Coppersmith, “Fast evaluation of logarithms in fields of characteristic two”, IEEE Transactions on Information Theory, Vol. 30, 1984, pp.587~594.

[12] D. Coppersmith, A. Odlyzko, and R. Schroeppel, “Discrete logarithms in $GF(p)$”, Algorithmica, Vol. 1, 1986, pp. 1~15.

[13] B. A. Lamacchia, and A. M. Odlyzko, “Computation of discrete logarithms in prime fields”, Designs, Codes, and Cryptography, Vol. 1, 1991, pp. 46~62.

[14] D. Gordon, “Discrete logarithms in $GF(p)$ using the Number Field Sieve”, University of Georgia, preprint 1992.

PEOPLE TRACKING VIA A MODIFIED CAMSHIFT ALGORITHM

Fahad Fazal Elahi Guraya[1], Pierre-Yves Bayle[2] and Faouzi Alaya Cheikh[1]
[1]Department of Computer Science and Media Technology, Gjovik University College
Gjovik 2802, Norway
[2]Department of Computer Science, Universite de Bourgogne
Dijon, France
Email: fahad.guraya@hig.no

ABSTRACT

The past decade has witnessed a rapid growth in the use of video cameras in all aspect of our daily life. Video cameras use for Surveillance purposes has also increased. This increased the demand of automatic video surveillance algorithms that can detect abnormal activity/events in the surveillance area and can raise the alarm. Most of the event detection applications relay on persons detection and tracking or object tracking algorithm. CAMSHIFT is a tracking algorithm that has already been used for face-tracking. It is, however, not being successfully used for people/person tracking. This paper presents a modified version of CAMSHIFT algorithm which can be used for person tracking. The proposed algorithm incorporates the motion information with CAMSHIFT algorithm, which allows it to track people successfully even in the case of occlusions. The experimental results prove the consistency of the proposed algorithm.

Keywords: Object Tracking, CAMSHIFT, Motion estimation tracking.

1. INTRODUCTION

Object tracking is used for many applications such as motion-based recognition, automated surveillance, video indexing and retrieval, human computer interaction and traffic monitoring etc. Object tracking can be divided into series of steps, such as object representation, feature selection for tracking, object detection, background subtraction, and object segmentation.

Objects could be represented with the help of a single point (e.g. centroid) [1] or a set of points [2], by primitive geometric shapes (rectangles or ellipses) [3], silhouettes and contours for tracking complex non-rigid objects[4]. Feature selection also plays an important role in object tracking. The most commonly used features are color, edges, optical flow and texture [10, 11]. Various color spaces other than RGB are used for tracking purposes such as HSV because RGB colorspace does not correspond to the color differences perceived by humans [5]. Commonly used edge detectors are Canny Edge detector [6]. Optical flow can also be used as feature for object tracking. Some of the most used optical flow methods are [7, 8, 9].

Object trackers can be categorized into three different categories such as point trackers, kernel trackers and silhouette trackers. Point tracking methods can be further sub-categorized into deterministic methods such as [12, 13] and statistical methods [14, 15]. Kernel based object tracking methods are Mean Shift tracking [16], and Continuously Adaptive Mean Shift tracking (CAMSHIFT) [17]. Eigen-tracking [18] and Support Vector Machines [19] are kernel based methods for multi-view appearance models. Silhouette tracking methods are divided into contour evolution methods [20, 21] and shape matching methods [22, 23].

Background modeling could be the basic step in many video analysis applications, used to extract foreground or moving objects from the video frames. Change in a scene or foreground objects could be extracted from a video sequence by subtracting the background image from each frame. In general the background is considered to be constant or slowly changing due to luminance changes. In practice the background pixels are always changing, for that reason we need a model which accounts for gradual changes. Several techniques have been proposed in the literature for modeling the variation of the background information [24, 25, 26]. In [24] modeling of each stationary background color pixel with a single 3D (YUV color-space) Gaussian was proposed. A single 3D (YUV) Gaussian however is not suitable for outdoor scenes [27] since at a certain location multiple colors can be seen due to repetitive motion, shadows or reflectance. Thus Mixture of Gaussians (MoG) for modeling a single pixel color was proposed in [28].

In this paper, we use MoG method for background modeling. Contour detection provided by Intel Open Source Computer Vision Library (OpenCV) [29] is used for object detection and representation. CAMSHIFT kernel based object tracking is used to track the object with color information using Hue Saturation Value color-space, considered as a better representative of color perceived by human vision [5]. Motion information is also incorporated with color information using Optical flow method provided in [32].

In the next section, the implementation of tracking algorithms is presented. Section 3 proposed the CAMSHIFT algorithm with motion detection. Conclusion and future work is given in section 4.

2. PEOPLE DETECTION AND TRACKING

2.1 Background Estimation and Post Processing

To perform people tracking, we first used Mixture of Gaussian (MoG) method to calculate a good background model. The OpenCV functions are an implementation of the Gaussians Mixture Model in [30]. In this implementation, the model assumes that each pixel in the scene is modeled by a mixture of K Gaussian distributions where different Gaussians represent different colors. The weight parameters of the mixture represent the time proportions that those colors stay in the scene. Thus, the probable background colors are the ones which stay longer and are more static. The probability that a certain pixel has a value of x_N at time N can be written as $p(x_N)$ as shown in eq.(1).

$$p(x_N) = \sum_{j=1}^{K} w_j . \eta(x_N ; \theta_j) \tag{1}$$

where w_j is the weight parameter of the j^{th} Gaussian component. $\eta(x_N;\theta_j)$ is the Normal distribution of the jth component. Static single-color objects trend to form tight clusters in the color space while moving ones form widen clusters due to different reflecting surfaces during the movement. The measure of this is called the fitness value [30]. The K distributions are ordered based on the fitness and the first B distributions are used as a model of the background of the scene where B is estimated as shown in eq.(2).

$$B = \arg\min_b \sum_{j=1}^{b} w_j > T \tag{2}$$

The threshold T is the minimum fraction of the background model. In other words, it is the minimum prior probability that the background is in the scene. Background subtraction is performed by marking a foreground pixel any pixel that is more than 2.5 standard deviations away from any of the B distributions. Figure.1 shows the original frame and the background obtained after applying Gaussian mixture model.

Figure 1. An example of the background results obtained with the Adaptive Gaussian Mixture Model.

Next, we need to remove the shadows in the foreground obtained with the previous background subtraction method. In [31] the authors assume that we can consider a pixel as shaded background or shadow if it has similar chromaticity but lower brightness than those of the same pixel in the background image. Thus, with an appropriate threshold T, we can remove shadows from the foreground image. Eq.(3) shows the decision either a certain pixel belongs to shadow or not.

$$Shadow\{x,y\} = \begin{cases} 1 & if\, brightness_{img} < brightness_{bg} \\ & \&\, chromaticity_{img} = chromacity_{bg} \pm T \\ 0 & otherwise \end{cases} \tag{3}$$

The shadow removal is applied on our sample video sequence. Figure.2 shows a frame from the video before and after shadow removal.

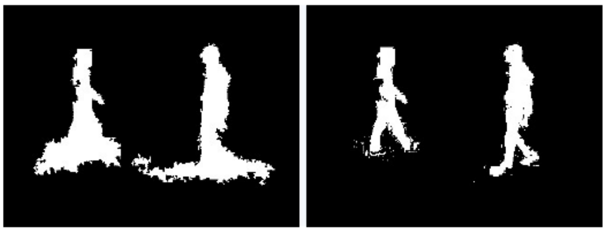


Figure 2. Foreground objects after background subtraction (left) and after shadows removal (right).

Finally, the morphological filtering is applied using closing operation. Unlike the opening which removes small objects, closing removes small holes. Figure.3 shows one frame of our sample video sequence before and after the application of morphological closing operation.

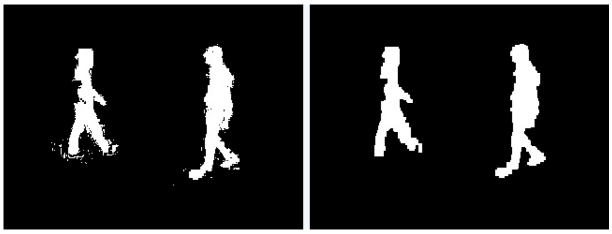


Figure 3. Foreground objects without morphological filtering (left) and after closing operation (right).

We can see that the closing operation removes small holes in the foreground, and also small points from the foreground around the person.

After the background subtraction, we apply a people detection algorithm, to detect all separate persons in the frame. For this, we proceed in two steps: As a first step a contour is detected. A contour is represented in OpenCV by a sequence of points. First, we use the OpenCV function to find contours around all separate foreground regions. In our case, we retrieve only the extreme outer contours and the function compresses horizontal, vertical, and diagonal segments, leaving only their ending points. Then, for each found contour, we define the smallest bounding box in which all the points of the contour are included, and we calculate its area. If this area is under some threshold, we remove this contour from the sequence to keep only contours of a potential person, and remove those that cannot be a whole person.

After that, we need to define a bounding box around the detected contour. The difficulty is that most of the time, the contour is not perfect, and we may have two distinct contours for the same person, for example, one for body and other for the head or the arms. These contours are in general very close or even overlapping. So, we just define a bounding box around the different contours, and detect which bounding boxes are overlapping or very close and merge them into one single box. We perform this operation for each contour, and then create a sequence of bounding boxes, corresponding to the different main areas in the foreground. The detected contour and bounding box in one of the frame of our sample video sequence is shown in figure.4.

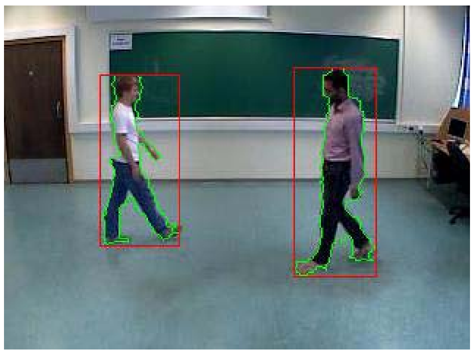


Figure 4. One frame with contours (green) and bounding boxes (red) around foreground objects.

2.2 PEOPLE TRACKING by CAMSHIFT

Now, we need to track the detected people, and for this, we will use the CAMSHIFT algorithm. The CAMSHIFT algorithm can be summarized in the following steps [29]:

(1) Set the region of interest (ROI) of the probability distribution image to the entire image.

(2) Select an initial location of the Mean Shift search window. The selected location is the target distribution to be tracked.
(3) Calculate a color probability distribution of the region centered at the Mean Shift search window.
(4) Iterate Mean Shift algorithm to find the centroid of the probability image. Store the 0th moment (distribution area) and centroid location.
(5) For the following frame, center the search window at the mean location found in step 4 and set the window size to a function of the 0th moment. Then go to Step 3.

The creation of the color histogram corresponds to steps 1 to 3. The first step is to define the region of interest (ROI) which is the bounding box corresponding to the detected person that we want to track. Then, we need to calculate the color histogram corresponding to this person. For that we use the HSV color space, and calculate a one dimensional histogram corresponding to the first component: hue. We also define a mask for the histogram calculation, which is the foreground image, to calculate the histogram only for the person, and not for the background inside the bounding box. But the results obtained with this method were not satisfying, because in the case where the background has almost the same color as the person, it is not possible to detect the difference between the two in the back-projection image. That is why we decided in a second time to use a three dimensional histogram, using the three components of the HSV color space: hue, saturation, value. With this method, we were able to find the location of the person in the whole frame, even with a similar background. In all cases the histogram bins values are scaled to be within the discrete pixel range of the 2D probability distribution image using eq.(4).

$$\left\{ \hat{p}_u = \min\left(\frac{255}{\max\{\hat{q}\}}\hat{q}_u, 255 \right) \right\}_{u=1\cdots m} \tag{4}$$

That is, the histogram bin values are rescaled to the range [0, 255], where pixels with the highest probability of being in the sample histogram will map as visible intensities in the 2D histogram back-projection image.

Now we have a histogram of the moving person, we need to find this person in all the frames (steps 4 to 5). For that end, we calculate the back-projection of the histogram in the subsequent frame. For each pixel of all input images we put, in the back-projection image, the value of the histogram bin corresponding to the pixel. In terms of statistics, the value of each output image pixel is the probability of the observed pixel that it is a pixel of the tracked object, given the distribution (histogram). Finally, using the previous location of the person, we detect the new position of the moving person and use it as starting search window for the next frame. The search window center could be computed from:

$$M_{00} = \sum_x \sum_y I(x,y) \tag{5}$$

$$M_{10} = \sum_x \sum_y xI(x,y) \tag{6}$$

$$M_{01} = \sum_x \sum_y yI(x,y) \tag{7}$$

$$x_c = \frac{M_{10}}{M_{00}}; y_c = \frac{M_{01}}{M_{00}} \tag{8}$$

Where M_{00} in eq.(5) is the zeroth moment, M_{10} in eq.(6) and M_{01} in eq.(7) are the first moments, these moments could be used to compute the next center position of the tracking window x_c and y_c as shown in eq.(8). Then, back to step 3, we calculate the new histogram of the person to update the previous one, using a slow update, to keep the difference between different persons if they are overlapping. The persons tracked and their corresponding tracking windows using CAMSHIFT algorithm are shown in figure.5.
But when two persons are crossing, it happens that both of the tracking windows follow the person in the foreground, for example if the two persons have similar colors.

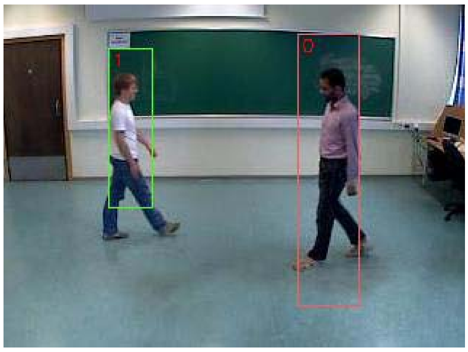


Figure 5. One frame with the tracking window of each tracked persons using CAMSHIFT tracking.

In order to solve this problem, we decided to add more information to the back-projection image, and we decided to use the motion coherence. When two persons are going in two opposite directions, the motion will allow us to follow the right person.

3. CAMSHIFT WITH MOTION (PEOPLE TRACKING USING OPTICAL FLOW)

In order to improve the results of the CAMSHIFT algorithm described above, we use an optical flow algorithm (Lucas-Kanade Method) [32] to determine the motion of the tracked persons.

3.1. Lucas-Kanade Algorithm

The basic idea of the LK algorithm rests on three assumptions:

(1) *Brightness constancy*: A pixel from the image of an object in the scene does not change in appearance as it (possibly) moves from frame to frame. For grayscale images, this means we assume that the brightness of a pixel does not change as it is tracked from frame to frame.
(2) *Temporal persistence or small movements*: The image motion of a surface patch changes slowly in time. In practice, this means the temporal increments are fast enough relative to the scale of motion in the image that the object does not move much from frame to frame.
(3) *Spatial coherence*: Neighboring points in a scene belong to the same surface, have similar motion, and project to nearby points on the image plane.
(4) The OpenCV function implements sparse iterative version of Lucas-Kanade optical flow in pyramids [32]. It calculates coordinates of the feature points

on the current video frame given their coordinates on the previous frame. The function finds the coordinates with sub-pixel accuracy.

So the aim of the optical flow method is, for a given set of points in a video frame find those same points in the next frame. Or for given point $[u_x, v_y]$ in frame F_1;find the point $[u_x + \delta_x, u_y + \delta_y]^T$ in image F_2 that minimizes ε as shown in eq.(9).

$$\varepsilon\left(\delta_x, \delta_y\right) = \sum_{x=u_x-w_x}^{u_x+w_x} \sum_{y=u_y-w_y}^{u_y+w_y} \left(F_1(x,y) - F_2(x+\delta_x, y+\delta_y)\right) \qquad (9)$$

Figure 6 presents the results of the Lucas-Kanade algorithm. The arrows are drawn between the previous and the next position of the pixels corresponding to the good features to track in the figure 6(right), and current frame is shown in figure 6(left).

To combine the previous histogram and the motion calculated by the LK optical flow method, we first calculate the previous histogram, and then calculate the back-projection image using this histogram. Before finding the new location of the person with CAMSHIFT, we update the back-projection image using the motion information. For that, we calculate the global motion of each person, and for each back-projection image calculated using the histogram, we update each point calculated with LK algorithm, and put a higher value for pixels going in the same direction of the person. Doing this, we will be able to follow two persons crossing, and know who is who when they keep going in two opposite directions. Using this method, we need also to use the first type of the histogram, namely the one dimensional histogram (hue component) because it was difficult to combine information from the 3-D histogram and motion. Using this method, we still have the problem of the background similar to the person color, and to solve it, we combined the back-projection image with the foreground mask, to keep only the back-projection of the foreground.

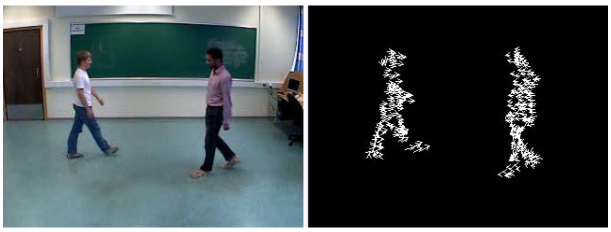


Figure 6. Original frame (left) and representation of the results of the LK algorithm (right).

3.2. Experimental Results

An example video is used to first test the simple CAMSHIFT algorithm without motion information. The CAMSHIFT without motion information failed to track the persons after the occlusion as can be seen in figure.7, which means color information is not enough to track the person in case of occlusion. The same video is used to test our proposed tracking algorithm using both color histogram and motion information. The same video frames shown in figure 7 are used in figure 8 to illustrate the effectiveness of the modified CAMSHIFT in resolving the confusion after the occlusion. We can see here an example of two persons crossing in the opposite directions; the motion in this case allows us to track both persons after the occlusion. The tracking windows keep the same tracking numbers as before and after the occlusion that shows the algorithm can correctly track each of the two persons even after heavy occlusion.

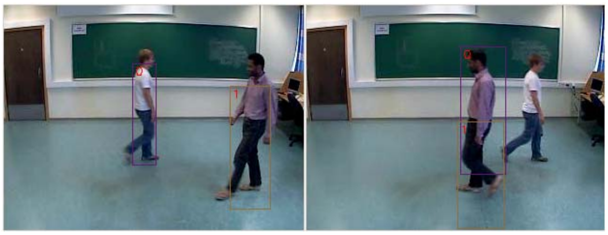


Figure 7. One frame with two persons before crossing (left) and same persons just after crossing (right) (Tracking using CAMSHIFT only).

Figure 8. One frame with two persons before crossing (left) and same persons just after crossing (right) (Tracking using CAMSHIFT plus Motion information).

4. CONCLUSIONS AND FUTURE WORK

In this paper, a modified CAMSHIFT algorithm is presented. This algorithm is using color feature similar to classical CAMSHIFT algorithm, motion or optical flow information is also added to it to make it robust against occlusions. The algorithm is verified for a set of videos. As future work the same algorithm will be extended to be used for multiple-cameras scenario also referred as multiple camera tracking.

REFERENCES

[1] Veenman, C. Reinders, M., and Backer, E. 2001. Resolving motion correspondence for densely moving points. *IEEE Trans. Patt. Analy. Mach. Intell. 23*, 1, 54–72.

[2] SERBY, D., KOLLER-MEIER, S., AND GOOL, L. V. 2004. Probabilistic object tracking using multiple features. In *IEEE International Conference of Pattern Recognition (ICPR)*. 184–187.

[3] COMANICIU, D., RAMESH, V., ANDMEER, P. 2003. Kernel-based object tracking. *IEEE Trans. Patt. Analy. Mach Intell. 25*, 564–575.

[4] YILMAZ, A., LI, X., AND SHAH, M. 2004. Contour based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Trans. Patt. Analy. Mach. Intell. 26*, 11, 1531–1536.

[5] PASCHOS, G. 2001. Perceptually uniform color spaces for color texture analysis: an empirical evaluation. *IEEE Trans. Image Process. 10*, 932–937.

[6] CANNY, J. 1986. A computational approach to edge detection. *IEEE Trans. Patt. Analy. Mach. Intell. 8*, 6, 679–698.

[7] HORN, B. AND SCHUNK, B. 1981. Determining optical flow. *Artific. Intell. 17*, 185–203.

[8] KANADE, T., COLLINS, R., LIPTON, A., BURT, P., AND WIXSON, L. 1998. Advances in cooperative multi-sensor video surveillance. *Darpa IU Workshop*. 3–24.

[9] BLACK, M. AND ANANDAN, P. 1996. The robust

estimation of multiple motions: Parametric and piecewisesmooth
flow fields. *Comput. Vision Image Understand. 63*, 1, 75–104.

[10] HARALICK, R., SHANMUGAM, B., AND DINSTEIN, I. 1973. Textural features for image classification. *IEEE Trans.Syst. Man Cybern. 33*, 3, 610–622.

[11] LAWS, K. 1980. Textured image segmentation. PhD thesis, Electrical Engineering, University of Southern California.

[12] SALARI, V. AND SETHI, I. K. 1990. Feature point correspondence in the presence of occlusion. *IEEE Trans.Patt. Analy. Mach. Intell. 12*, 1, 87–91.

[13] VEENMAN, C., REINDERS, M., AND BACKER, E. 2001. Resolving motion correspondence for densely moving points. *IEEE Trans. Patt. Analy. Mach. Intell. 23*, 1, 54–72.

[14] BROIDA, T. AND CHELLAPPA, R. 1986. Estimation of object motion parameters from noisy images. *IEEE Trans.Patt. Analy. Mach. Intell. 8*, 1, 90–99.

[15] STREIT, R. L. AND LUGINBUHL, T. E. 1994. Maximum likelihood method for probabilistic multi-hypothesis tracking. In *Proceedings of the International Society for Optical Engineering (SPIE.)* vol. 2235. 394–405.

[16] COMANICIU, D., RAMESH, V., ANDMEER, P. 2003. Kernel-based object tracking. *IEEE Trans. Patt. Analy. Mach. Intell. 25*, 564–575.

[17] Bradski, G. R. 1998. Real Time Face and Object Tracking as a Component of a Perceptual User Interface. In *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (Wacv'98)* (October 19 - 21, 1998). WACV. IEEE Computer Society, Washington, DC, 214.

[18] BLACK, M. AND JEPSON, A. 1998. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *Int. J. Comput. Vision 26*, 1, 63–84.

[19] AVIDAN, S. 2001. Support vector tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 184–191.

[20] ISARD, M. AND BLAKE, A. 1998. Condensation - conditional density propagation for visual tracking. *Int. J.Comput. Vision 29*, 1, 5–28.

[21] RONFARD, R. 1994. Region based strategies for active contour models. *Int. J. Comput. Vision 13*, 2, 229–251.

[22] HUTTENLOCHER, D., NOH, J., AND RUCKLIDGE, W. 1993. Tracking nonrigid objects in complex scenes. In *IEEE International Conference on Computer Vision (ICCV)*. 93–101.

[23] KANG, J., COHEN, I., ANDMEDIONI, G. 2003. Continuous tracking within and across camera streams. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 267–272.

[24] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780-785, 1997.

[25] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, “Background modeling and subtraction of dynamic scenes,” Oct. 2003, pp. 1305-1312 vol.2.

[26] M. Irani and P. Anandan, “Video indexing based on mosaic representations,” *Proceedings of the IEEE*, vol. 86, no. 5, pp. 905-921, May 1998.

[27] X. Gao, T. Boult, F. Coetzee, and V. Ramesh, “Error analysis of background adaption,” vol. 1, 2000, pp. 503-510 vol.1.

[28] C. Stauffer and W. Grimson, “Learning patterns of activity using real-time tracking,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 747-757, Aug 2000.

[29] Intel Corporation (2001): Open Source Computer Vision Library Reference Manual.

[30] P. KaewTraKulPong and R. Bowden, (2001): *An improved adaptive background mixture model for real-time tracking with shadow detection*, in Proc. 2nd EuropeanWorkshop on Advanced Video-Based Surveillance Systems.

[31] Andrea Prati, Ivana Mikic, Mohan M. Trivedi, Rita Cucchiara, (2003): *Detecting Moving Shadows: Algorithms and Evaluation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 918-923.

[32] Jean-Yves Bouguet (2000): Pyramidal Implementation of the Lucas Kanade Feature Tracker – Description of the algorithm, Intel Corporation Microprocessor Research Labs.

REALIZATION OF SCENE MAPPING SYSTEM BASED ON VC++ *

Dan Liu, Long Zhang, Jiang Chang
Department of Forensic Science & Technology, China Criminal Police University
Shenyang, Liaoning110035, P.R.China
Email: dliudlmu@gmail.com

ABSTRACT

The criminal scene map plays an important role in criminal cases, a computer-aided mapping system for drawing the ichnography of a criminal scene and the specific methods to develop it are put forward, the system development tools is the basic class library of Visual C++ which has the characteristics of expansion and succession, rich scene model make the scene mapping simple and practical. The experiments show that the system can be widely used in the drawing of a variety of criminal scene ichnographies.

Keywords: Scene Mapping, Ichnography Mapping, Mapping System, CAD, VC++.

1. INTRODUCTION

According to People's Republic of China Criminal Procedure Law, the Procedure Requirements of Public Security Organs for Criminal Cases and the Rules of Public Security Organs Investigating the Crime Scene, the criminal scene graph is an important part of investigating the criminal scene cases and checking the working records.

Criminal Scene Mapping is that criminal and technical personnel use the principles or methods of cartography to describe the crime scene with an appropriate proportion by a variety of symbols and text.

Criminal scene graph can clearly reflect the track activities of criminals and the overall state of the crime scene. It is very important in incarnating the actual situation of criminal evidence and it plays an unusual role on studying the criminal case, confirming the statements of criminal suspects and restoring the original appearance of crime scene.

According to the reflected content and form, the scene graph can be separated into the scene ichnography, the three-dimensional scene map and the scene perspective drawing. According to the range size, the ichnography for the scene can be separated into the direction map, the internal full view of the scene, the local ichnography and the expand map for the scene [1].The ichnography mapping system called Criminal Cases Ichnography Mapping System is developed, on the basis of the basic class library (Microsoft Foundation Class, referred to as MFC) [2] [3] of Visual C + + (hereinafter referred to as VC).

Criminal Cases Ichnography Mapping System is a software system that is used to draw the crime scene. It provides a graphical editing environment to depict crime scene by computer. The software system can draw the ichnography and print it quickly and accurately even on the criminal scene using the computer. At the same time, the system can also store ichnography as a bitmap file in order to save, search and view.

* This paper is supported by China Criminal University Science Research Project (0209023).

2. COMPUTER GRAPHICS AND OBJECT-ORIENTED DEVELOPMENT ENVIRONMENT

Computer Graphics (Computer Graphics, referred to as CG) is a kind of science which uses mathematical algorithm to transform two-dimensional graph or three-dimensional graph into a grid form of computer display [4].

In 1962, Ivan E. Sutherland in Massachusetts Institute of Technology Lincoln Laboratory published a doctoral thesis entitled Sketchpad: a man-machine interactive graphical communication system. In the thesis, He used the term "Computer Graphics" for the first time and proved the interactive computer graphics is a feasible and useful research area and the independent status of computer graphics as a new branch of science was determined [5].

The major contents of Computer Graphics include how to present graph in computer and the relevant principles and algorithms of calculation, processing and display of graph by computer. Graph is composed of geometric elements, such as point, line, surface, three-dimensional object and other non-geometric attributes, such as gray level, color, line type and line width. According to the processing technology, graph is mainly separated into two categories: one is showed by line information, such as engineering drawings, contour maps and wire-frame map of surface; the other is shading map, which is usually referred to realistic graph [6].

The rapid development of computer graphics impels the accelerated emergence of various computer graphics systems. In the past decade, drawing crime scene graph by computer are applied in more and more cases within Public Security System. However, majority popular CAD software mainly supports the engineering drawing mapping and output [7]. Taking AutoCAD as an example, although interactive graphics is its most common function , but AutoLisp is needed as the programming language in order to carry out a perfect drawing, and sometimes other high-level language must be used. So, it isn't very convenient and has strong proprietary. As for the mapping software of Public Security System, they mostly realize a single graphics and spend more time on mapping. There didn't realize the intelligent processing of computer and the high efficiency of drawing the scene graph.

Generally, scene ichnography can be drawn with the method of vertical projection. When drawing, the terrain and surface features are drawn to the graph on standard scale which is selected according to the actual situation. Scene ichnography includes the major part of the scene, the relevant objects of criminal activity and the symbols and text which depict the objects at the scene.

Using the basic principles of computer graphics, in accordance with the idea of object-oriented development system, the variables included in the system which is described by objects and classes can realize the scene ichnography completely. User interface uses Windows mode. Chinese primary menu, the pull down menu and toolbar make up the simple and efficient operation. Graphics Template is provided for the left mouse button to drag when mapping and man-machine conversation is realized via dialog box [8].

As a visual, object-oriented programming tools, Visual C + + 6.0 is the most powerful, widely used and efficient application development tool for Windows currently. In addition to the general application structure class, its MFC class library also provides a set of specialized Visual Object Class, such as Class Menu, Class Button, and Class View. Object-oriented programming language describes things as a entirety which is known as the object. Including a series of its own attributes and operation methods, objects can be seen as examples of classes. Class is entity model which package the data and function together as a type [9].

3. SYSTEM DESIGN

The system call CDC device context object classes, CDC object provides device context member functions to deal with the display or printer and device context customers to deal with the window client display area. Device context operation class provides member functions to deal with drawing tools. Secure Graphics Device Interface (GDI) object collects and deals with color and palette. It also acquires and sets up graphics attributes, projects, processes view point, extends window, draws the text, deals with fonts, converts coordinate, deals with region, cuts, pastes and draws straight line, simple ellipse and polygon. The CDC member functions can achieve all functions of the graphics [10] [11].

Using the object-oriented idea, System design and realize class point, class curve, class line, class rectangle, class circle, class text, and class combination as the composition of the various elements of the scene graph.

According to the classification of scene graph elements and the requirements of system design objectives, the system designs the following modules: the graphics module, the information module, the model module, the image module, input and output modules [12].

Graphics modules: The variety of graphics rendering is the core of the system, the purpose of introducing this module is to select the brush style and the primitive type more easily, achieve the character, terrain and geographical object and physical evidence accurately, and reflect a real criminal scene and criminal activities. Class Shape is designed in the system on the basis of device context class. As the basic class of primitives, Shape class inherits from the basic class Cobject in MFC and contains the basic properties and operations of primitives. On the basis of inheriting class shape, class line, class curve, class arc, class circle, class rectangle and other simple graphics classes redefine their virtual function combined with its own characteristics and increase the necessary attributes and methods.

Information modules: One of the requirements of the scene graph is a concise text description. This module mainly realizes text function, that is, the basic information and notes are marked with the name, the time and the location of the case. System designs Class Font on the basis of device context class to set up display font and size.

Model modules: Model module is one of the characteristics of the system. Using the frequently-used legend which system provides in the toolbar, technical policeman can quickly draw the criminal scene. This function can be made up by graphics modules. For the sake of convenience, system set up class combination. Class combination is the basic class which includes characteristic points array and the basic graphics class. In this class, the graph generating function can create composite figure in the order of storage sequence, graph type and characteristic points array. By calling the mouse position function, the location information is converted into graphical rotation and offset rules. After graphic positioning and mapping, user can still move all the basic primitives of combination class for the actual needs. Graphics modules include some common objects which are related to the criminal activities, such as vestige, exhibits, trails, cadaver, and tools for criminal purpose [13].

Image module: In order to overcoming the general shortcomings of traditional scene graph, this system inserts local scene photos in scene graph in a manner of label. System calls document class Cfile and windows picture viewer, sets up member function document management and adds event handling code to the function label. The system also realizes the integration of criminal scene graph and criminal scene photos [14].

Input and output modules: The system can save criminal scene graph as vector graphics files and read it. On the basis of calling Bmp format file, system calls the function open and storage to complete the operation. System also provides serialization output operation. Serialization object class is derived from class Cobject, it realizes the function serialize in order to complete the geometric preservation of the criminal scene vector graphics[15].

4. EXPERIMENTAL RESULTS

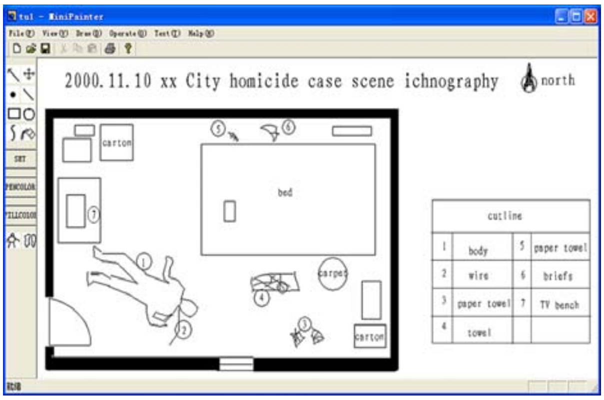


Figure 1. a criminal case scene ichnography

5. CONCLUSIONS

Criminal Cases Ichnography Mapping System discovers from the actual needs of the detection of criminal cases. The entire design mainly focused on software programming, it makes a good use of VC software in ample library resources, strong underlying control and flexible interface preparation. After drawing the actual scene to test, the software run properly and has made good results, which can basically meet the mapping needs of the majority police personnel. But in order to adapt to the special requirements of various cases, there are many places need to be improved and perfected, for example, the function needs to be further expanded and the interface needs to be further beautified.

REFERENCES

[1] Ou Huanzhang et al, Crime Scene Investigation Tutorial, Beijing: Police Education Publishing House, 2007(in Chinese).

[2] Xu Xiaogang, Gao Zhaofa, Wang Xiujuan et al, Visual C++ Entry and Improving, Beijing: Tsinghua University Press, 1999(in Chinese).

[3] Hou Junjie et al, Head First MFC, Hubei: Huazhong University of Science and Technology Publishing House, 2001(in Chinese).

[4] Zhang Caiming, Yang Xingqiang, Li Xueqing et al, Computer Graphics, Beijing: Science Press, 2005(in Chinese).

[5] Ivan Edward Sutherland, Sketchpad: A man-machine graphical communication system, Massachusetts Institute of Technology Lincoln Laboratory, 1963.

[6] Yao Haigen, Shi Liqin et al, Computer Graphics Application, Beijing: Science Press, 2004(in Chinese).

[7] Huang Woming, More “Use of Computer Mapping on Criminal Scene”, Road Traffic Management, No.12, 2000, pp. 8~9(in Chinese).

[8] Tang Yangshan, Li Jiang, Jiao Cheng-wu et al, “Accident Scene Graph CAD Based on MFC”, Journal of Engineering Graphics, Jan 2006, pp. 9~12(in Chinese).

[9] Li Yujian, Visual C++ Practice and Improve－Graphics and Image Articles, Beijing: China Railway Publishing House, 2001, pp.55~78(in Chinese).

[10] Yuan Quan, Pei Jianping, Li Yibing, “Road Traffic Accident Scene Mapping System for Accident Reconstruction”, Computer Engineering, March 2005, pp.206~207(in Chinese).

[11] Zeng Ruixing, Song Jun, Huo Jianjun, “Using of CAD in Crime Scene Mapping”, Criminal Technology, March 2002, pp.42~43(in Chinese).

[12] Kruglinski D J, Wingo S, Shepherd G, Programming Visual C++ 6.0 Technology Lowdown, Beijing: Hope Electronic Press, 1999(in Chinese).

[13] Chen Shixian et al, Crime scene analysis, Beijing: Masses Press, 2003(in Chinese).

[14] Donald Hearn M.Pauline Baker, Computer Graphics C Version Second Edition, Beijing: Tsinghua University Press, May 1998.

[15] Lin Junjie et al, Visual C++6 Program Design Classic, Beijing: Science Press, 1999(in Chinese).

A DISTRIBUTED IMAGING FRAMEWORK TO EXTRACT, ANALYSE AND VISUALIZE CELL FEATURES IN MULTI-DIMENSIONAL BIO-IMAGE DATASETS

Tony Mclay, Andreas Hoppe, Darrel R Greenhill, Souheil Khaddaj
Faculty of Computing, Information Systems and Mathematics, Kingston University, London, KT1 2EE, UK
Email: S.Khaddaj@kingston.ac.uk

ABSTRACT

Current imaging technologies can generate datasets of immense size and new image processing applications are applying ever more complex algorithms to these datasets. Parallel computing resources are commonly employed to obtain results in reasonable timescales. This paper outlines the development of a distributed framework to provide support for concurrent processing across a parallel system, and to apply this framework to the 3D image processing and visualization domains. Specifically, image processing to identify and extract cell image features in large multi-dimensional bio-image datasets, and visualization of the results. A novel aspect of this integrated framework is to allow feature extraction on the fly for the analysis of spatio-temporal events and their immediate visualisation, which greatly enhances the interpretation of the data.

1. INTRODUCTION

The combination of light microscopy and image processing led to the development of quantitative image analysis techniques in cell biology. Image processing in microscopy is particularly useful for the automatic analysis of large repetitive data sets such as generated by time-lapsed observations. More recently, imaging systems to automatically screen large numbers of cells have become available. These High Content Screening (HCS) systems are an efficient approach for assessing cellular signalling and morphology [Lit-Hsin 2007]. HCS uses automated microscopy to collect images of cultured cells from a large number of cell culture dishes. The images are processed with image segmentation algorithms to identify cellular structures or to quantify their morphology, for thousands of individual cells. Moreover, it is increasingly important to study dynamic processes in living cells in three dimensions at high temporal resolution. These data volumes are an order of magnitude higher and the 3D image processing techniques are substantially more computationally expensive.

In the last few decades, continued advances in the design of the single CPU, and the innovative organisation of these units into parallel clusters has provided ever increasing processing performance. Indeed, in the mid 90's, the High Performance Computing community shifted to using clusters of these 'volume economic' single CPU's in place of the then prevalent vector processing architectures

Furthermore, in recent years there has been a profound shift to designing 'scaled in' parallelism. The rate of performance increase of single CPU designs has slowed dramatically due to significant technological challenges such as design and manufacture at ever decreasing scales, the 'power wall', and limits to Instruction Level Parallelism (ILP) (Wall, D. W., 1991). This has resulted in an inflection point in the industry, and the emergence of multi-core processor designs, and active research into many-core units (Gschwind, M., 2006). Where in the past, the complexities of parallel processing could be sidestepped by relying on Moore's law and waiting for a faster single CPU, or kept somewhat manageable by building clusters from single CPU components, we have now entered the true 'parallel age'. Indeed, vector processing cores are now back in the mix of 'volume economic' components, due to the success of the STI (Sony, Toshiba & IBM) alliances Cell Broadband Engine in the Sony Play station, and IBM's derived PowerXCell 8i processor (Walsh, R. et al., August 2008).

The universal acceptance of parallel processing technology has been slow for a number of reasons. The complexity of parallel processing, the need for specific skills and the increased software production costs has been a major deterrent, as has the plethora of competing compute architectures and choice of connection topologies. On the other hand, the generation of huge amounts of data from scientific and engineering applications can only be processed in reasonable timescales by leveraging parallel processing technologies. Technological advancement is driven by this need, the landscape is changing, and future software must fully embrace the challenges and opportunities of ubiquitous parallel processing (Asanovic, K. et al., 2006).

In the following sections, we outline research underway to harness parallel processing resources through the use of a distributed framework that is built on top of MPI. To demonstrate and test the framework, we are developing an application that applies feature extraction, analysis and visualization techniques to multi-cell 3D image stacks.

2. FEATURE EXTRACTION AND ANALYSIS

Many pre-processing, image segmentation, feature extraction and feature analysis techniques are well suited to parallelisation and so provide good diverse examples upon which to test our framework. In this work, segmentation techniques to isolate cells and to extract morphological features are applied. Internal cellular structures such as stress fibres, and vesicles can be made visible through fluorescent tagging which then allows assessing the orientation or density of these internal elements. Behaviour characteristics such as motility activity are assessed as spatio-temporal processes. The acquired features are used to form feature vectors in a corresponding feature space upon which feature analysis techniques can then be applied. To ensure sensible comparisons, these vectors are normalized and de-correlated.

The objective of the analysis is to identify important cell characteristics. The success of the analysis hinges on identifying the pertinent features for cell type classifications, for example classifying malignant and benign cancer cells. Cell classification in this research uses simple distance classification techniques (k-means clustering). Principle Component Analysis techniques will also be useful in both the extraction and analysis stages, and also in the visualization stage.

3. VISUALIZATION

Having segmented the cells and extracted the prominent characteristics, they can then be visualized. As with feature extraction and analysis, many visualization techniques are ideal candidates for parallelisation. Indeed, this has been a significant driver in the design and development of the parallel architectures of today's GPU's.

In 3D visualization, image data translation is commonly converted into multi-polygon graphics representations that a graphics engine can then render efficiently using hardware accelerated techniques for representing such entities, including clipping algorithms and hidden surface removal. In order to test our framework, we use a simple pre-processing method, such as threshold, and direct volume rendering (Hege, H. C., Hollerer & Stalling, 1993).

Direct volume rendering describes techniques to directly render 3D images without the need to convert to a polygonal representation, although volume rendering of scenes of polygonal objects is also a popular technique for shading realism (Appel, 1968). We specifically use image order volume rendering (ray casting) techniques, which are very computationally intense but appealing for there ability to produce realistic shading. We also aim to provide a variation where instead of detecting iso-surfaces, intensity samples will be taken at regular intervals as the ray penetrates the volume, and these values accumulated according to some ray function to compose the final image (Razdan et al., 2001).

The parallelization of volume rendering has received considerable attention in the literature, with various techniques being devised to exploit the inherent parallelism. These techniques generally use domain knowledge to optimisation data and task partitioning in parallel implementations, and to align decompositions to specific algorithms to improve load balancing (Gao et al., 2005). When compositing images on a cluster, a popular choice is to use a 'sort last' approach (Molnar, S. et al., 2008). In this approach, partial volume rendering is processed on arbitrary nodes, and only then are values communicated to particular processors responsible for compositing portions of the image space. A particular example is (Yu, H., Wang & Ma, 2008) which describes a technique that combines the 'direct send' and 'binary swap' compositing methods for sort last parallel volume rendering.

4. PARALLELISATION AND DATA DECOMPOSITION

In order to achieve reasonable results in interactive timescales, the aggregated resources of multiple processors can be brought to bear on compute intensive applications. However, parallelisation introduces problems that do not occur in sequential programs. The synchronization of concurrent computing processes becomes necessary. Lock and barrier constructs are employed to manage data integrity, and extreme care is required to avoid introducing deadlocks. The indeterminate timing of concurrent communicating processes means that time dependent problems are particularly difficult to resolve. In addition, one has to consider how to best partition and manage work and data when parallelising an application.

In symmetric multiprocessor architectures, communication uses the shared memory paradigm. Cache coherence models manage the integrity of data between local processor caches and shared main memory. These cache coherence schemes can be an impediment to scaling this sort of architecture, a point that designers of Multi-core and many-core systems have to consider. Distributed memory architectures commonly rely on the 'message passing' paradigm and remote memory access.

The Message Passing Interface (MPI) (University of Tennessee, Knoxville, Tennessee, 1995) (Gropp, Lusk and Skjellum, 1999) specifies the communication protocols for point to point and collective communications on the basis of the message passing paradigm. MPI introduces the concept of a communicator which provides a context to multiple communications and is also suited for library development and fault tolerance (Geist et al., 1994). It also provides a feature for an application to provide topological information. The current version (MPI-2) adds dynamic process, remote memory access, and parallel I/O capabilities (Gropp, Lusk and Thakur, 1999).

Parallelisation implies the division of tasks and/or data. Data decomposition across multiple nodes in a cluster is particularly attractive for very large datasets as it allows the accommodation of these datasets across a cluster's aggregate memory, whilst the memory installed on a single node can remain modest. However, there are problems with this approach. Division and aggregation implies extra processing, communication, and account must be made for load balancing to fully utilize the available resources (e.g. see (Marchesin, S., Mongenet & Dischler, 2006)).

Communication is a common bound of parallel systems, being tied to system bandwidth and latency. Choices to minimise communication include moving processing to nodes containing data subsets, moving data to processing nodes or some hybrid scheme dependent upon the application (Nebel, J. C., 1998).

Our proposed distributed framework will form a basis for many image processing research projects that want to take advantage of parallel processing, and so we have chosen to build our framework on top of MPI, as it is a mature portable, standards based, ubiquitous message passing model, with well understood semantics. The framework implementation will leverage the parallel I/O capabilities of MPI-2, to improve data loading onto nodes (Peterka et al., 2008). Further improvements in this area can be obtained by de-clustering the data across a disk farm to improve access times (Scheuermann, Weikum & Zabback, 1998), but our initial focus is on real time interaction with data already loaded across the cluster.

It may appear sensible to more fully utilise the parallel resources of a GPU in our application, especially for computation within the visualization stage and thus creating a hybrid solution. However, it is expected that the bounds on processing speed will generally be inter processor and I/O communications rather than computations, justifying the deployment and processing of image data on a cluster (Ross et al., 2008). We restrict our use of a GPU's vector processing capability to manipulating and rendering already computed graphical constructs (e.g. vertex lists) generated on the cluster.

5. THE PROPOSED FRAMEWORK

Parallelism is an intrinsic feature of many physical systems. Identifying that parallelism and mapping to a parallel environment is key to optimizing the use of parallel computing resources.

The purpose of the proposed framework is to provide a simple high level abstraction to facilitate the usage of parallel processing in imaging applications by shielding the user from

any low level complexity. The image-processing pipeline is broken up into individual processing entities, which correspond to imaging filters for pre-processing, image segmentation and feature extraction as shown in figure 1.
In order to assist in this effort, we are developing a distributed imaging framework, which is based on the following:

- Imaging filters that represent dynamically distributable processing entities within the framework
- Adaptive distribution of processing entities based on efficient load balancing of processing loads.
- Load balancing strategies are based on feedback metrics that can be used to auto-tune the resource utilisation
- Effective data flow management through data allocation, partitioning and instancing.
- Instrumentation to monitor the performance of parallel programs, and provide feedback on operation and efficiency.
- A modular design that supports plug-in type imaging extensions for the development of new filters
- Debugging capabilities for the users, for instance to detect timing dependent deadlocks and minimise the impact of synchronisation problems

From a research perspective, an essential aspect of the framework is the fact that the system can create several instances of the same processing entity on demand. Concurrency is thus established and managed through processing parallel instances of data streams or data partitions. The control unit is responsible for creating and resourcing those processing entities. The framework will assist the user by taking care of such aspects that are common to many distributed applications. The system will provide data flow management, which is performed using a number of partitioning strategies, namely temporal partitioning and spatial partitioning:
Temporal partitioning is based on processing time sequences simultaneously. For example, the same noise removal filter is applied to all images in an image sequence. Spatial portioning is based on processing individual sub-volumes simultaneously from a large dataset. For example, one instance of a filter is applied to all sub volumes to achieve de-blurring. In most cases, the system will

A priori knowledge is intrinsically incorporated in the composition of the filter pipeline, which is determined by the user. The process will require no further intervention from the user as the user is building the application by defining the connections between filters, which automatically yield dependencies.

An important consideration is the level of abstraction that a framework provides. This could range from the very abstract, where the developer is less aware (or unaware) of the parallelism employed, to the low level explicit and more complex parallel programming paradigms (Skillicorn & Talia, 1998). Our framework provides a higher level abstraction (application) and a simple lower level abstraction for developing distributable filters, with research being predominantly focused on the higher level abstraction.

At the higher level, users of the framework define filter objects that implement ‘Message’ and ‘Task’ interfaces defined by the framework, and register a factory object that maps these object classes to ids. The Ids must be unique to a particular factory, but different factories can use the same ids – the framework will disambiguate. This code is then made available to the framework using a plugin architecture (i.e. as dynamically loaded shared libraries). Once a library plugin is loaded on each node, a client can send a ‘specification’ to the framework, and the framework will interpret and distribute the workflow according to the specification. The system autonomously provides information on input, output, control parameters and the particular code to run, and will resolve dependency constraints when running multiple tasks.

At a lower level the framework is to be used as a development library to create distributable image filters. The deconstruction of the objects, packing and sending using MPI’s point to point and collective operations will then be performed transparently by the framework. Similarly, on some other node, the framework will receive and reconstruct objects using the classIds as guidance, and pass to the application code. The user will not be concerned about where the code is executed, only that parallel resources are being used efficiently.

Figure 1. shows a simplified high level schematic of distributed processing with regard to pre-processing, feature extraction and visualization. The client workstation external to the cluster controls processing on the cluster. Note that rendering is shown as being performed both on the cluster (direct volume image compositing – dotted nodes), and on the workstation. OpenGL is used on the workstation to directly render pixel images composited on the cluster, and also to render generated vertex lists that describe iso-surfaces, bounding boxes etc. (OpenGL Architecture Review Board & Shreiner, 2004). This hybrid approach will allow processing entities with a high degree of communication to run on a closed coupled shared memory system, for example, the graphics memory and GPU.

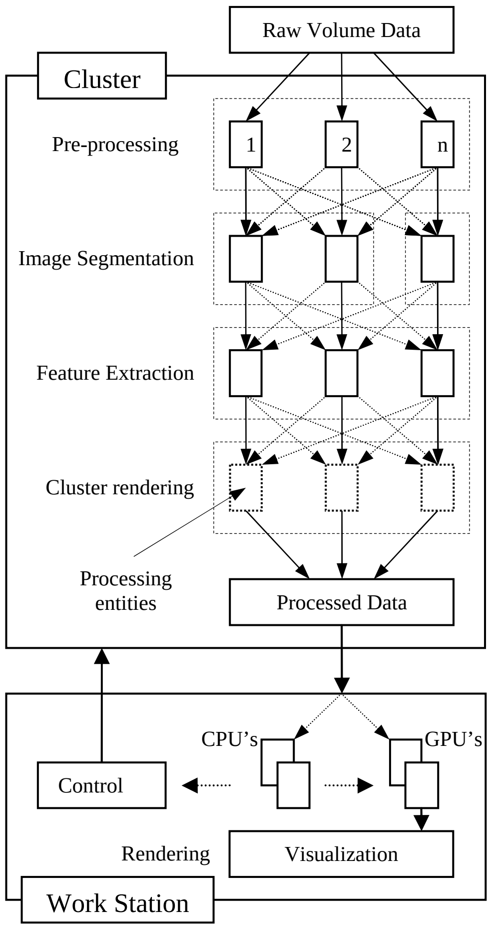


Figure 1. A Simple Schematic of the Distributed Processing Arrangement.

6. PROGRESS AND PRELIMINARY RESULTS

We have tested a prototype framework on a simple search program, using the MPICH2 implementation of MPI (Argonne National Laboratory, 2008). This parallel program uses a master/slave model. A search word or phrase is sent to all nodes. Each node then requests a parcel of work. The master then divides up the text to be searched, and sends a parcel to each requesting node. Each node processes its received work and either finds the search term, part of the search term (which will generate a further search at this point), or no search term. Results are returned to the master, which then sends further work until the full text has been processed. Investigations are ongoing to assess the impact of work packet size on efficiency using non-blocking point to point operations.

We have implemented thresholding and ray casting algorithms to detecting and render iso-surfaces within an image volume. Our technique is to form a bounding box around a 3D image, and transform that to a generic cube. Then, given the orientation of the camera, and for each pixel to be rendered, we compute whether a ray will intersect that bounding box. If it does, then a further processing stage is required. In this second step, a ray is stepped into the volume using a variant of “A Fast Voxel Algorithm for Ray Tracing” (Amana tides, J. & Woo, 1987). Normals are computed at each intersection, and the intensity value to be rendered is computed using the dot product of the normal and lighting vectors. See figure 2.

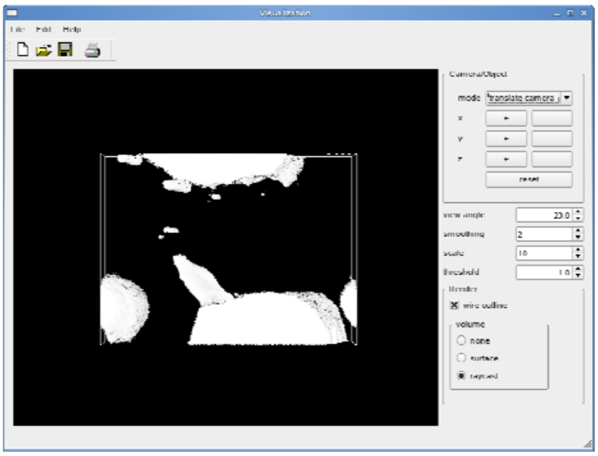


Figure 2. Volume Ray Casting to Extract an Iso-surface Using Cell Data (Prototype User Interface Build Using QT3.3 (Trolltech, 2008)).

These algorithms are currently being parallelised to use the framework. Note that in the above search program the master is sending data to each node. This will not scale, and our visualization port to parallel will isolate control instructions from image data movement to alleviate this bottleneck. In the feature extraction and analysis stages, communication patterns will be between neighbour nodes.

Visualization using ray casting is well suited to task parallelisation, individual rays or groups of rays forming the tasks. As individual rays traverse only a small subset of the image volume, it would seem that data decomposition is also well suited. However, an interesting aspect here will be to optimise the decomposition keeping in mind that the 3D image will be rotated. A partitioning that is optimal for one orientation of the image might not be optimal for another orientation, where rays are incident on different data partitions and in a different order. The framework will allow us to investigate this sort of scenario, and provide further insight into distribution and communication patterns.

Furthermore, we will decompose the image volume for the feature extraction and analysis algorithms, as well as the decomposition for visualization mentioned above. The optimal decomposition strategy for the parallelised image processing algorithms will likely be different to that of the visualization decomposition. In order to reduce communication overhead (I/O and inter-processor), and where the same dataset is used at various stages of an application, we will seek to identify any possible compromise decomposition that will maximize efficiency across those stages, and evaluate this against the optimum decomposition for each stage.

7. CONCLUSIONS

Our proposed framework aims to provide a simple high-level abstraction to facilitate the usage of parallel processing in imaging applications by shielding the user from any low level complexity. The image-processing pipeline is broken up into individual processing entities, which correspond to imaging filters for pre-processing, image segmentation and feature extraction as well as visualisation. We intend to utilise the framework’s instrumentation to gain feedback on patterns of communication in our example application, and expect to derive valuable insight into load balancing and communication pattern optimisations that will maximise efficiency in an application whose parallelisation requirements are varied. The framework should provide considerable productivity benefits to users that want to take advantage of parallel processing without the need of low level programming. The modular design of our framework will encourage its use across many image and data analysis projects that require parallel processing capabilities across distributed systems.

REFERENCES

[1] Amanatides, J. & Woo, A. (1987) A fast voxel traversal algorithm for ray tracing.

[2] Appel, A. (1968) 'Some techniques for shading machine renderings of solids.', AFIPS 1968 Spring Joint Computer Conference, pp. 37-45.

[3] Argonne National Laboratory. (2008) MPICH2. Available at: http://www.mcs.anl.gov/research/projects/mpich2/index.php Argonne National Laboratory. (Accessed: 05/10).

[4] Asanovic, K. et al. (2006) The Landscape of Parallel Computing Research: A View from Berkeley. EECS Department, University of California, Berkeley.

[5] Gao , J. et al. (2005) 'A parallel multiresolution volume rendering algorithm for large data visualization', Parallel Comput., 31 (2), pp. 185-204.

[6] Geist , A. et al. (1994) PVM: Parallel virtual machine: a users' guide and tutorial for networked parallel computing. Cambridge, MA, EUA: MIT Press.

[7] Gropp, W., Lusk, E. and Skjellum, A. (1999) Using MPI (2nd ed.): portable parallel programming with the message-passing interface. Cambridge, MA, USA: MIT Press.

[8] Gropp, W., Lusk, E. and Thakur, R. (1999) Using MPI-2: Advanced Features of the Message-Passing Interface. Cambridge, MA, USA: MIT Press.

[9] Gschwind, M. (2006) Chip multiprocessing and the cell broadband engine. Ischia, Italy. New York, NY, USA: ACM.

[10] Hege, H. C., Hollerer, T. & Stalling, D. (1993) Volume Rendering Mathematical Models and Algorithmic Aspects. Berlin, Germany: Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB).
[11] Marchesin, S., Mongenet, C. and Dischler, J. (2006) Dynamic Load Balancing for Parallel Volume Rendering. LSIIT, UMR CNRS-ULP 7005, Louis Pasteur University, Strasbourg, France.
[12] Molnar, S. et al . (2008) A sorting classification of parallel rendering. Singapore. New York, NY, USA: ACM.
[13] Nebel, J. C. (1998) A New Parallel Algorithm Provided by a Computation Time Model. Eurographics Workshop on Parallel Graphics and Visualisation. Rennes, France.
[14] OpenGL Architecture Review Board & Shreiner, D. (2004) OpenGL Reference Manual: The Official Reference Document to OpenGL, Version 1.4. 4th edn. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
[15] Peterka , T. et al. (2008) 'Parallel volume rendering on the IBM blue Gene/P', pp. 73-80.
[16] Razdan , A. et al. (2001) 'Volume visualization of multicolor laser confocal microscope data', Computers and Graphics, 25 (3), pp. 371-382.
[17] Ross , R. B. et al. (2008) 'Visualization and parallel I/O at extreme scale', Journal of Physics: Conference Series, 125 pp. 012099 (10pp).
[18] Scheuermann, P., Weikum, G. and Zabback, P. (1998) 'Data partitioning and load balancing in parallel disk systems', The VLDB Journal, 7 (1), pp. 48-66.
[19] Skillicorn, D. B. & Talia, D. (1998) 'Models and languages for parallel computation', ACM Comput.Surv., 30 (2), pp. 123-169.
[20] Trolltech. (2008) Trolltech home page. Available at: http://trolltech.com/ Trolltech. (Accessed: 11/20).
[21] University of Tennessee, Knoxville, Tennessee (1995) MPI: A Message-Passing Interface Standard. 1.1st edn. Knoxville, Tennessee: University of Tennessee.
[22] Wall, D. W. (1991) Limits of instruction-level parallelism. Santa Clara, California, United States. New York, NY, USA: ACM.
[23] Walsh, R. et al. (August 2008) White Paper: With Its New PowerXCell 8i Product Line, IBM Intends to Take Accelerated Processing into the HPC Mainstream. Global Headquarters: 5 Speen Street Framingham, MA 01701 USA: IDC (www.idc.org).
[24] Yu, H., Wang, C. and Ma, K. (2008) Massively parallel volume rendering using 2--3 swap image compositing. Singapore. New York, NY, USA: ACM.

REAL-TIME ROI-BASED BACKGROUND SKIPPING ALGORITHM FOR TEMPORAL SCALABLE VIDEO CODING

Qiong Liu
Department of Computer Science and Media Technology, Gjøvik University College
Po. Box. 191, 2802 Gjøvik, NORWAY
Email: liuq_whu@163.com

ABSTRACT

The Scalable Video Coding (SVC) standard allows flexible Region of Interest (ROI) scalability of video bit streams. It enables the user to preferably obtain the higher quality or frame rate at the desired region and background information can be skipped in temporal enhancement layers. However, improper background skipping leads to obvious display quality decrease. An adaptive background skipping algorithm for region-of-interest (ROI) scalable video coding is proposed in this paper. For real-time applications, the coding scheme is based on hierarchical P-pictures structure with zero delay. The background skipping algorithm is based on the estimation of motion activity. It dynamically decides the background content in which temporal layers can be skipped. Therefore, the skipping algorithm could make a good trade-off between video quality and bit-rate cost. Furthermore, the saving bits are used to enhance the ROI spatial quality. The experimental results indicate that the overall perceptual quality is significantly improved since human vision tends to pay more attention to the ROI.

Keywords: Region of Interest, Scalable Video Coding, Background Skipping, Video Coding

1. INTRODUCTION

Nowadays, video transmission applications require that video content can be accessed by different users from a various collection of terminals and networks. Scalable Video Coding (SVC) [1] is being developed to encode video signal once, but enables decoding from partial streams with respect to the specific rate required by a certain application. SVC support a mechanism that permits region-of-interest (ROI) scalability, which enables the user to preferably obtain the higher quality or frame rate at the desired region. The video frames are segmented into regions of interest and non-ROI regions (background) by the visual attention model. Encoding ROIs at a high frame rate can ensure smooth motion, and the background can be skipped in temporal enhancement layers to save bit-rate cost. As the background is paid less attention, and contains no complex motion and evident edges usually, a proper skipping would not cause quality decrease. The concept of adaptive background skipping has attracted a lot of attention in past literature. In [2], it proposed an optimized σ -domain bit allocation scheme to reallocate bits from non-ROI to ROI by adaptively skipping non-ROI. In [3], the motion intensity and complexity of ROI and background are computed by perceptive motion energy spectrum (PMES) to decide the ratio of the frame rate of ROI to that of background. However, to the best of our knowledge, there is little work about the background skipping based on ROI scalability. The conventional background skipping strategies does not consider about the hierarchical prediction structure of SVC and the skipping decision can not base on the temporal layer unit. In this paper, we propose a real-time ROI-based adaptive background skipping algorithm for ROI scalable video coding. It is suitable to low bit-rate and real-time applications, such as wireless video surveillance. The background skipping decision is made by layer unit. So the strategy is easy to integrate into the traditional scalable video coding scheme. The adaptive skipping decision is depend on the motion activity of background. It can make a good trade-off for the perceived quality and bit-rate cost. Furthermore, the bits saved by skipping background could be used to enhance the ROI spatial quality.

2. REGION-OF-INTEREST TEMPORAL SCALABLE STRUCTURE

For hybrid video codec, temporal scalability can generally be enabled by restricting motion-compensated prediction to reference pictures with a temporal layer identifier that is less than or equal to the temporal layer identifier of the picture to be predicted. The set of pictures between two successive pictures of the temporal base layer together with the succeeding base layer picture is referred to as a Group of Pictures (GOP).

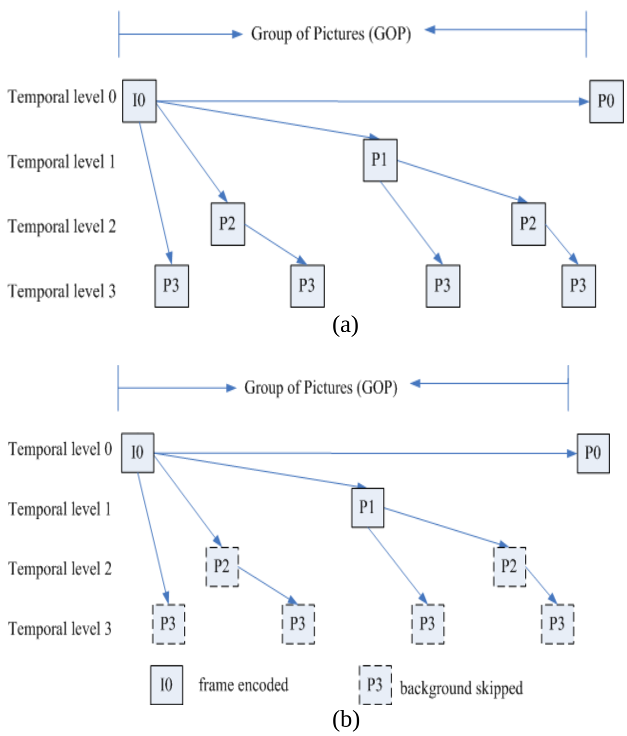


Figure 1. Hierarchical prediction structures.(a) Hierarchical P-pictures with 4 temproal layers and it with structural encoding/decoding delay of zero. (b) ROI based hierarchical P-pictures with non-ROI skipped in the two higher layers.

As illustrated in Fig. 1(a), the concept of hierarchical P-pictures [1] can provide dyadic temporal scalability. The advantage of hierarchical P-pictures is low delay. The base layer is a usual sequence of P or I pictures. The enhancement layer, however, consists only of P pictures. Since a P-picture can be decoded without backward prediction, the enhancement layer P-pictures can be decoded as soon as they are transmitted. Therefore, the structure of Fig. 1(a) has zero delay. It is valuable for the real-time video transmission. In this paper, the temporal ROI scalability is exploit the hierarchical P-pictures structure considering about the delay restriction.

As illustrated in Fig. 1(b), it is an example of skipping background in the two higher temporal layers. The ROIs are encoded at base and enhancement layers to obtain a high frame rate, while the background can be skipped in enhancement layers to save bits. It should be note that the background in lower temporal layer is forbidden to skip unless the background in higher layers is decided to skip. It is because of the hierarchical prediction dependency. At decoder, the background content of background-skipping frame is replaced by reference frame.

3. ADAPTIVE BACKGROUND SKIPPING DECISION ALGORITHM

Although skipping background in more layers could save more bits, unsmooth display or jitter will occur at decoder. Consequently, we propose a content-based adaptive algorithm to select proper temporal layers to skip background. The goal is obtain a good trade-off between the temporal perceived quality and bitrate cost.

Generally, the proper background skipping has high correlation with the temporal motion activity of sequence. If the background is almost still, the background in most enhancement layers can be skipped. Otherwise, too much background skipping will cause obvious unsmooth display. So the skipping decision should depend on the background motion activity. There are many ways to estimate the amount of motion contained in a video sequence[3~5], such as the motion compensation error (MCE), motion vector field and so on. Most of these measures need motion estimation (ME) process. The complexity is extremely high. In this paper, a low-complexity measure for motion activity is pre-computed before video encoding. Let the temporal layers be identified by a temporal layer identifier *T*, which starts from 0 for the base layer and is increased by 1 from one temporal layer to the next. For each natural number *K*, the temporal layer *K* is predicted from temporal layer with identifier *T* smaller than *K*. As the skipping background will be replaced by the reference frame at decoder, the difference between current background and reference background is a good indication for skipping. As illustrated in Fig. 1(a), in a GOP unit, the first frame of each enhancement layer is predicted from the first frame of GOP. The size of GOP is less than 32 for common applications. Therefore, the frames inside a GOP have a very tiny distance in time. In the visual system, there are temporal interactions among stimuli that do not overlap in time. This phenomenon is temporal masking [6]. Benefited from the temporal masking, human visual system could not detect the background replacement if the difference of neighboring frame is minor. Since the ordinary video motion is consecutive in a short time, the background difference between the first frame of each enhancement layer and the first frame of GOP can indicate the background difference of enhancement layer and reference layer. Therefore, the motion activity of layer T, denoted by MAT, could be generated by (1):

$$MA_T = \frac{\sum | BG_T - BG_0 |}{N} \qquad (1)$$

where BG_T is the background in the first frame of temporal layer *T*, and BG_0 is the first frame of base layer. *N* is the pixel number of background region. Therefore, MA_T indicates the average pixel difference. It makes the skipping algorithm could be adaptive with different size of background area. The background in temporal layer *T* could be skipped if (2) is satisfied and its higher layers are already decided to skip.

$$MA_T < \eta \qquad (2)$$

According to a mount of objective and subjective results, the value of η can be set as 10 to achieve a good trade-off for performance. The adaptive background skipping algorithm could be performed for each GOP. The overall flow of proposed algorithm is showed in Figure 2.

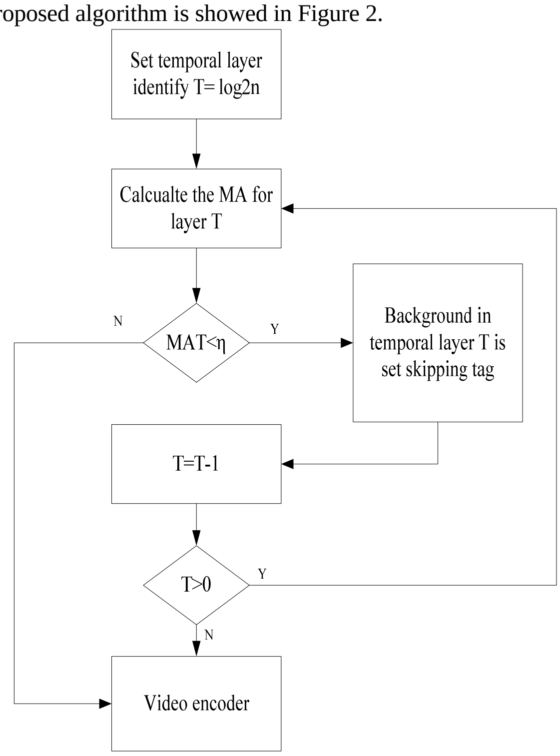


Figure 2. The Overall Flow Of Proposed Algorithm

4. EXPERIMENTAL RESULTS

A number of experiments have been conducted with the CIF sequences *akiyo*, *foreman*, and *coastguard*, which stand for sequences from light motion to big motion. We choose the region of face as ROI in sequences *akiyo* and *foreman*, and choose the region of moving boat as ROI in sequence *coastguard*. The region could be changed frame by frame as the object moving. Only the first 100 frames of these sequences are used. We implemented our codec using the proposed background skipping scheme and the basic coding algorithms specified in H.264/SVC reference model. The number of reference frame is 1 and GOP size is 4, so there are 3 temporal layers in total.

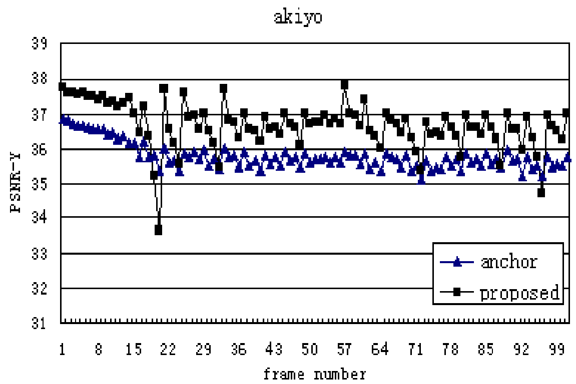
The objective performance is represented by the *Peak Signal to Noise Ratio* (PSNR) of output video of decoder. The "anchor" is the reference model without background skipping algorithm. From Fig. 3 (a), we can see that proposed algorithm could obtains about 1~2db gains for the sequence *akiyo*. That is because the background data in enhancement layers is skipping and the saving bits is used to promote the spatial quality. In Fig. 3 (b) and (c), the curve of proposed algorithm is changed a lot compared with the anchor. The proposed algorithm could make a proper decision on skipping for different video content. The differences of neighboring frames will not affect the subjective quality obviously because of the temporal masking of HVS. Human could not perceive the changes in so little time interval. The PSNR value drops because the background replacement. It not equal to worse perceived quality in this situation. On the other side, because the background of reference frame has a higher quality, the replacement could achieve better subjective quality to anchor.

5. CONCLUSIONS

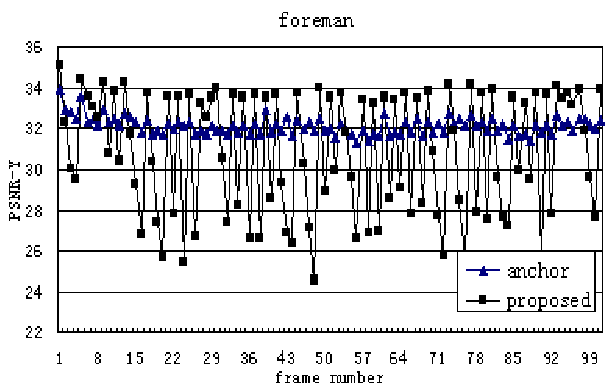
In this paper, a content-based background skipping algorithm for ROI scalable video coding is proposed to achieve better perceived quality with less bitrate cost. The ROI temporal scalable structure exploits a zero-delay hierarchical structure. And the average difference between the enhancement temporal layer and base layer in a GOP is used to indication the motion activity of scene. A threshold is used to decide the background in which temporal level could be skipped. The proposed algorithm could be adaptive with different video content and achieve better perceived quality.

AKNOWLEDAGEMENT

Most of the work has been done when the first author was affiliated with Wuhan University. This work was supported by Nation Natural Science Foundation of China (NO. 60772106).



(a)



(b)

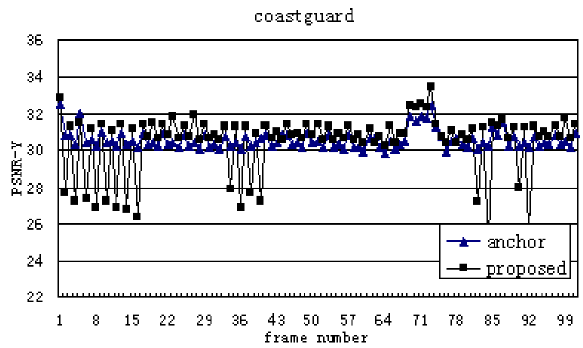


Figure 3. Comparison Result of PSNR for Each Frame at the Same Bitrate Condititon

REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 17, pp. 1103-1120, 2007.

[2] W. Haohong and K. El-Maleh, "Joint Adaptive Background Skipping and Weighted Bit Allocation for Wireless Video Telephony," in Proceeding of International Conference on Wireless Networks, Communications and Mobile Computing, 2005. pp. 1243-1248.

[3] L. Wei, G. Xiao-Dong, W. Ren-Hua, D. Li-Rong, and Z. Hong-Jiang, "A region based multiple frame-rate tradeoff of video streaming," in Proceeding of International Conference on Image Processing, 2004. Vol. 3.pp. 2067-2070.

[4] M. Yu-Fei and Z. Hong-Jiang, "A new perceived motion based shot content representation," in Proceeding of Image Processing, 2001. Proceedings. 2001 International Conference on, 2001, vol.3.pp. 426-429.

[5] F. Pan, X. Lin, S. Rahardja, "Proactive frame-skipping decision scheme for variable frame rate video coding," in in Proceeding of International Conference on Multimedia and Expo, 2004, Vol.3.pp. 1903-1906.

[6] R. N. Habe, The Psychology of visual Perception: New York Holt, 1973.

Qiong Liu Received Ph.D degree in Computer School of Wuhan University, 2009. She is a researcher in Gjøvik University College, focusing on media technology. She was with National Engineering Research Center for Multimedia Software from 2003 to 2008, severed as Ph.D candidate. She worked actively for Surveillance Video and Audio Standard of China. Her research interest includes optimization of H.264, scalable video coding and region-of-interest video coding.

AN IMPROVED IMAGE DENOISING ALGORITHM BASED ON GREY RELATIONAL ANALYSIS *

Gang Li[1,2] , Xinping Xiao[1] , Rui Chen[2]
[1]School of Science, Wuhan University of Technology
Wuhan, Hubei 430063, China
[2]School of Science, Hubei University of Technology
Wuhan, Hubei 430068, China
Email: 87475853@qq.com

ABSTRACT

The problem of image denoising is a primary operation in the stage of image pre-processing. In recent years, combining the image filter with grey system theory, some scholars present an adaptive added-weigh mean filter algorithm which is a optimizing process to obtain the value of weigh decided by comparing the similarity between the grey levels of sample pixels in the filter window and the mean grey levels of pixels in the window based on the grey relational degree. Compared with the traditional mean filter, it has achieved certain progress. However, because of the defects of the algorithm of its own, when the noise density reaches a certain level, the processing results are unsatisfactory. The main part unreasonable is that when filtering the image we do not make a clear distinction between the noise points and normal ones, which will inevitably lead to the edge blurred phenomena. In this paper, we first use a new detector to distinguish between the noise and normal points. Then we just do filter operation to the pixels which have been sentenced to noise points using an improved filter algorithm based on grey relational analysis. Finally, the experimental results demonstrate that the algorithm improved is significantly better than the classical mean filter, median filter and the traditional image filter algorithm based on the grey relational analysis.

Keywords: Image Denoising, Grey Relational Analysis, Noise Detect, Neighborhood Window, Filter.

1. INTRODUCTION

The image denoising has been a hot topic in the field of image processing. Many scholars have put forward all kinds of algorithms on the noise reduction, which can be summed up in two categories: linear filters and nonlinear filters. The representative of the former is mean filter which is mainly used for the elimination of Gaussian noise, and the latter was represented by median filter, which is mainly used to eliminate impulse noise. Most of the current algorithm is improved and perfected based on the two filters [1-4]. In recent years, with a number of new academic disciplines and interdiscipline emerging and developing, in particular the flourishment in system science has brought the information science new vitality and energy. As image information of its own is grey, it comply with the requirements of uncertainty that “Some of the information known, partial information unknown”. Some scholars apply the grey system theory to the field of image processing and have achieved good results [5-11]. Based on the combination of grey relational analysis(GRA) and median filter, the literature[12-13] give us a new method to deal with the salt and pepper noise in the image. Considering the work done by the scholars before, we improve the algorithm and try to make it more perfect.

The paper is organized as follows. Section II describes the traditional added-weigh mean filter method. In section III, the defect of the method before is represented, and we analyze the cause of it and make it improved. In section IV, the performance of the proposed method is verified using the MATLAB, and the experimental result show that our new method is correct and rational. Finally, conclusion is drawn in section V.

2. THE TRADITIONAL ADD-ED MEAN FILTER METHOD

2.1 The Basic Conception of GRA

The procedure of a general grey relational analysis consists of the following 4 steps [12]:

Step 1:

Select factors $X_0^{(k)}$, $X_i^{(k)}$ as the reference sequences and comparative sequences respectively, namely,

reference sequence:

$$X_0^{(k)} = \left\{ X_0^{(1)}, X_0^{(2)}, \cdots, X_0^{(k)} \right\} \tag{1}$$

compare sequence:

$$X_i^{(k)} = \left\{ X_i^{(1)}, X_i^{(2)}, \cdots, X_i^{(k)} \right\} \quad \text{for} \quad i = 1,2\cdots,\text{n}. \tag{2}$$

Step 2:

In order to facilitate comparison of the various sequences, all the data should be in the same order of magnitude. We usually adopt mean method, namely,

$$X_i^{(k)} = \frac{X_i^{(k)}}{\overline{X_i}}, \quad \text{where} \quad \overline{X_i} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(k)} \tag{3}$$

Step 3:

For the multiple sequences, compute the different sequence

$$\Delta_{0i}^{(k)} = \left| X_0^{(k)} - X_i^{(k)} \right|, \tag{4}$$

maximum difference of two grades

$$\max_i \max_k \left| X_0^{(k)} - X_i^{(k)} \right|, \tag{5}$$

and minimum difference of two grades

$$\min_i \min_k \left| X_0^{(k)} - X_i^{(k)} \right|. \tag{6}$$

Step 4:

Calculate the grey relational coefficients $\varepsilon_{0i}^{(k)}$ according to Professor Deng’s method:

$$\varepsilon_{0i}^{(k)} = \frac{\min_i \min_k \left| X_0^{(k)} - X_i^{(k)} \right| + \rho \max_i \max_k \left| X_0^{(k)} - X_i^{(k)} \right|}{\left| X_0^{(k)} - X_i^{(k)} \right| + \rho \max_i \max_k \left| X_0^{(k)} - X_i^{(k)} \right|} \tag{7}$$

Step 5:

Let grey relational coefficient as corresponding weigh, and we can obtain the added-weigh mean value $y(i, j)$ of the sequence $\{x(i-1, j-1), x(i-1, j), \cdots, x(i+1, j), x(i+1, j+1)\}$ as below:

$$y(i,j) = \frac{\varepsilon_1 x(i-1,j-1) + \varepsilon_2 x(i-1,j) + \cdots + \varepsilon_8 x(i+1,j) + \varepsilon_9 x(i+1,j+1)}{\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \cdots + \varepsilon_7 + \varepsilon_8 + \varepsilon_9} \tag{8}$$

* This work was supported by Specialized Research Fund for the Doctoral Program of Higher Education (Grant No.200804970005)

2.2 The Added-weigh Mean Filter by Using GRA

In the traditional mean filtering process, the noise points and non-noise points are given the same weight, which is unreasonable. As a result, a filter method of automatically adjusting the weight associated with development of the grey system theory came into being [12].

Suppose an image with the size of $M \times N$ can be expressed as $f(i,j)(i=1,\cdots,M;j=1,\cdots,N)$ .For any pixel of the image excluding some pixels which is located on the edge of image frame, $f(i,j)(i=1,\cdots,M-1;j=1,\cdots,N-1)$, select the pixels around its neighborhood as a window for

$$\begin{bmatrix} f(i-1,j-1) & f(i-1,j) & f(i-1,j+1) \\ f(i,j-1) & f(i,j) & f(i,j+1) \\ f(i+1,j-1) & f(i+1,j) & f(i+1,j+1) \end{bmatrix}. \tag{9}$$

Let the average value of pixels in the window be v, that is

$$\begin{aligned} v = mean\{ & f(i-1,j-1), f(i-1,j), f(i-1,j+1), \\ & f(i,j-1), f(i,j), f(i,j+1), \\ & f(i+1,j-1), f(i+1,j), f(i+1,j+1)\} \\ = & \frac{1}{9}\sum_{t=i-1}^{i+1}\sum_{s=j-1}^{j+1} f(t,s). \end{aligned} \tag{10}$$

Suppose reference sequences are

$X_0 = \{v,v,v,v,v,v,v,v,v\}$, and comparative sequences are

$$\begin{aligned} X_1 = \{ & f(i-1,j-1), f(i-1,j), f(i-1,j+1), f(i,j-1), \\ & f(i,j), f(i,j+1), f(i+1,j-1), f(i+1,j), f(i+1,j+1)\} \end{aligned} \tag{11}$$

As the comparative sequences here is single dimensional, we can get the difference sequences:

$$\Delta(k) = |X_0(k) - X_1(k)| \; (k = 1,2,\cdots,9) \tag{12}$$

Maximum difference of two grades:

$$M = \max_k \Delta(k) \tag{13}$$

Minimum difference of two grades:

$$m = \min_k \Delta(k) \cdot \tag{14}$$

The grey relational coefficients are

$$\varepsilon_k = \frac{m + \rho M}{\Delta(k) + \rho M} \tag{15}$$

for $k = 1,2,\cdots,8,9$, where ρ is a discrimination coefficient, generally set by $\rho = 0.5$.

Let the grey relational coefficients be the weight of the pixels in the window, and replace weigh of the traditional mean filter with ε_i , and let the sum of these weigh substitute denominator of the filter, we get

$$\hat{f}(i,j) = \frac{\varepsilon_1 f(i-1,j-1) + \varepsilon_2 f(i-1,j) + \cdots + \varepsilon_8 f(i+1,j) + \varepsilon_9 f(i+1,j+1)}{\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_8 + \varepsilon_9} \tag{16}$$

2.3 Shortcomings of The Traditional Filter Algorithm Based on GRA

Compared with the traditional mean filer, the filter based on grey relational analysis above is to calculate the grey relational coefficients as the weigh for each pixel in the window rather than make the ordinary arithmetic mean value of all the pixels in the window. Generally speaking, the normal pixels comply to a certain degree of continuity in the spatial distribution; when the noise has low density, the mean value of pixels in the window is very near to the true value, then the filter degenerate into the traditional mean filter. As we select the mean value as the reference sequences, the pixels which are more approximate to the mean value can get the bigger coefficient, that is they occupy a greater weight; whereas the pixels which deviate from the mean value can acquire a less coefficient, that is, they occupy a less weight. To optimize grey levels in the filter window, the algorithm make the weigh of the noise point changed dynamically. It can reduce the influence of noise point on the result image. Experimental result demonstrates that the algorithm can be more efficient in the noise reduction than traditional mean filter algorithm, and it is a feasible method. However, this method is generally applicable to the image with lower density noise. When the noise reaches a high density, it fails to deal with it. By analyzing the principle and mechanism of the algorithm, it is easy to find the limitations of the algorithm itself:

1) Reference sequence selected was the mean of pixel in the window. When the noise density become larger, the average of the pixels in the window will gradually deviate from the true value of the pixel in the center of the window, and the greater the noise density is, the more serious the deviation is. So choosing the average value of pixels in the window for reference sequences can not accurately portray the essential characteristics of the central pixel;

2) The pixels selected as comparative sequence are all the pixels in the neighborhood of the central pixel in the window. It is clear that using the pixels containing a lot of noise to forecast the value of the central pixel is unscientific when there is high-density noise distribution in the window.

In view of this, this paper analyzes the mechanism of the realization of the algorithm, and puts forward a novel filter method based on grey relational analysis, which overcomes the above-mentioned two defects, and greatly enhances the effectiveness of the algorithm to deal with. Especially for high-density noise pollution it has made even more superior results.

3. AN IMPROVED IMAGE FILTER ALGORITHM BASED ON GREY RELATIONAL ANALYSIS BY USING THE NOISE DETECTION METHOD

3.1 The Background of the New Algorithm Proposed

The main problem existing in the algorithm described above is that we do not discriminate between normal pixels and noise pixels, which is contrary to the original intention of the algorithm. When the noise density is low, this error is not particularly visible, but when the noise reaches a certain level, relying on grey relational analysis algorithm itself to adjust the weigh of pixels dynamically is far from sufficient. So the noise detection operation is added before we make grey relational analysis, and improved operation is indispensable.

In the traditional mean filter method, we treat all the pixels as equal no matter whether they are noise or not. Therefore, when the central pixel in the window is a normal point, the mean filter is bound to result in image distortion; when it is noise point, the use of the pixels mixed by noise in the neighborhood of window will inevitably lead to significant forecast error. In the added-weigh mean filter algorithm based on grey relational analysis, the grey value of pixels close to the average value in the window have the right of access to major weight, vice versa. This is an added-weigh mean method by using calculating the relational coefficients. Since the range of values of relational coefficients is $(0,1]$, that is the impact from noise pixels on the results will never equal to zero, the result of an error leaded by algorithm itself is inevitable.

3.2 A Useful Method of Noise Detection Carried out Before Filtering

In this paper, the improvement points are as follows:

Firstly, a method is designed to detect the noise pixels[1].

(a) As the noise is usually the extreme value, which is located at the two ends of the range [0,255].Tests shows that the grey value of noise changes in the range of [0 10] or [245,255], that is

$$f(i,j) \in [0,10] \cup [245,255] \quad (17)$$

(b) In the neighborhood of pixels in the 5×5 window, there are 25 pixels in total. The pixels sorted can be expressed as $f_1, f_2, \cdots, f_{24}, f_{25}$. If each of 25 pixels locates before T or after T, which means it is extreme value, we can judge the pixel is a possible noise point(After extensive testing, for the 5×5 window, we can let T be 10),that is

$$f(i,j) \in [f_1, f_{10}] \cup [f_{16}, f_{25}]. \quad (18)$$

(c) Once the pixel in window meets the two conditions above (17) and (18) simultaneously, we can judge it is noise point, or it is determined to be normal point. Let a tag matrix F with the same size as the original image, and initialize F, which make all the elements zeros. When the image is detected from up to down and left to right, once $f(i,j)$ is found to be noise, we can change zero in the F to one in the corresponding location, that is $F(i,j)=1$, which can be used as a tag of noise.

The main process is as follows:

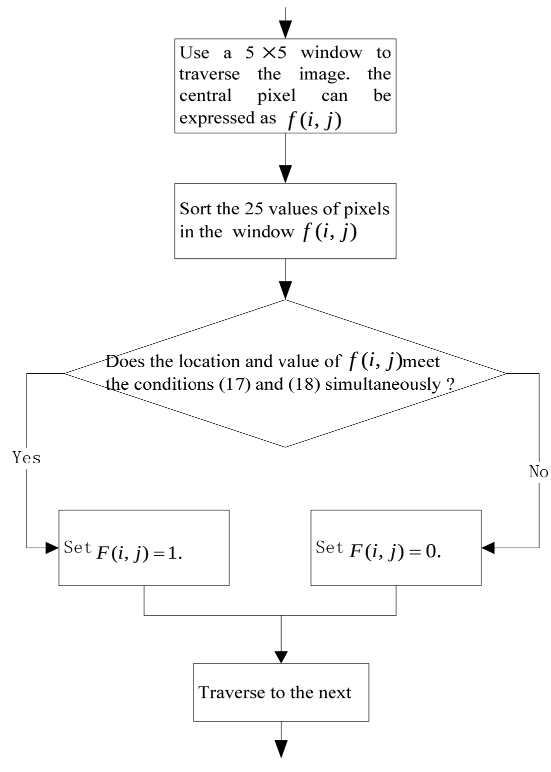


Figure 1. Noise Detection Flow Chart.

3.3 The Whole Detailed Steps of the Improved Method

The entire process of algorithm can be divided into two stages, which is described as follows:

The fist stage is to detect the noise:

Step1, traverse the image from up to down and left to right, and use the algorithm described above to detect whether the pixel is noise point or not .Meanwhile, use zeros to express normal point and one to express noise point in the tag F .

The second stage is to filter the image:

For each pixel in the image, if it is a normal point due to the conclusion acquired from the first stage, nothing need to be done with it; or it is decided to be noise point, we can do it like these:

Select the normal points around its neighborhood of the central pixel in the 3×3 window as the comparative sequences, and calculate the median value of these normal point rather than mean value of all the points in the window.

In fact, for each central point $f(i,j)$ in the 3×3 window, when we traverse the image in an orderly manner along the direction of up to down and left to right, some pixel points on the up and on the left of the window, such as $f(i-1,j-1)$, $f(i-1,j)$, $f(i-1,j+1)$, $f(i,j-1)$,which have been dealt in steps before, are real-time updates. It is clear that they are the new data including no noise, so we should make full use of these points instead of the initial data.

Let $f(i,j)$ express the initial pixel, then the new data dealt already in the same location can be given by $\hat{f}(i,j)$.

For example, suppose a window is given by Eq.(9), the corresponding window in tag matrix is

$$\begin{bmatrix} \hat{F}(i-1,j-1) & \hat{F}(i-1,j) & \hat{F}(i-1,j+1) \\ \hat{F}(i,j-1) & F(i,j) & F(i,j+1) \\ F(i+1,j-1) & F(i+1,j) & F(i+1,j+1) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (19)$$

Then, reference sequences and comparative sequences are given by

$$X_0 = \{v, v, v, v, v, v\} \quad (20)$$

where

$$v = median\{f(i-1,j-1), f(i-1,j), f(i-1,j+1), f(i,j-1), f(i+1,j-1), f(i+1,j+1)\}$$

$$X_1 = \{\hat{f}(i-1,j-1), \hat{f}(i-1,j), \hat{f}(i-1,j+1), \hat{f}(i,j-1), f(i+1,j-1), f(i+1,j+1)\} \quad (21)$$

Other steps can be carried out in accordance with the requirements of Section II.

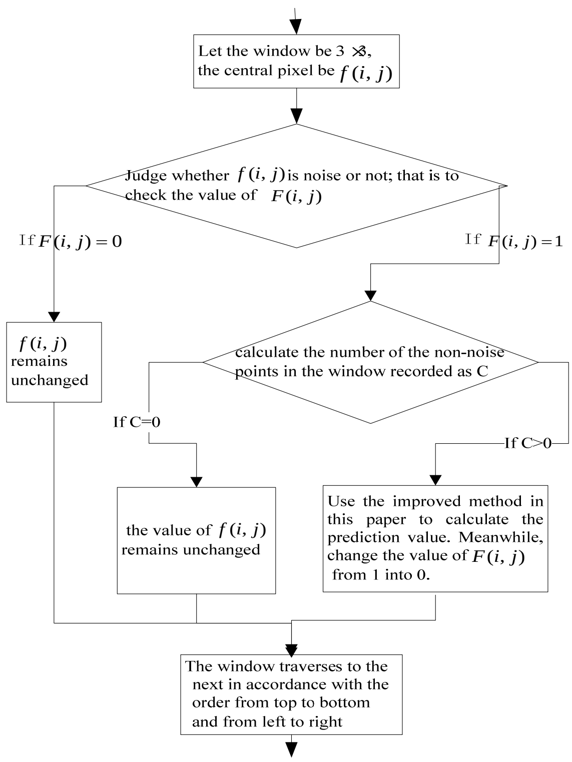


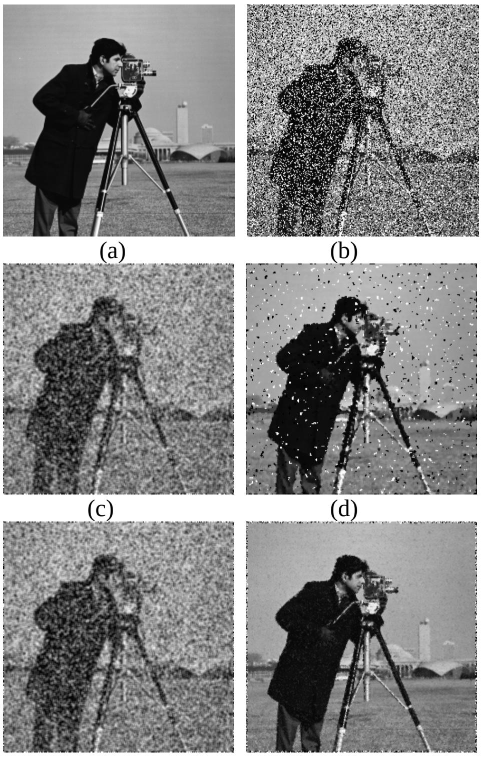
Figure 2. The Main Process of the Whole Algorithm Improved in this Paper.

Note that the tag matrix F itself is not involved in the filter operation, and it just only set for a tag or mark to help us identify the normal pixels. At the end of the operation, F will change from a matrix whose elements are all one into a null one.

4. SIMULATION STUDY

By using MATLAB, the image named "cameraman.tif ", which contains salt & pepper noise ,is processed. We add all kinds of density noise to the experimental image, and several filters are used to compare the effectives of algorithms.

(1) Subjective evaluation
Evaluate whether an algorithm for a kind of image effective or not, we just need to put experimental images together, and observe the difference between them. The images below are all polluted by the same kind and rate of noise, and they are dealt with the traditional mean filter, median filter, added-weigh mean filter based GRA, and the improved method proposed in this paper respectively except the original image.



(a) (b) (c) (d) (e) (f)

Figure 3. Cameraman images:(a) original image; (b) Image with noise; (c)Image dealt with the mean filter;(d)Image dealt with median filter; (e)Image dealt with the added-weigh mean filter based on the grey relational analysis in[12];(f)Image dealt with the improved filter based on the grey relational analysis proposed in this paper.

As shown from the experimental result (Fig.3): The original image in Fig.3.(a) without any noise is very clear, and seems very satisfactory; The image containing noise in Fig. 3.(b) is not smooth, and the noise seems very sharp, and the quality of the entire menu is the worst; As observed in Fig. 3.(c), the noise is reduced with the traditional mean filter, and the image becomes smooth. Grey levels in the image change relatively not strongly, but it makes edges of the object blurred; In Fig. 3.(d), the noise in image is decreased in number but bringing some new big noise; In Fig. 3.(e), the whole image blurred, which illustrates that the filter based on the GRA is not very fit for the images polluted seriously, although it may be successful when the noise density is light; Effect in Fig. 3.(f) is better than that in Fig. 3.(e), which shows that the method using an improved adaptive grey relational degree not only remove the noise but also strengthen the edge of the object, and has made the most satisfactory effect.

(2) Objective evaluation
In order to evaluate the effect of methods above objectively, there were many kinds of methods put forward ago. Among them, a universally recognized approach has been done by computing the Peak Signal/Noise Ratio (PSNR) as criterion to evaluate the performance of filters. The specific approach is like this: set $s(i,j)$ to be the original image without any noise, $r(i,j)$ to be the image made by denoising, and the size of the image is $M \times N$, thus the PSNR formula is given by

$$PSNR = 10\lg \frac{255^2}{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[s(i,j)-r(i,j)]^2}. \qquad (22)$$

PSNR can be used to quantify the effect that the quality of image processing. Generally speaking, the greater the PSNR is, the higher the quality of image denoising is. We can compare the value of PSNR to find the better one in various methods. In table1, we give the experimental result that the value of PSNR after processing by the four kinds of filters when the cameraman image is added the rates of 0.1, 0.3, 0.4, 0.5,0.7 noise.

Table 1. Comparison on PSNR among mentioned algorithms under different noise intensity.

Noise density	0.1	0.3	0.4	0.5	0.7
The mean filter	21.5884	17.1465	15.7211	14.4482	12.6545
The median filter	26.0781	20.9594	17.3820	14.1203	9.5031
The filter based on GRA	21.5849	17.1403	15.7163	14.4412	12.6489
The improved filter in this paper	28.9377	23.8500	22.3768	20.8556	18.0208

From table 1, we can see that, although the values of PSNR decrease with the noise increasing, the improved method proposed in this paper have the biggest PSNR value in all the algorithms when the rate of noise is equal, which shows the new algorithm is superior to any filter listed here.

5. CONCLUSIONS

Applying the GRA into image enhancement is a feasible direction, and a large progress has been made in this area, but there are still some shortcomings existing here to be overcome. From the mechanism of GRA, this paper improves the traditional added-weigh filter based on GRA. Some innovation can be summarized as follows:

(1) It proposes a noise detection stage before the filter is carry out. We just deal with the pixels decided to be noise point in order to decrease the noise, and left the normal pixels nothing to be done for keeping the details in the edge and preventing blurred;
(2) We choose the effective data (normal pixel point) as

comparative sequences in the neighborhood of central pixel rather than all the pixels in the window;

(3) Let the median value of the comparative sequences be the reference sequences rather than the mean value of that, for the former is more close to the true value than the latter;

(4) When calculating the value of central pixel in the window, we make full use of the pixels on the top and left sides in the window for they have been dealt and have no noise.

REFERENCES

[1] Wang Xiangju, "The Algorithm Research of Image Denoising Based on Median Filter and Wavelet Transform", Xi'an: Xi'an University of Science and Technology of outstanding master's thesis, April 2008, pp.44-48.

[2] Samuel P. Kozaitis, and Sunghee Kim, Denoising using Higher-order Statistics, Proc. SPIE, Vol. 5102,2003, pp.37-44

[3] A. B. Hamza, and H. Krim, Image Denoising: A Nonlinear Robust Statistical Approach. IEEE Trans. on Signal Processing, 49(12), 2001, pp.3045-3054

[4] Wang Hongmei, Research on the Technology of Multisensor Image Fusion and Its Preprocessing. Northwestern Polytechnical University, Dissertation for the Doctor Degree. Xi'an. 2005

[5] D. Z. Feng,J. Yan and P. F. Huang, "A novel De-noise Method Based On the Grey Relative Analysis." Proc. 2006 IEEE,ICIA, pp.1385-1389

[6] Miao Ma, YanNing Zhang,Hongpeng Tian, "Grey Theory based Evaluation for SAR Image Denoising", 2008 Congress on Image and Signal Processing,2008, pp.608-612

[7] Ma Miao,Zhang Yanning, Sun Li,Yuan Hejin,Zhou Tao, "SAR Image Despeking Using Grey System Theory", Nanjing,Chia: Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services, November 18-20,2007, pp.458-462.

[8] LU Yao,WANG Jing, HU Lei, "A Survey on Application of Grey System Theory in Image Processing" , Automation & Applicatioan, Vol.26,No.6, 2007, pp.49-64.

[9] MA Miao, TIAN Hongpeng, ZHANG Yanning, "Research Progress on the Application of Grey Theory in Image Engineering", Journal of Image and Graphics,Vol.12,No.11,Nov.,2007,pp.1944-1955

[10] Ma Miao, Fan Yangyu, Xie Songyun, et al. A Novel Algorithm of Image Edge Detection based on Gray System Theory. Chinese Journal of noises. Image and Graphic (Chinese Version). Vol.8,2003, pp. 1136-1139

[11] Zhang zi-hua, Chen jia-zhen,Chen Li-yong, " An Algorithm of Image Edge Detection based on Gray Correlation Analysis", Journal of Fujian Normal University(Natural Science Edition,Vol.20,No.4, pp. 20 -23

[12] Gui, YF; Chen, GT; Dong, H; Li, G; Lu, SQ ; Tao, JF; Xu, XY, "The Noise Reduction of Digital Image Based on Grey Relational Aanalysis", SDEC 2006, pp. 1423-1425

[13] Zhang Su-wen, Chu Nai-qiang, "a filtering algorithm Based on Grey Correlative Degree for Salt-Pepper Noise Images",Infrared Techonology,Nol.30, No.11, 2008, pp.651-654

Li Gang is a doctoral student in Wuhan University of Technology. He is also a teacher in Hubei University of Technology and he is deputy director of Department of Computer Accounting. He graduated from China Three Gorges University in 2003; from Wuhan University of Technology in 2006 with specialty of Applied Mathematics. He has participated in two National Natural Science Foundation project, and has published over 10 papers in core journals. His research interests are in distributed parallel processing, image processing, and grey system theory.

Xiao Xinping is a full professor and also a tutor of a Ph.D student, dean of the science college of Wuhan University of Technology. He graduated from Huazhong Normal University in 1990, from Huazhong University of Science and Technology in 2002 with a doctor's degree of system analysis engineering. He is a holder of the Award of Huo Yingdong education endowment which is just for the elite of the young teacher, Visiting professor of university of Manitoba from 2006 to 2007. He is one of the organizer of scholar of operational research and management science conference, was a chairmen of 2007 IEEE International Conference on Grey System and Intelligent Services. He has published two books, over 60 Journal papers, edited two academic journals, one is《Journal of Grey System》.His research interests are in Operations research and management science, prediction theory and technology, the number of economic analysis and decision-making.

PUBLIC MONITORING PHOTOGRAPH: DEFORMATION AND CORRECTION

Song Li，Xiangjiang Dong, Tao Tang ,Jiangning Shi
Nanjing Public Security Bureau,
Nanjing, Jiangsu, China
Email: li_song1972@sina.com

ABSTRACT

In area of investigating and solving criminal cases, the suspect image taken by public monitoring, particularly taken by ATM camera is somewhat deformed, making it hard to identify the suspect. Based on experiments and theoretical analysis, we can explore the causes and correction of the image deformation by monitoring camera includes size and spherical deformations. By correcting the parameters we are able to achieve the desired effect of identifying the suspect.

Keywords: Public monitoring, Suspect image, Experiment, Deformation, Correction

1. INTRODUCTION

With the increasing installation and application of public monitoring, monitoring plays an even more important role in investigating and solving criminal cases. Nevertheless, in reality, we often find that the suspect image taken by the monitor camera is somewhat deformed, particularly the image taken by ATM cameras, which is so dramatically deformed that it often makes it hard to identify the suspect.

A better image will contribute to improved identification if we work out how the deformation is caused and what can be done to make the necessary correction. A number of causes can be listed which result in deformed image taken by monitor cameras; both the external environment and the device may precipitate deformation. This paper addresses the causes and correction of image deformation of public monitor devices by discussing ATM monitoring.

2. AN EXPERIMENT OF ATM IMAGE DEFORMATION

The experiment involves a self-made plank as image source. Fix square rulers on a plank, forming a scaled square frame. A cube is stuck at the center of the square. The cube is also scaled by using similar square rulers. Thus, the plank has two square frames with the smaller one relieved against the larger one. ATM monitor records the different images of the plank taken from varying angles, directions as well as distances. Another set of data are collected with a different ATM monitor for the sake of comparison.

Experiment results

A. Compared with the real object, the recorded ATM image is deformed both horizontally and vertically. The extent of the deformation is related to the recording size of the ATM. The usual recording size is 352mmX288mm, with the result that the vertical dimension is stretched 10 percent more than the horizontal dimension.

B. ATM image also produces dramatic spherical deformation. The closer the distance, the more obvious the deformation. In addition, the edges of the image are more clearly deformed than the central part of the image.

3. THEORETICAL ANALYSIS OF THE IMAGE DEFORMATION

A. The focal length of most present ATM monitors is not more than 3.7mm; and the focal length of the fish-eye lens is usually 6-13mm. From our knowledge of the relationship between focal length and image, ATM image-recording is apt to generate marked spherical deformation. The closer the distance, the more deformed the image is; and the edge parts are progressively deformed.

B. The video camera principle stipulates that the ideal initial image should have horizontal to vertical (H/V) ratio of 4:3. This H/V ratio guarantees that the outcome image is true to the real object. In other words, there is no distortion of size. Nowadays, the terminal output sizes of public monitoring cameras are numerous while the dominant size is 352mmX288mm. Take this size for example. The initial output image is 10 percent longer. The image is more or less deformed in size if output ratio is anything but 4:3.

4. IMAGE CORRECTION EXPERIMENT

A. From the images taken by the above ATM monitor camera, we used the Photoshop tool to change the image size of those images of scaled object with a front view so that the revised (H/V) ratio is 4:3. The consequent result was the true-to-life images.

B. After the images have been corrected in terms of H/V ratio, we again used Photoshop spherical correction tool to make a maximized spherical correction. Experiment findings showed that great improvement was achieved in spherical correction. For more exact rectification of the images, other relevant parameters are needed.

5.THEORETICAL ANALYSIS OF THE DEFORMATION CORRECTION

A. For the size deformation caused by the image output facility, change the terminal output size ratio to 4:3, the initial image ratio, and the size deformation correction is well done.

B. For the spherical deformation caused by focal distance, if the front view of the object is perpendicular to the optical axis, there is well-developed solution that makes possible the relatively precise spherical correction. If, however, the object is at an angle to the optical axis, the degree of difficulty of correcting responds to the angle of inclination. Provided this, it is still possible to dramatically improve the correction effect.

6. DISCUSSIONS

A. The sensitive target of the monitor CCD chip is set the H/V ratio as 4:3, implying that monitored image has the initial size ratio of 4:3, and the image is normal. After the initial image has been encoded and compressed by the control part to make it terminal monitor document, this document is formatted into varying sizes according to the different equipment, with 352X288 being the dominant size. Thus, H/V is rarely 4:3 and the initial image is altered. The photograph taken from such document is therefore deformed. To correct the deformation, it is only necessary to restore the size of the photograph to the initial ratio of 4:3 (for example, change 352X288 to 400X300). When the ratio is the same with the initial ratio, the deformation is eliminated.

B. The lenses of most monitors have short focal distance, with ATM not more than 3.7mm. The shorter the focal distance, the more is spherical deformation. This means that spherical deformation in the monitors is only a matter of degree. Moreover, this kind of deformation is also closely connected with shooting distance and location in the photograph. ATM has very short shooting distance, causing apparent deformation. Other monitors generally have longer shooting distance and the deformation is less severe. The spherical deformation has a limited negative effect in practice, which has been confirmed in our work. Therefore, as far as the spherical deformation is concerned, the main problem that needs to be solved by the police is ATM monitors.

C. ATM monitors are apt to cause extensive spherical deformation that can be corrected by applying deformation correction technique to approach image of object with plane perpendicular to the optical axis, including photoshop spherical correction function. In reality, the image of the suspect—the facial view may not necessarily be the front view, perpendicular to the optical axis. This leads to varying deformation owing to varying distance of the face to the camera, which, in turn, brings not a little difficulty in correction. The photograph taken is mainly used for the purpose of identifying the suspect, the correction technique is sufficient to satisfy the need. Experiment demonstrates that, for the ATM photograph, substantial improvement in deformation correction is obtainable .Although it is not possible to completely eliminate deformation, the correction is good enough for the benefit of identification of the face.

7. CONCLUSIONS

The image taken by most of public monitoring has the size deformation, and it can be completely corrected. ATM camera produces the serious spherical deformation, and it is hard to be completely eliminated, but the correction is good enough for the benefit of identification of the face.

Attention: In the process of correction, the photograph should not be artificially cut or trimmed before the completion of image correction. And size correction should precede spherical correction.

8. PRACTICAL APPLICATIONS

Case one: Two scaled cubes of differing size are stuck together, the smaller one on top of the larger one with their fronts on different planes. After the photographs are taken by ATM monitoring camera, there are apparent size and spherical deformations. By adopting the above methods of correction, we find the size deformation is completely rectified and the spherical deformation is remarkably improved.

Figure 1. Original image by ATM camera

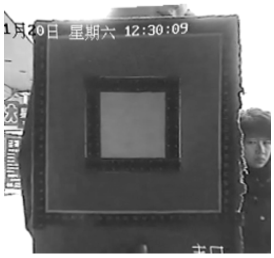


Figure 2. Image effect after correction

Case two: In 2008, there was a grave crime of fraud in Nanjing. The image of the suspect was recorded and obtained by ATM video. The size of the video was 352X288. There were both size and spherical deformation. After the correction process, the size deformation was entirely corrected and the spherical deformation was greatly improved.

Figure 3. Original spherical deformed image

Figure 4. Image effect after correction

REFERENCES

[1] Li Lian industry, Su Chang, edited by Wang Chao-cheng, TV monitoring system qualification course, China Electric Power Press. pub, 2006.
[2] R.Y. Tsai and T.S. Huang.Multiple frame image restoration and matching. in Advances in Computer Vision and Image Processing. Greenwich, CT: JAI Press Inc., 1984, pp. 317-339.
[3] J. J. Clark, M R. Palmer, and P.D. Laurence. Atransformation method for the reconstruction of functions from nonuniformly spaced samples. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-33, 1985, pp. 1151-1165.
[4] Huanfeng Shen, Pingxiang Li, Liangpei Zhang, Yindi Zhao, A MAP Algorithm to Super-Resolution Image Reconstruction. Proceedings of the Third International Conference on Image and Graphics. 2004 IEEE.
[5] Haitao Shen. Analysis and Simulation of Super-Resolution Reconstruction from Compressed Video Based on Bayesian Theory.
[6] CHEN Guangsheng,LI Shutao. Reconstruction of Super - r esolution Image using MAP and POCS Algor ithms. Science Technology and Engineering, Vol. 6 No. 4 Feb. 2006
[7] Xu Zhong-qiang, Zhu Xiu-chang. Super-resolution Reconstruction Technology for Compressed Video. Journal of Electronics & Information Technology, Vol.29No.2, Feb. 2007

A MOTION DETECTION ALGORITHM BASED ON THREE FRAME DIFFERENCE AND BACKGROUND DIFFERENCE

Mingming Yang, Qingping Guo
Department of Computer Science and Technology, Wuhan University of Technology
Wuhan , Hubei 430063, China
Email: yangmmgood@163.com; qpguo@mail.whut.edu.cn

ABSTRACT

The frame difference method is simple and meets real time request, so it is used to implement the video surveillance system, but the frame difference method may cause some errors, such as: two peaks, holes. The background difference method do not exist holes but it is sensitive to the environment**.**

In this paper, we use three frame difference method and background difference method to implement the motion detection, three frame difference method fits for environmental change effectively, it can overcome the lack of the background difference method, on the other hand, background difference method does not exist two peaks and holes, so, we combine three frame difference with background difference, in the process, we use open operation to eliminate the very small object and use close operation to fill the tiny hole, then, we can obtain the more fully and appropriately result. It has been found that it can efficiently detect the moving objects, which lays the foundation for the further study.

Keywords：Motion Detection, Optical Flow Method, Frame Difference Method, Background Difference Method, Three Frame Difference Method

1. INTRODUCTION

With the development of science and technology, the society is changing rapidly, surveillance system is more and more widely used in daily life, which plays an important role in modern life. If we assign someone to do it, it must be not only baldness but also waste human resource, in addition, sometimes, it may be unreliable. So, we choose computer instead, when exception has happened, it will give an alarm, which effectively reduces labor intensity and improves the quality of life of the staff.

In the process, motion detection is the most important technology, in this paper, we use the three frame difference method and background difference method to research and implement the motion detection.

1.1 Motion Detection

Motion detection divides the moving region from the image sequence, it is the key technology of the image processing and analyzing. However, due to the video surveillance systems application environment is quite complex, such as the weather, light etc, makes motion detection become a key in video surveillance systems. At present, optical flow method、background difference method and frame difference method are the common methods of motion detection[1-3].

(1) Optical flow method
The advantage of optical flow method is that we can detect the moving object without any information of the scene, and it allows the camera to move. Owe to the impact of noise、multisource、shadow and transparency, the distribution of optical flow is not accurate and reliable, in addition, the optical flow is time-consuming, so it is hard to be put into practice.

(2) Background difference method
Background difference method is a way to compare the input image with the background model, according to judging the statistical information of gray scale or histogram, it can infer whether an exception has happened. This method is simple, under the condition of known background, it can extract the whole target image, but it is sensitive to the external condition, such as weather、light, etc.

(3) Frame difference method
Frame difference method is a way that subtracts two or three adjacent frames of the Motion image sequence, then, judge from the absolute difference of the image gray-scale whether is greater than the threshold, to determine whether there is a moving object. Generally, this method cannot be extracted from the full features of all relevant pixels, but it is not sensitive to the light, so this method applies to the dynamic environment. Owing to its simplicity、fast of testing, it is widely applied as basic algorithm to rapidly judge the moving object in surveillance system.

1.2 Open Operation and Close Operation

In this paper, we use open operation to eliminate the very small object and use close operation to fill the tiny hole.

1.2.1 Erosion and Dilation

(1) Erosion
We obtain Sx by Structural elements S translates x, if Sx is included in X, we record this point, all of the points that meet the above conditions make up of a congregation, which is called the result of X eroded by S. It is shown in the formula 1-1.

$$x \otimes s = \{x \mid S + x \subseteq X\} \tag{1-1}$$

(2) Dilation
We obtain Sx by Structural elements S translates x, if Sx unites x is not equal to null, we record x. All of the points that meet the above conditions make up of a congregation, which is called the result of X dilated by S. It is shown in the formula 1-2.

$$x \oplus s = \{x \mid Sx \cup x \neq \varnothing\} \tag{1-2}$$

1.2.2 Open Operation and Close Operation

(1)Open operation is the operation that we use the erosion first and then dilation. It is shown in the formula 1-3.

$$OPEN(X, B) = (X \otimes B) \oplus B \tag{1-3}$$

(2)Close operation is the operation that we use the dilation first and then erosion. It is shown in the formula 1-4.

$$CLOSE(X,B) = (X \oplus B) \otimes B \qquad (1\text{-}4)$$

2. MOTION DETECTION

2.1 Noise Removal

As the video collection device made noise in the process of collecting and the changes of environment also bing noise, the video data must be polluted, in addition, the changes of environment also bring noise. While obtaining the difference image, the noise is brought at the same time, so we should remove the noise at next step.The noise follows the Gaussian distribution [8], so we can use the characteristic of Gaussian distributions to remove the noise.

The characteristic of Gaussian noise is noise energy most concentrated in the three times scope of the variance, however, the gray distribution of the moving region is often dispersed. The distribution of the noise influences by the moving region, but the area of moving region is finite, so the distribution of the noise still follows Gaussian distributions. In the whole picture, the noise is uniform and quite different from the moving region. So we can use a single threshold to divide the moving region and the noise, in this paper, we use the distribution range of noise instead of the threshold [5].If the moving region is small , we can neglect it, then, we obtain the distribution by subtracting two frames, thereby, we can estimate the mean and variance, and remove the noise. When the moving region becomes bigger, we should consider the process of computing distribution.

In the moving region, as the interference of the moving object, the gray value changes a lot, so the variance grows, in the region without moving object, the distribution of noise is steady, the variance of this region is much less than the variance of the moving region, the overall variance belongs to the range. Assume that the overall variance is detTotal, the variance of the moving region is detMotion, noise region is detNoise, then,

detNoise <detTotal<detMotion (2 - 1)

So, we can use the variance as the criterion to divide the moving region and the noise.

In this paper, we partition the frame difference image into pieces, then, we eliminate the piece that contains the moving region, which makes the moving region vanished. After the partition of image, the image divides into three groups:

(1) Only contains the noise
(2) Only contains the moving region
(3) Contains both noise and moving region

As the variance between noise and moving region is significant different, it is very easy to distinguish. For the third group, the partition is vague, we can use the closeness of variance to partition them, according to the formula 2-1, we take the overall variance as the boundary value. As the affect of the moving region, the boundary value get bigger, it is not accurate to partition them by boundary value, we need to modify this value, according the current variance, we partition the piece, then, we get a new moving region and noise, the partition continues until the change between new variance and the last variance is very little, then, we take the variance as the final estimated result.

The flow of getting estimated parameter shown in figure 2-1.

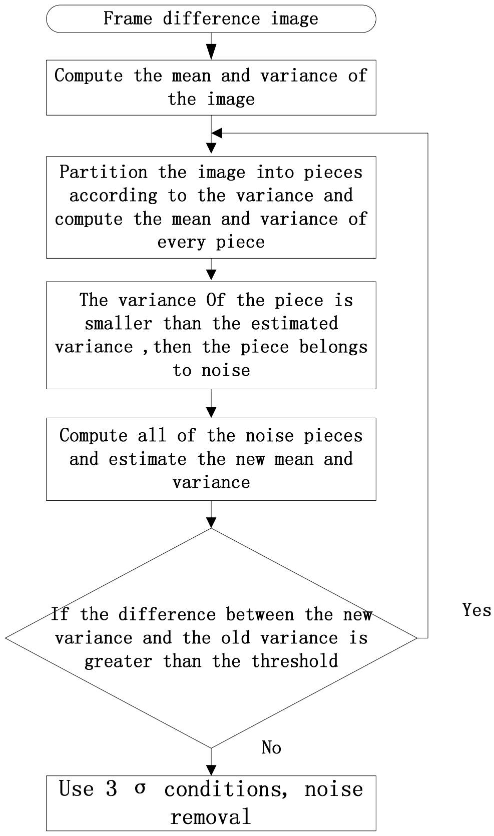


Figure 2-1. The Flow of Noise Removal

2.2 Three Frame Difference Method

In the image processing, frame difference method is to compute two adjacent frames to obtain the moving region, this method can measure the moving object rapidly, but the moving target that obtains by this method is greater than the full size, then, it may cause two peaks, in addition, this method can not measure the overlapping part, it may causes holes. If we use three adjacent frame method, the three adjacent frames will compute again, so it can effectively measure the profile of the moving target.

When we detect the monitor scene by using frame difference method, the luminance component of image sequence includes most of the information, so we can use the gray image sequences to motion detection. In this paper, we use the formula 2-2 to convert the collected image to gray image[4].

Y=0.299R+0.587G+0.114B (2 - 2)

After gray image, we should compute difference image, according to the formula 2-3.Assume that $I_i(x,y)$ is the image sequence, thereinto, (x,y) indicates the pixel position, i indicates the I frame, we choose three adjacent frames $I_{i-1}(x,y)$, $I_i(x,y)$, $I_{i+1}(x,y)$, then, we obtain the difference image $d_{(i,i-1)}(x,y)$ and $d_{(i,i+1)}(x,y)$:

$$\begin{cases} d_{(i,i-1)}(x,y) = |I_i(x,y) - I_{i-1}(x,y)| \\ d_{(i,i+1)}(x,y) = |I_i(x,y) - I_{i+1}(x,y)| \end{cases} \qquad (2-3)$$

After computing, we use the method which is introduced in the 2.1 to remove the noise, if the pixel point is noise, then, the pixel value become zero.

Then, we implement image binaryzation according to the formula 2-4:

$$\begin{cases} b_{(i,i-1)}(x,y)=\begin{cases}255 & d_{(i-1,i)}(x,y)\neq 0\\ 0 & d_{(i,-1,i)}(x,y)=0\end{cases} \\ b_{(i,i+1)}(x,y)=\begin{cases}255 & d_{(i,i+1)}(x,y)\neq 0\\ 0 & d_{(i,i+1)}(x,y)=0\end{cases}\end{cases} \qquad (2-4)$$

After obtaining the $b_{(i,i-1)}(x,y)$ and $b_{(i,i+1)}(x,y)$, we use the formula 2-5 to compute them, and then obtain the binaryzation image $B_i(x,y)$:

$$B_i(x,y)=\begin{cases}255 & b_{(i,i-1)}(x,y)\cap b_{(i,i+1)}(x,y)=255\\ 0 & b_{(i,i-1)}(x,y)\cap b_{(i,i+1)}(x,y)\neq 255\end{cases} \qquad (2-5)$$

2.3 Background Difference Method

In practice, light、weather can cause the change of the background, in this case, it is important to update the background in time, we optionally update the background, in this paper, we use the formula 2-6 to update the background, in fact, we just need update the region which does not contain the moving region.

$$B_{i+1}(x,y)=\begin{cases}B_i(x,y) & C_i(x,y)=1\\ \beta B_i(x,y)+(1-\beta)\ I_i(x,y) & C_i(x,y)\neq 1\end{cases} \qquad (2-6)$$

$B_i(x,y)$ is the background value of the current frame, $B_{i+1}(x,y)$is the background value of the next frame, $I_i(x,y)$is the gray value of the current frame, $\beta \in (0,1)$is the modulus of update, which controls the speed of update, in this paper, β =0.006.

After getting background image B_i, according to the formula 2-7 and 2-8, we obtain d by substracting background B_i from current frame I_i, if the d is greater than the threshold T, this pixel belongs to moving object.

$$d=|\ I_i(x,y)-B_i(x,y)\ | \qquad (2-7)$$

$$DB_i(x,y)=\begin{cases}255 & d\geq T\\ 0 & d<T\end{cases} \qquad (2-8)$$

$I_i(x,y)$is gray value of current frame, $B_i(x,y)$ is gray value of background frame, $DB_i(x,y)$is gray value of difference image, T is the threshold.

2.4 Open Operation and Close Operation

For the moving target obtained by background difference, the profile and internal is integrated, but there will be a lot of noises. So we use the open operation, first we use erosion to eliminate the tiny point, isolated point, isolated region, and then use dilation to fill the gap and hole, in this case, target size is constant.

For the moving target obtained by three frame difference, the profile and internal isnot integrated. So we use close operation, first we use dilation to fill the gap and hole, and then use erosion to eliminate the tiny object, in this case, target size is constant too.

2.5 The Implement of Motion Detection

After open operation and close operation, according to the formula 2-9, we compute the result of the three frame difference and the result of the background difference by using or operator, then, we can obtain the more fully and appropriately result.

$$C_i(x,y)=\begin{cases}255 & DB_i(x,y)\cup B_i(x,y)=255\\ 0 & DB_i(x,y)\cup B_i(x,y)\neq 255\end{cases} \qquad (2-9)$$

$DB_i(x,y)$is the result of the background difference, $B_i(x,y)$is the result of the three frame difference, $C_i(x,y)$is the moving target.

3. PERFORMANCE

In the testing stage, we have got the hardware platform as following:

CPU : Intel(R) Pentium(R) 4 CPU 2.40GHz
Memory : 1GB

In this stage, the software we use is visual c++6.0, the three adjacent frame is shown in the figure 3-1、figure 3-2 and figure 3-3. The figure 3-4 is the result processed by frame difference, and the figure 3-5 is the result processed by the method of the paper introduced.

Compare the figure 3-4 with figure 3-5, we can find that the method that we introduce in this paper can obtain the moving object much more effective and reliable, and the profile is much more integrated.

Figure3-1. The First Frame

Figure3-2. The Second Frame

Figure3-3. The Third Frame

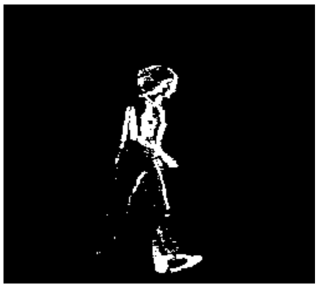


Figure3-4. TheResult which is Processed by Frame Difference

Figure3-5. The Result which is Processed by Three Frame Difference and Background Difference Method

4. CONCLUSIONS

This paper mainly uses the noise removal、three frame difference method and background difference method and open operation、close operation, it has been found that it can quickly and effectively detect the moving target, however, in order to better meet the requirements of surveillance, we have to be further research and improvement in system reliability, stability, bandwidth utilization on the network.

REFERENCES

[1] Hao Rongxia, Xu Xudong. Research on algorithm of motion detection in a video surveillance system [J],Computer Applications and Software,Vol.25,No.4,Apr 2008,pp.190-191

[2] Kang li, Lang Xihua,A Motion Detection Research Based on Three Frame Difference and Background Difference[J], Science and Technology Information,2008,pp.417-417.

[3] Wang hui. The Processing of Background Difference Images[D], Harbin University of Science and Technology,2005,pp.12-15.

[4] Zhang Ruiping.Research and Development of Digital Video Surveillance System Based on DirectShow[D], Wuhan university of technology,2008,pp.28-31.

[5] Wang hui.The Processing of Background Difference Images[D], Harbin University of Technology Engineering,2005,pp.3-7,12-15.

[6] Zhang yu. Facial Feature Locating Algorithm Based on Inter-Frame-Differential and Skin Color-Space Segmentation [J], Journal of UEST of China, 2005,pp.497-500.

[7] Chen xi,Zhang xiaoyu.A New Method of Moving Targets Detection Based on Background Subtraction[J], Journal of Chengdu University(Natural Science Edition), 2008,pp.137-140.

[8] Wang de yang. Research on Key Technology of Intelligent Video Surveillance[D], Xidian University,2005,pp.9-10.

Mingming Yang is a master candidate of school of computer science and technology, Wuhan university of technology, focuses his study on parallel computing and multimedia
.

SURVEY OF CLOTH MODELING IN COMPUTER GRAPHICS

Xinrong Hu, Yan Bai, Shuqin Cui, Xiaoqin Du, Zhongmin Deng
School of Computer Science, Wuhan University of Science and Engineering
Wuhan, China
Email: hxr@wuse.edu.cn

ABSTRACT

Many approaches are used to model cloth drape: physical schemes and geometrical schemes. A variety of cloths modeling techniques are reviewed in this paper. First we introduce some classic cloth modeling schemes and analysis the features of each scheme. Then the efficiency of each scheme is discussed separately. At last, we expect the cloth model in development should be able to model various textile materials.

Keywords: Cloth model, Geometrical model, Continuous model, Discrete model

1. INTRODUCTION

More and more researchers in textile field and graphics field put their efforts in the cloth design, modeling and animation with various approaches. Both physical and geometrical modeling schemes are proposed to express the properties of cloth drape. Weil [1] proposed the first cloth modeling scheme in 1986: he described the cloth with geometrical model. Feynman [2] improved it and proposed the first physical model in 1986: he described the cloth with elastic shell theory. In 1988, Moore et al[3] proposed a scheme to solve some key techniques in flexible object animation, such as conflict detection. After that, many researchers found many ways to express cloth behavior with accuracy, stability, efficiency and collision handling.

In the past time, many textile researchers focus on the mechanical properties of cloth such as stretching, bending, shearing and cloth anisotropic behavior. The first physical model—particle model was proposed by Breen et al[4-6] in 1992. In Breen's model, he use the KES experimental data to fit the simulation curve and describe the inherent properties of fabric by acquiring some successive piecewise linear functions. After that, many researchers used physical model to simulate the mechanical property of fabric. In 1995, Provot [7-8] proposed a Spring-Mass model to model cloth. There are many critical techniques in cloth simulation, such as cloth model, numerical integration, collision and self-collision handle. Cloth model has much influence on the cloth simulation. In this paper, we concerned more on the development of cloth models and discussed some typical cloth models based on the graphics research.

2. CLOTH MODELS

In the early days, the research about cloth modeling focused on the mechanical property of fabric. Peirce [9] was the first one who was interested in the cloth bending behavior and the measurement for cloth properties. His model can be applied to solve some problems because it only described simple geometrical relations among the yarns of the cloth. After that, from the textile views, some researchers also used different methods to analyze the cloth model, such as Abbot[10], Cooper[11], Dahlberg[12], Lindberg[13], Grosberg[14], Konopasek [15], Kawabata[16], Lloyed [17], Brown [18] et al. However, these researchers pay more attention on the micro structure of the fiber. The physical properties can not be expressed with these methods. So many graphics researchers focused on the cloth physical properties.

There are three main cloth models in computer graphics: geometrical model, continuous model and discrete model.

2.1 Geometrical Model

In later 1980's, the early geometrical models were used to reproduce the cloth geometrical features. Weil[19] generated the cloth 3D drape images with some geometrical constraints. His model can produce the wrinkles and folds of the cloth; however, it can not simulate the cloth motion. It is suitable for generating a static cloth shape. Conquillart [20] proposed a geometrical simulation scheme for deforming surface. The scheme started with a planar surface which is embedded into a cylindrical lattice and created the deformation surface with folds effects. Aono et al.[21-24] proposed dart insertion and mapping algorithm to fit a 2D cloth onto a 3D surface.

These geometrical models is lack of flexibility, that is, it can create creases and folds easily at one moment but hard to do it at another moment. They are purely geometrical models for fabric but limited in application. In garment application, most of these techniques need to feature the body or generate the cloth mesh. Moreover, geometrical cloth model does not consider the mechanical properties or empirical relations. It will result mapping distortions.

Hinds et al.[25] first used a technique to define a 3D garment panel as surface and apply it in garment design. Dai et al.[26] proposed a geometrical method to model cloth drape with a few shape parameters which are predicted according to the pattern structure and mechanical properties of cloth. This method is first used to model flared skirt and visualize the 3D drapeability of cloth. Chiricota[27-28] proposed a fast 3D approximation of cloth from pattern picture. His method is based on the parameterization of 2D polygonal curves and measurements. The method is applied to model certain elements inherent to the field of garment, such as sleeves and collars.

Turquin[29-30] proposed a simple method for designing a basic garment. The idea of the algorithm is that garment is generated according to the silhouette which is drawn from the front or back of a garment. Wang[31-32] present another scheme for 3D garment made. The solution is a parametric modeling method which enables the garment fit for different body shapes. It can represent the complex geometrical garment models with the freeform surface.

2.2 Continuous Model

The geometrical modeling techniques focus on creating realistic appearance such as folds and ceases, however, it did not consider the physical properties of the cloth.

Terzopoulos[33-35] is the first one of the researchers who model the freeform surface with physical techniques. The freeform cloth is considered as zero thickness in his solution. Terzopoulos applied the elastic theory to simulate the cloth like deformable object. While animating the deformable object, it is assumed that each object has a deformable energy and is discretized uniform meshes. Deformable energy is represented with numerical integration equations and is solved with a semi-implicit method.

Carignan[36] further developed the Terzopoulos's work. Based on the Terzopoulos's mechanical model, he applied triangular meshes on the garment panels which would be sewn together. Carignan is the first one to detect self-collision in the cloth.

Continuous models are continuous, but the solutions for continuous are discrete in computer simulations. So a suitable computational method is needed to solve this problem. FEM(Finite Element Method) is proposed to meet this need. Various element types are adopted in FEM, such as plates, shells and beams.

Continuous shell elements are used to model cloth with FEM, which can simulate the physical properties of cloth: bending, shearing and stretching in a formulation. Collier[37] proposed a nonlinear FEM with plates. It treated the curvature relationships with different responses happened in the weft and wrap directions. Chen et al.[38] extend the plate model with some geometrical constraints. He assumed that the length of the fabric cannot be changed whether motion or not.

Ascough[38] proposed a beam lattice model that consider the fabric as a lattice. His approach illustrates a good solution for large displacements to simulate cloth drape feature. Moreover, it is more representative for the microstructure of fabric and has better robustness.

Hu et al.[39-42] proposed a new fabric mechanical model with FVM(Finite Volume Method), which became an attractive research field. Hu's experiments proved that his model is an efficient model.

In all of the continuous models, the great achievement is that the implicit integration is used to solve linear simulation formulations.

2.3 Discrete Model

As mentioned above, computer only can deal with discrete signals. After continuous model, more and more discrete models are presented one after another. Particle system is one of typical discrete models.

Particle system considered the particle is the primary element of the cloth and the cloth is constructed with polygonal or triangular meshes. The mechanical behavior of the cloth can be represented with the motion of the particles. These particles moved with the external forces or internal forces between particles.

Breen et al[4-5] simulated cloth with a set of particles according to Newton Motion Law. In his model, cloth surface is constructed with polygonal or triangular meshes where each vertex is a particle with physical properties. Realistic surface can be obtained with some KES experiment data applied in their model. Weber et al.[43] developed the Breen's model to model specific cloth properties such as anisotropic and hysteresis. Ng et al.[44] used polygonal meshes and EMM(Energy Minimization Method) to create the cloth shape with several energy equations, which include the bending constant, elasticity and density etc. Thalmann et al.[45] animated cloth from two aspects: one is dynamic accuracy of the cloth model, and the other is the techniques of the real-time visualization. Plath[46], Jojic[47], Oshita[48], HeeJung[49], Dohta[50] and Boxerman[51] proposed a similar new technique to simulate the dynamic behavior of cloth.

The well-known particle system is mass-spring system. It is a physical based system and used in modeling deformable objects widely. In this system, each particle is a mass, which connected with the other mass by three springs: bending spring, stretching spring and shearing spring. An advantage of mass-spring systems is that they do not require the mesh resolution as fine as the FEM while simulating the cloth's folds and ceases. The mass-spring is widely used for its simple implement.

Provot[7-8] adopted the mass-spring system to model the cloth behavior. In his scheme, he applied Explicit Euler's method to solve the cloth motion equations. He also presented a new solution to solve the super-elassticity problem with a position modification algorithm to the end of over elongated springs. However, repositioning can result over elongation of the other springs and may require more iteration steps to animate the cloth. Vassilev et al.[52-55] extend this method by modifying the mass speed instead of the position. Some similar schemes further developed and simulated the dynamic cloth. In Bridson's work[56], he expressed some concepts such as mass-spring modeling, animation, collision detection and collision handling etc.

Choi et al[57] proposed an immediate buckling model and a semi-implicit cloth simulation technique for essentially simulating realistic cloth animations without introducing buckling instability. An attempt to incorporate the bending model as a part of cloth drape simulation is also made. Denise[58] proposed a differential model to solve the cloth hysteretic property. It incorporated many fabric parameters in cloth model and reproduced some KES tests with traditional mass-spring system.

Other helpful works have been done[59-61]. In these literatures, some adaptive meshes generation scheme are proposed and some formulas for estimating the curvature at the mass can be found. Such as Volkov [62-64] presented a method to introduce adaptive meshes into the most elaborated cloth simulation approach based on irregular triangular meshes.

3. IMPORTANT APPLICATION ARIA FOR DISTRIBUTED PARALLEL PROCESSING

In the paper, we discuss the cloth modeling methods. Geometrical modeling techniques do not consider the physical properties of cloth. Continuous model and discrete model belong to physical modeling technique. Continuous approaches offer a strong theoretical background but are usually time consuming and not easy to use for dynamic simulation for complex models with interaction between different types of tissues. Discrete approaches can build complex models and are more efficient than continuous. The physical based models usually express the cloth model with triangular or rectangular meshes which is linked with points of finite mass at the intersections.

Since those models have massive data to process， it is necessary to design efficient parallel algorithms and using distributed computing resources to deal with them. Also visualization of cloth modeling is an important issue in the electronic design. Those are further works of our research team.

REFERENCES

[1] J.Weil. The Synthesis of Cloth Objects, Computer Graphics (Proc. SICCRAPH): 49-54, 1986.

[2] Carl. Feyman. Modeling The Appearance of Cloth, Master Thesis, Massachusetts Institute of Technology, 1986.

[3] M. Moore and J. Wilhelms. Collision Detection and Response for Computer Animation, Computer Graphics (Proc. SICCRAPH), 1988.

[4] D. Breen, D. House and M. Wozny. Predicting the Drape of Woven Cloth Using Interacting Particles. Computer Graphics (Proc. SICCRAPH): 365-372, 1994.

[5] D. Breen, D. House and P.Getto. A Physical-Based Model of Woven Cloth, The Visual Computer, Vol(8): 264-277, 1992.

[6] D. Breen, D. House and M. Wozny. A Particle-Based Model for Simulating the Drape Behavior of Woven Cloth, Textile Research Journal, Vol(64):663-685, 1994.

[7] X. Provot. Deformation Constraints in a Mass-Spring Model to Describe Rigid Cloth Behavior, Graphics Interface’95, 1995.

[8] X. Provot. Collision and Self-Collision Handling in Cloth Model Dedicated to Design Garments, Graphics Interface’97, 1997.

[9] F.T. Peirce. The Handle of Cloth as a Measurable Quantity, Journal of The Textile Institute, Vol(21): 377-416,1930.

[10] N.J. Abbott. The Measurement of Stiffness in Textile Cloths, Journal of Textile Research, Vol(21):435-444,1951.

[11] D.N. Cooper. The Stiffness of Woven Textile, Journal of The Textile Institute, Vol(51): 317-335,1960.

[12] B. Dahberg. Mechanical Properties of Textile Cloth (Part II), Journal of Textile Research, Vol(31):94 -99,1961.

[13] J. Lindberg, B. Behre and B. Dahberg. Mechanical Properties of Textile Cloth (Part III), Journal of Textile Research, Vol(31):99-122,1961.

[14] P. Grosberg and N.M. Swani. The Mechanical Properties of Woven Cloth, Journal of Textile Research, Vol(36):338-345,1966.

[15] M. Konopasek and J.S. Hearle. Computational Theory of Bending Curves, Journal of Fibre Science and Technology, Vol(5):1-28,1972.

[16] S. Kawabata, M. Niwa and H. Kwai. The Finite Deformation Theory of Plain Weave, Vol(2):21-85,1973.

[17] D.W. Lloyd. The Analysis of Complex Cloth Deformations, Mechanics of Flexible Fibre Assemblies, Sijthoff & Noordhoff International Publishers B.B,1980.

[18] P.R.Brown, D.R.Buchanan and T.G. Clapp. Large-Deflection Bending of Woven Cloth for Automated Material Handling. Journal of The Textile Institute, Vol(21): 377-416,1930.

[19] J.Weil. The Synthesis of Cloth Objects, Computer Graphics (Proc. Of SIGGRAPH): 49-54,1986

[20] S. Coquillart. Extended Free-Form Deformation: A Sculpturing Tool for 3D Geometrical Modeling, Computer Graphics, Vol(24):1990.

[21] M. Aono, D.E. Breen and M.J. Wozny. A Computer-Aided Broadcloth Composite Layout Design System, Proc. Of Geometrical Modeling for Product Realization:1993.

[22] M. Aono, D.E. Breen and M.J. Wozny. Fitting a Woven Cloth Model to a Curved Surface: Mapping Algorithm, Computer-Aided Design, Vol(26):278-292,1994.

[23] M. Aono, P. Denti, D.E. Breen and M.J. Wozny. Fitting a Woven Cloth Model to a Curved Surface: Dart Insertion, Computer Graphics and Application, Vol(16):60-70,1996.

[24] M. Aono, D.E. Breen and M.J. Wozny. Modeling Methods for The Design of 3D Broadcloth Composite Parts, Computer-Aided Design, Vol(33):989-1007,2001.

[25] K. Hinds and J. Mccartney. Interactive Garment Design, Journal of The Visual Computer Graphics, Vol(6):53-60, 1990.

[26] Xiaoqun Dai, Takao Furukawa and Shigeru Mitsui. Drape Formation Based On Geometric Constraints and Its Application to Skirt Modeling, Journal of Clothing Science and Technology, Vol(13):23-31,2001.

[27] Y. Chiricota. Geometrical Modeling of Garments, Journal of Clothing Science and Technology, Vol(13):38-52,2001.

[28] Y. Chiricota. Three-Dimensional Garment Modeling Using Attribute Mapping, Journal of Clothing Science and Technology, Vol(15):346-358,2003.

[29] E. Turquin. A Sketching Interface for Garment design, 2004.

[30] E. Turquin, M.P. Cani and J.F. Hughes. Sketching Garments for Virtual Characters, Eurographics Workshop on Sketch-Based Interfaces and Modeling, August,2004.

[31] C.L. Wang, Y. Wang and M.F. Chen. Feature Based 3D Garment Design Through 2D Sketched, Journal of Computer-Aided Design, Vol(35):659-672,2003.

[32] C.L. Wang, Y. Wang and M.F. Chen. Design Automation for Customized Apparel Products, Journal of Computer-Aided Design, Vol(375):675-691,2005.

[33] D.Terzopoulos, A.Barr, J. Platt and K. Fleischer. Elastically Deformable Models, Journal of Computer Graphics, Vol(21):205-214,1987.

[34] D.Terzopoulos and K. Fleischer. Modeling Inelastic Deformation: Viscoelasticity, Plasticity, Fracture, Journal of Computer Graphics, Vol(22):269-278,1988.

[35] D.Terzopoulos and K. Fleischer. Deformable Models, Journal of The Visual Computer, Vol(4): 306-331,1988.

[36] M. Carignan, Y. Yang,N.M. Thalmann and D. Thalmann. Dressing Animated Synthetic Actors with Complex Deformable Clothes, Journal of Computer Graphics, Vol(26):99-104,1992.

[37] J.R. Collier. Drape Prediction by Means of Finite-Element Analysis, Journal of Textile Institute, Vol(82):96-107,1991.

[38] J. Ascough, H.E. Bez and A.M. Bricis. A Constrained Finite Element Method for Modeling Cloth Deformation, Journal of The Visual Computer, Vol(15):90-99,1999.

[39] J.L. Hu and J.G. Teng. Computional Cloth Mechanics:Present Status and Future Trends, Finite Elements in Analysis and Design, Vol(21):225-237,1996.

[40] J.G.Teng, S.F. Chen and J.L. Hu. A Finite-Volume Method for deformation Analysis of Woven Fabric, International Journal for Numerical Methods in Engineering, Vol(46):2061-2098,1999.

[41] S.F. Chen, J.L. Hu and J.G. Chen. A Finite Method for Contact Drape Simulation of Woven Fabrics and Garments, Journal of Finite Elements in Analysis and Design, Vol(37):513-531,2001.

[42] J.L. Hu, S. Chung and M.Luo. Effect of Seams on Fabric Drape, International Journal of Clothing Science and Technology, Vol(9):220-227,1997.
[43] Weber, B. Eberhardt and W.Strasser. A Fast, Flexible Particle-System Model for Cloth Draping, IEEE Computer Graphics and Applications, Vol(16):52-59,1996.
[44] H.N. Ng, R.L. Grimsdale and W.G. Allen. A System for Modelling and Visualtion of Cloth Material, Journal of Computers & Graphics, Vol(19):423-430,1995.
[45] N.M. Thalmann, P. Volino and F. Cordier. Avenues of Research in Dynamic Clothing, Proc. Of Computer Animation:19-21,2002.
[46] J. Plath. Realistic Modeling of Textile Using Interacting Particle Systems, Journal of Computer &Graphics, Vol(24): 897-905,2000.
[47] N. Jojic, Gu J., Shen T.S. and Huang T.S. Computer Modeling, Analysis, and Synthesis of Dressed Humans, Journal of Circuits and Systems for Video Technology, Vol(9):378-388,1999.
[48] M.Oshita and A. Makinouchi. Real-time Cloth Simulation with Sparse Particles and Curved Faces, Proc. Of Computer Animation,:7-8,2001.
[49] B. Heejung, R. Kwyanwoo and J. Byungtae. Procedural Approach to Generate Real Time Motions of Cloth, Proc. Of Shape Modeling:2002.
[50] T. Dohta, M. Hattori and T. Takamori. Ana Intuitive Operation System for Cloth Animation of Computer Graphics, Proc. of The 41st SICE Annual Conference: 5-7,2002.
[51] E. Boxerman. Speeding Up Cloth Simulation, University of McGill: 2003.
[52] T. Vassilev, B. Spanlang and Y. Chrysanthou. Cloth Modeling for Real-time Animation, Proc of Computer Graphics Forum: 2001.
[53] T. Vassilev, B. Spanlang and Y. Chrysanthou. Fast Cloth Animation on Walking Avatars, Proc of Computer Graphics Forum: 2001.
[54] T. Vassilev. Dressing Virtual People, Proc of SCI'2000: 23-26,2000.
[55] T. Vassilev and B. Spanlang. Efficient Cloth Model For Dressing Animated Virtual People, Proc of SCI'2000: 23-26,2000.
[56] R. Bridson, R. Fedkiw and J. Anderson. Robust Treatment of Collisions, Contact and Friction for Cloth Animation, Proc of Eurographics/SIGGRAPH Symposium on Computer Animation: 2003.
[57] K.J. Choi and H.S. Ko. Stable but Responsive Cloth, Proc of ACM Trans. Graph, Vol(21): 604-611, 2002.
[58] J. Denise, D. Reversat and A. Gagalowicz. Modeling Hysteretic Behaviour of Fabrics, Proc of Mirage 2003: 10-12,2003.
[59] E. Grinspun, P. Krysl and P. SchrOder. Charms: A simple Framework for Adaptive Simulation, Proc of ACM tans.: 2002.
[60] J. Villard and H. Borouchaki. Adaptive Meshing for Cloth Animation, Proc of The 11th International Meshing Roundtable: 2002.
[61] Jianzhen Lin, Min Tang and Jinxiang Dong. Cloth Simulation Based On Local Adaptive Subdivision and Merging, Proc of the 8th International Conference on Computer Supported Cooperative Work in Design: 2004.
[62] V. Volkov and L. Li. Adaptive triangular Meshes for Cloth Simulation: 2005.
[63] V. Volkov and L. Li. Real-Time refinement and Simplification of Adaptive Triangular Meshes, Proc of IEEE Visualization 2003: 2003.
[64] L.Li and V. Volkov. Cloth Animation with Adaptively Refined Meshes, Proc of 28th Australasion Computer Science Conference: 2005.

Xinrong Hu, is an associate professor of Wuhan University of Science & Engineering. She graduated from Wuhan University of Technology in 1997, and a Ph.D candidate in Pattern Recognition and Intelligent Control of Huzhong University of Science & Technology. Her research interests include numerical analysis of fabric drape and simulation, dynamics and control of flexible mechanisms and physics properties analysis in cloth simulation, virtual reality.

DETECTION-BASED ALGORITHMS FOR LOCATING INITIAL SHAPE ON FACE

Dai Wei, Wang Shitong
School of Information Technology, Jiangnan University
Wuxi, Jiangsu 214122, China
Email: loveispdvd@gmail.com

ABSTRACT

In this paper, we introduce three detection-based algorithms for locating initial shape on face. These three algorithms are all employ single face detector or some facial local feature detectors, and combine with statistical model derived from training set to infer initial shape. From uni-detector to multi-detector, from mean to Least-Squares Regression (LSR), they improve gradually, using more detectors instead of single detector to impose more constraints on initial shape, using more sophisticated statistical model to find the relationship between detected features and initial shape more precisely. The algorithm that uses multi-detector and LSR model shows the best performance on precision in our experiments.

Keywords: Shape Analysis, Face Detector, Feature Points Location, Facial Local Feature, Least-Squares Regression.

1. INTRODUCTION

Locating facial features, which searches the locations of facial feature points in given regions of images or image sequences, is one of the core issues in face recognition. The locating algorithms complete the conversion from lower-level semantic features to high-level semantic features in face recognition.

Active Shape Model [1] and Active Appearance Model [2, 3] are two algorithms widely used for locating facial features, but these two algorithms have the same shortcoming: convergence speed and precision rely heavily on initial shape. If initial shape starts far away from real shape, then not only will the number of iterations increase, but the possibility of falling into local minimum increase. Therefore, there is an urgent need for high real-time algorithms for locating initial shape.

Face detection must be carried out to determine whether faces exist in an image before locating facial features. The Viola-Jones detector [4] and the Fröba-Küblbeck detector [5] are both well-developed real-time face detection algorithms and can be used to detect eyes, nose, mouth and other local features aside from face at high speed. Consequently, we could detect face and local features first, and then locations of feature points are estimated through locations and sizes of detected face and local feature.

In this paper, we propose three detection-based algorithms for locating initial shape on face and compare their performances. Section 2 presents the three algorithms. Section 3 gives some experimental results. Section 4 concludes this paper.

2. DECTION-BASED ALGORITHMS FOR LOCATING INITIAL SHAPE ON FACE

Before introducing the three algorithms, some mutual concepts in the three algorithms must be defined below.

(1) Let training set have m samples and each sample have n feature points. Then shape s of a sample is defined as all x-coordinates followed by all y-coordinates of n feature points in this sample.

$$\mathbf{s} = (x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) \tag{1}$$

(2) Face rectangle or local feature rectangle can be obtained as result of detector. Both face and local feature rectangle have four parameters: (x, y, w, h), where x and y are x- and y-coordinate of the rectangle center, w and h are width and height of the rectangle.

(3) Original shape can be normalized with regard to face or local feature rectangle to normalized shape.

$$\begin{aligned} x_i &= \frac{x_i - x}{w} \\ y_i &= \frac{y_i - y}{h} \end{aligned} \tag{1}$$

(4) Normalized shape also can be reverted with regard to face or local feature rectangle to reverted shape.

$$\begin{aligned} x_i &= w \times x_i + x \\ y_i &= h \times y_i + y \end{aligned} \tag{2}$$

2.1. Global Mean Shape Method

Global Mean Shape Method (GMSM) is the simplest one in detection-based algorithms for locating initial shape. The algorithm is described below.

A. Training stage:

(1) Face detector is used to obtain face rectangles $\mathbf{F}_i$ of all samples in training set, where $i = 1 \ldots m$.

(2) Every shapes $\mathbf{s}_i$ is normalized with regard to corresponding $\mathbf{F}_i$.

(3) Mean shape $\bar{\mathbf{s}}$ is calculated from all normalized shapes.

$$\bar{\mathbf{s}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{s}_i \tag{3}$$

B. Inferring stage:

(1) Face detector is used to obtain face rectangle F of an unseen sample.

(2) The mean shape $\bar{\mathbf{s}}$ obtained in training stage is reverted with regard to F. The reverted shape is the initial shape inferred by GMSM.

GMSM move and resize the mean shape with regard to the detected face rectangle. Although GMSM imposes global constraint on initial shape and can produce the initial shape which fits well the real shape on the whole, it has weak ability to locate locally initial shape while variations of expression and pose exist in samples.

2.2. Local Mean Shape Method

Due to possible large local deviation caused by GMSM, we consider using some local feature detectors instead of a single face detector to locate locally initial shape. It is necessary to find an automatic method for grouping feature points according to different detectors. The algorithm Local Mean Shape Method (GLSM) is described below.

A. Training stage:

(1) *k* different local feature detectors are used to obtain local

feature rectangles $\mathbf{F}_{ij}$ of samples in training set, where $i = 1 \ldots m$, $j = 1 \ldots k$.

(2) n feature points is divided into k groups. Automatic grouping method is generally the following: First, the sum d_j of distances between a given feature point of all training samples and the jth local feature rectangle is calculated. For all k sums { d_j }, if the lth sum d_l is the least, then the given feature point is classified to the lth group.

(3) For all m training samples and all k local shapes, local shape $\mathbf{s}_{ij}$ is normalized with regard to corresponding $\mathbf{F}_{ij}$.

(4) For all k normalized local shapes and the jth normalized local shape, local mean shape $\overline{\mathbf{s}}_j$ is calculated from all local shapes in training set.

$$\overline{\mathbf{s}}_j = \frac{1}{m}\sum_{i=1}^{m}\mathbf{s}_{ij} \tag{4}$$

B. Inferring stage:

(1) k different local feature detectors are used to obtain k local feature rectangles $\mathbf{F}_j$ of an unseen sample, where $j = 1 \ldots k$.

(2) The k local mean shapes { $\overline{\mathbf{s}}_j$ } obtained in training stage are reverted with regard to $\mathbf{F}_j$, and then the feature points in k local mean shapes are rearranged with the order before grouping. The shape composed of the rearranged feature points is the initial shape inferred by LMSM.

LMSM is a generalization of GMSM. While k = 1 and the one detector is face detector, LMSM degenerates into GMSM. Comparing with GMSM, LMSM places some local constraints on local shapes, so LMSM locates initial shape more precisely than GMSM. However, LMSM is unsatisfying while in-plane rotation of face exists in samples.

2.3. Least-Squares Shape Method

Least-Squares Shape Method (LSSM) is different statistical model built from the two algorithms above in statistical model, which is least-squares regression instead of mean model. The algorithm is described below.

A. Training stage:

(1) Face detector is used to obtain face rectangles of samples in training set, and every shape is normalized with regard to corresponding face rectangle.

(2) k different local feature detectors are used to obtain local feature rectangles $\mathbf{F}_{ij}$ of samples in training set, where $i = 1 \ldots m$, $j = 1 \ldots k$.

(3) For all m training samples and all k local shapes, $\mathbf{F}_{ij}$ is normalized with regard to corresponding face rectangle. The way to normalize rectangle is the same as shape.

(4) The coordinates of k local feature rectangles constitute a $2k\times1$ vector $\mathbf{f}_i$. All $\mathbf{f}_i$ constitute a $m\times2k$ matrix $\mathbf{S}$.

(5) All shapes in training set constitute $m\times2n$ matrix $\mathbf{F}$.

(6) Solving linear equation

$$\mathbf{F}\times\mathbf{X} = \mathbf{S} \tag{5}$$

(7) The linear coefficient matrix $\mathbf{X}$ is found by

$$\mathbf{X} = \mathbf{F}^{\dagger}\times\mathbf{S} \tag{6}$$

where $\mathbf{F}^{\dagger}$ is the generalized inverse of $\mathbf{F}$.

B. Inferring stage:

(1) k different local feature detectors are used to obtain k local feature rectangles $\mathbf{F}_j$ of an unseen sample, where $j = 1 \ldots k$.

(2) Face detector is used to obtain face rectangle of the unseen sample, and every local feature rectangle is normalized with regard to the face rectangle.

(3) The coordinates of k local feature rectangles constitute a $2k\times1$ vector $\mathbf{f}$.

(4) Shape $\mathbf{s}$ is calculated by

$$\mathbf{s} = \mathbf{f}\times\mathbf{X} \tag{7}$$

(5) $\mathbf{s}$ is reverted with regard to the face rectangle. The reverted shape is the initial shape inferred by LSSM.

As all local feature rectangles and shapes are normalized to corresponding face rectangles, it is unnecessary to add constant element 1 to $\mathbf{f}_i$ as common least squares regression. LSSM fits the relationship between local features and feature points, and it has remarkable ability to locate initial shape.

3. EXPERIMENTAL RESULTS

We test the three algorithms introduced in section 2 on IMM face database. The experiment repeats ten times on 100 training samples and 10 test samples randomly chosen from IMM face database. Let the initial shape $\mathbf{s}_p = (x_{p1}, x_{p2}, \ldots, x_{pn}, y_{p1}, y_{p2}, \ldots, y_{pn})$ and the real shape $\mathbf{s}_r = (x_{r1}, x_{r2}, \ldots, x_{rn}, y_{r1}, y_{r2}, \ldots, y_{rn})$, the precision of locating initial shape is showed in figures and measured on the mean distance between the initial shape $\mathbf{s}_p$ and the real shape $\mathbf{s}_r$

$$MeanDistance = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(x_{pi}-x_{ri})^2+(y_{pi}-y_{ri})^2} \tag{8}$$

Figure 1. Results of Locating Initial Shape. Top-left is Real Shape, Top-right is Initial Shape Located by GMSM, Bottom-left LMSM, Bottom-right LSSM

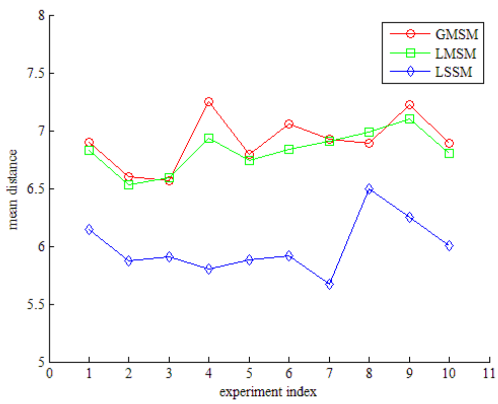


Figure 2. Comparing Precision of the Initial Shapes Located by the Three Algorithms

From **Figure 1** and **Figure 2**, we learn that the precision of LSSM is the highest, and LMSM a little higher than GMSM.

4. CONCLUSIONS

In this paper, we have proposed three detection-based algorithms for locating initial shape on face. In the three algorithms, LSSM puts in an excellent performance; GMSM and LMSM are just fair.

There is still much room for improvement in LSSM, which is the best algorithm in this paper. As LSSM is based on regression, all improvements to regression also can apply to LSSM. Kernel methods can fit the non-linear relationship between local features and feature points. Partial least squares regression can remove the dependence in local features while fitting. Ridge regression and the lasso can impose constraints on variations of local features, and can control and explain the models better.

REFERENCES

[1] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application", Computer vision and image understanding, 61, 38-59 (1995)

[2] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models", IEEE Transactions on Pattern Analysis and Machine Intelligence, 23, 681-685 (2001)

[3] T. Cootes and C. Taylor, "Statistical models of appearance for computer vision", World Wide Web Publication, February, (2001)

[4] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple", *Computer Vision and Pattern Recognition*, 1 (2001).

[5] B. Froba and C. Kublbeck, "Real-time face detection using edge-orientation matching", Lecture notes in computer science, 78-83 (2001)

Dai Wei is a graduate student of School of Information Technology, Jiangnan University. He graduated from Beijing University of Technology in 2005. He is interested in computer vision, pattern recognition, artificial intelligence and machine learning.

Wang Shitong is a Full Professor and a dean of School of Information Technology, Jiangnan University. He graduated in 1982 and received his MSc in Computer Science in 1988 from Nanjing University of Aeronautics and Astronautics. He was a visiting professor in some universities of English, Japan and Hongkong. He researches mainly on neural network, fuzzy system and fuzzy artificial intelligence.

IMAGE'S EDGE DETECTION OF PESTS IN STORED GRAIN BASED ON GREY THEORY

[1]Shuang-de Zhang, [2]Long Zhou
Wuhan Polytechnic University
Wuhan 430023, P. R. China
Email: zhangshd@whpu.edu.cn

ABSTRACT

The method of pests detection in stored grain is always investigated. Researching of the harmful stored grain insects on-line examination system, not only has the important academic value, but also has the broad application prospect. The method based on image recognition is often discussed. With development of computer technology, information processing, pattern recognition, intelligence detection, detection method based on image recognition develops fast and becomes main direction of grain pests intelligence detection. This paper puts forward an edge detection algorithm based on grey relation analysis. At first the reference series and compare series are defined. Then the relevant coefficients between the reference series and compare series are calculated to every pixel. Finally, the edge detection is processed and its application in image's of pests in stored grain is discussed. The examples show that the method can better detect the image's edge of pests in stored grain.

Keywords: Grey Relation Analysis, Pests in Stored Grain, Image Recognition

1. INTRODUCTION

China is the biggest grain production, preservation and consumption nation in the world, No doubt, it is of great significance to make good grain preserved. In recent years, our country's grain annual output has amounted to 500 billion kilograms, the reserve reaches as high as more than half of the annual output. [1]Even more, the sorts and density of pests in stored grain present a rising trend in recent years, and this leads to a greater loss of stored grain. In order to prevent and cure the pests effectively, it is important to forecast their occurrence trend, number, species development and their latent harms. Moreover, the different prospective results made by different measures and strategies of prevention and cure should be evaluated. The outstanding problem that exists in traditional detection method of food insect (method of normal sampling, sampling and entrapment) is that the domestic grain-situation detection system can't detect the pests, and the foreign food insect sound detection technology can't realize the number detection of pests. Aimed at this situation, we found that the machine visual technology which develops rapidly in recent years can solve the problem. [2-3]

During the course of food insect image processing, restraining the noise is a very important part. The edge outline of the food insect images have many important information , in order to analyze and identify the sort of food insect precisely , a food insect imager with clear edge outline is needed . Therefore, the first step of analyzing a food insect image contaminated by noise is to restrain the noise. [4-5]

In the traditional method of restraining signal noise based on Fourier Transform, if the frequency domain of signal and noise are overlapped, the effect will be bad. Because the noise spectrum of food insect image is distributed in high frequency end, the signal and noise can't be separated well apart only with Fourier Transformation; therefore the function of restraining noise can't work. Fourier Transformation takes the sine curve as Orthogonal function , but many important features of images are highly localized in spatial position , these image elements aren't similar to Fourier basic function , their transform coefficients aren't distributed compactly either, so Fourier transformation coefficients can not get the best expression when the images with a lot of localized elements are analyzed . [6-9]

There are many methods for edge detection, such as template matching, differential coefficient method, statistics method and outline matching etc. , and there are also many classical basic algorithm for edge detection , such as Sobel operator , grads operator , Marr operator , Robert operator , Prewitt operator, Laplace operator, Gauss partial differential filter and Canny edge detector etc. . A good method for edge detection should be provided with the right ability to restrain sorts of noise, and to keep the self-contained edge character, but the classical edge detection algorithm can't achieve this effect. [10-11]

Currently, main on-line pest detection methods include sampling method, bait method, near infrared method, stridulation and image recognition method, etc. With development of computer technology, information processing, pattern recognition, intelligence detection, detection method based on image recognition develops fast and becomes main direction of grain pest intelligence detection.

2. IMAGE'S EDGE DETECTION OF PESTS IN STORED GRAIN BASED ON GREY RELATION ANALYSIS

In the on-line examination system, the image which gains through CCD (Charge Coupled Devices) does not make us satisfy extremely, because of illumination non-uniformity, dust influencing CCD photographic camera and so on, it is necessary to carry on the noise and smoothening pretreatment to the image. Obviously, the pretreatment is an important foundation in image analysis such as characteristic forms, withdraws, compression and so on image analysis.

The essential principle of grey relation analysis is: distinguishing the relational degree of the factor in the system according to the similarity degree of geometrical shape of the statistical serial curve. The more similarity of geometrical shape of the serial curves, the bigger relation degree value they have. During an image's edge detection, it may be considered as a grey system and analyzed by using grey relation.

Edge is the part of image which has the most remarkable change of brightness, and a point whose pixel value is same as or has little difference with that of its neighborhood is defined as a non-edge point. Then the pixel value of an ideal non-edge point is completely same as that of its neighborhood. Consequently, this pixel and its neighborhood may form a reference sequence. If they have the same value, then this point is a non-edge point. For convenient and available computation, the reference sequence is a 9 points sequence whose neighborhood values are all equal to 1(Figure 1); the compare sequence consists of each pixel and its 8 neighborhood pixel points. For convenient comparing and analysis and guaranteeing the equivalent of each sequence, the pixel point would be interval-valued and the value of compare sequence would be converted into that of the interval-value [0,1], that is:

$$x_{i,j} = \left(x^*_{i,j} - x_{i,jmin}\right) / \left(x^*_{i,jmax} - x^*_{i,jmin}\right) \quad (1)$$

Where the $x^*_{i,jmax}$ is the maximum of the comparative sequence and the $x^*_{i,jmin}$ is the minimum of the comparative sequence.

Accordingly, the reference sequence and comparative sequence of an image whose size is $m \times n$ (Figure 2) may be gotten as follow:

$$x_0 = \{1,1,1,1,1,1,1,1,1\} \quad (2)$$

$$x'_{i,j} = \{x_{i-1,j-1}, x_{i-1,j}, x_{i-1,j+1}, x_{i,j-1}, x_{i,j}, x_{i,j+1}, x_{i+1,j-1}, x_{i+1,j}, x_{i+1,j+1}\} \quad (3)$$

where $i = 1,2,\cdots,m$, $j = 1,2,\cdots,n$

For easy expressing, $x'_{i,j}$ is represented by x_k in the comparative sequence, where $1 \le k \le m \times n$.

1	1	1
1	1	1
1	1	1

$x_{i-1,j-1}$	$x_{i-1,j}$	$x_{i-1,j+1}$
$x_{i,j-1}$	$x_{i,j}$	$x_{i,j+1}$
$x_{i+1,j-1}$	$x_{i+1,j}$	$x_{i+1,j+1}$

Figure 1. Reference Sequence **Figure 2.** Comparative Sequence

Relation coefficient of the comparative sequence x_k which is formed by centering on the pixel point k in the h-th point and the reference sequence x_0.

$$\xi_k(h) = \left[\min_{1\le k\le m\times n}\min_{1\le h\le 9}\left|x_0(h)-x_k(h)\right| + \rho\times \min_{1\le k\le m}\min_{1\le h\le 9}\left|x_0(h)-x_k(h)\right|\right] /$$
$$\left[\left|x_0(h)-x_k(h)\right| + \rho\times \min_{1\le k\le m\times n}\min_{1\le h\le 9}\left|x_0(h)-x_k(h)\right|\right]$$

(4)

Where $1 \le h \le 9$ means the sequential number of a 9-point sequence; $1 \le k \le m \times n$ means the comparative sequence; ρ is differentiation coefficient whose value may be selected in $0.1 \sim 1.0$, usually $\rho \le 0.5$.

c. Calculating the grey relation degree of the comparative sequence x_k which is formed by centering on the pixel point k and the reference sequence x_0

$$r_k = \frac{1}{9}\sum_{h=1}^{9} \xi_k(h) \quad (5)$$

The closer it is to the reference sequence's changing tendency, the bigger relation degree it has. If r_k is greater than a certain threshold value φ, usually $\varphi \ge 0.8$, it means this point has the same characteristic as the reference sequence, and it is not an edge point; contrarily, it is considered as an edge point.

The main advantage of grey relation analysis is small required sample size, which causes a low computation quantity.Figure 3 is the original image of pests in stored grain whose size is 2048×1536, Figure 4, Figure 5 and Figure 6 are the results of edge detection by using Sobel operators, Robert operators and grey relation analysis respectively, and fig 5 is the result of edge detection based on grey relation analysis, where φ=0.9**.**

From comparison between Figure 4, Figure 5 and Figure 6, the edge detection results based on Sobel operators have much noise and fuzzy pick-up edges. However, the method based on grey relation analysis has the advantage of detecting the characteristic edge of pests in stored grain clearly and continuously, building a solid foundation for pick-up of their further characteristics.

On-line pest detection method based on image recognition is the development trend of condition forecasting of the pests in stored grain. Edge detection arithmetic is the key to recognition. Starting from the basic thought of gray system, this paper applies grey relation analysis to edge detection of the image of pests in stored grain. The research results show that it is simple, utility, efficient arithmetic and may build a good base for further characteristics pick-up and pattern recognition of pests in stored grain.

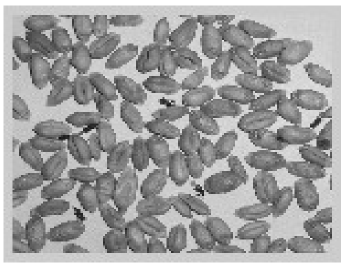


Figure 3. Original Image of Pests in Stored Grain

Figure 4. Edge Detection Result Based on Sobel Operators

Figure 5. Edge Detection Result Based on Robert Operators

Figure 6. Edge Detection Result Based on Grey Relation Analysis

3. CONCLUSIONS

Based on the pattern recognition, the on-line grain insect examination is the development tendency of the harmful stored grain insect measuring and reporting, the pre-processing of the grain insect image is the key and important foundation of recognition. The Calculation results show that application of grey relation analysis to restraining noise of the image and edge detection not only can keep the edge characteristics of the image with good visual effect but also can get clearer food insect image character, and lay a good foundation on distilling the food insect image character.

4. ACKNOWLEDGEMENT

This paper is supported by National Natural Science Foundations under Grant 79970025 of China, Hubei Provincial Science Foundation under Grant 2007ABA408, Hubei Provincial Department of Education under Grant Q200618002.

REFERENCES

[1] Zayas I Y，Flinn P W ．Detection of insects in bulk wheat samples with machine vision[J]．Trans．ASAE，1998(3)：883—888．

[2] Wilkin D R,Fleural-Lessard F,The detection of insects in grain using conventional samplings spears[C]. Proceeding 5th Int.Wkg,Conf. On Stored-Product Protection, Bordeaux, France,1990，1445-1450。

[3] Haykin S.Neural Networks[M].Macmillan College Publishing Company,1994,7-16.

[4] Chambers J,Conve I A,Van Wyk C B,et al, NIR analysis for the detection of insect peats in cereal grains[C],Proc.Int.Conf.on Diffuse Spectroscopy,MD USA, 1992, 96-100.

[5] Osumna E,Freund R,Girosi F.Training support vector machines:An application to face detection[C].Proceeding of CVPR'97,Puerto Rico,1997,78-81.

[6] Ju-long Deng. Grey forecasting and decision-making [M].Huazhong University of Science and Technology Press, 1986,134-140.

[7] Xu Fang, Qiu Daoyin, pattern recognition in granary harmful insect examination aspect applied research, Zhengzhou engineering college journal. 2001.2,47-50.

[8] Wan Zhengqun, the current our country science guarantees the grain,The preserve of grain. 1996.2,44-48.

[9] Mr. Dai, Mathematics morphology in imagery processing application ,Foshan science and technology university (natural sciences version). 1998,46-49.

[10] Zhoulong, Application of two-dimension wavelet transform in image process of pets in stored grain . 2008 International Conference on Machine Learning and Cybernetics, 2790-2793,2008.

[11] Wang Xiaodan. Image process based of MATLAB，Xi'an university of electrical science and technology press，2006,16-19.